frontiers | Research Topics

Large Language Models for medical applications

Edited by
Ariel Soares Teles, Alaa Abd-alrazaq, Thomas F. Heston
and Rafat Damseh

Coordinated by
Livia Ruback

Published in
Frontiers in Medicine

ISSN 1664-8714
ISBN 978-2-8325-6448-6
DOI 10.3389/978-2-8325-6448-6

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Large Language Models for medical applications

Topic editors

Ariel Soares Teles — Federal Institute of Education, Science and Technology of Maranhão, Brazil
Alaa Abd-alrazaq — Weill Cornell Medicine, Qatar
Thomas F. Heston — University of Washington, United States
Rafat Damseh — United Arab Emirates University, United Arab Emirates

Topic coordinator

Livia Ruback — State University of Campinas, Brazil

Citation

Teles, A. S., Abd-alrazaq, A., Heston, T. F., Damseh, R., Ruback, L., eds. (2025). *Large Language Models for medical applications*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-6448-6

Table of contents

frontiers | Frontiers in Medicine

TYPE Editorial
PUBLISHED 30 May 2025
DOI 10.3389/fmed.2025.1625293

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Beatriz S. Lima,
Research Institute for Medicines (iMed.ULisboa), Portugal

*CORRESPONDENCE
Ariel Soares Teles
ariel.teles@ifma.edu.br

RECEIVED 08 May 2025
ACCEPTED 14 May 2025
PUBLISHED 30 May 2025

CITATION
Teles AS, Abd-alrazaq A, Heston TF, Damseh R and Ruback L (2025) Editorial: Large Language Models for medical applications. *Front. Med.* 12:1625293. doi: 10.3389/fmed.2025.1625293

Editorial: Large Language Models for medical applications

Ariel Soares Teles[1]*, Alaa Abd-alrazaq[2], Thomas F. Heston[3], Rafat Damseh[4] and Lívia Ruback[5]

[1]Campus Araioses, Federal Institute of Maranhão, Araioses, Maranhão, Brazil, [2]AI Center for Precision Health, Weill Cornell Medicine - Qatar, Doha, Qatar, [3]Department of Family Medicine, University of Washington, Seattle, WA, United States, [4]Computer Science & Software Engineering Department, United Arab Emirates University, Al-Ain, Abu Dhabi, United Arab Emirates, [5]School of Technology, University of Campinas, Limeira, São Paulo, Brazil

KEYWORDS

Large Language Models, LLMs, natural language processing, NLP, medical education, decision support systems

Editorial on the Research Topic
Large Language Models for medical applications

1 Introduction

The advent of Large Language Models (LLMs) marks a transformative moment in the evolution of Artificial Intelligence (AI), particularly in their capacity to process and generate human language with remarkable fluency and contextual awareness. These models, trained on vast and diverse corpora, have demonstrated state-of-the-art performance across a range of Natural Language Processing (NLP) tasks. In the medical domain, their potential is especially compelling: from synthesizing complex biomedical literature and supporting clinical decision-making to enhancing patient communication and enabling more equitable access to health information.

This Research Topic was launched to explore the multifaceted role of LLMs in transforming healthcare delivery and medical research. The objective was to gather interdisciplinary contributions that investigate both the capabilities and the limitations of LLMs when applied to clinical decision support, patient engagement, precision medicine, and beyond. We aimed to foster a comprehensive dialogue that includes technical innovations, ethical reflections, and practical case studies. The collection features fifteen published articles, reflecting a diverse range of perspectives and methodological approaches. This Research Topic aspires to illuminate the pathways for integrating LLMs into medical practice while addressing the critical questions that accompany their adoption.

2 Clinical decision support and diagnostics

One of the most promising applications of LLMs in medicine lies in their potential to support clinical decision-making and diagnostic reasoning. Ríos-Hoyo et al. assessed GPT-3.5 and GPT-4 on 75 complex diagnostic cases and found that GPT-4 included the correct diagnosis in 68% of cases and ranked it among the top three in 42%. The study highlighted GPT-4's superior accuracy and consistency compared to GPT-3.5, though both models showed limitations. Notably, diagnostic success was more strongly associated with literature prevalence than disease incidence, reinforcing that LLMs should currently be viewed as decision support tools rather than standalone diagnostic systems.

Yin et al. conducted a comparative assessment of four language models, including GPT-4.0, in answering pediatric asthma-related questions. GPT-4.0 showed the highest scores across dimensions such as accuracy and completeness, although all models had limitations in addressing treatment-specific questions. Lee et al. explored the use of GPT-4 as a simulated digital twin for neurological history-taking in cases of headache, stroke, and neurodegenerative disease. Their tripartite model demonstrated 81% overall accuracy in retrieving history of present illness details, supporting the potential of LLMs in structured pre-consultation workflows. Liu et al. presented MED-ChatGPT CoPilot, an AI-assisted system that leverages prompt engineering and GPT-4 to extract structured medical case data from scientific literature, build a vector-based local knowledge base, and deliver diagnostic and therapeutic suggestions through a chatbot interface.

In the oncology domain, Peng et al. developed an interpretable machine learning model for predicting survival in aggressive prostate cancer using SHAP-based explanations. Among nine algorithms tested, LightGBM offered the best prognostic performance, with 1-, 3-, and 5-year AUCs exceeding 0.77. The use of SHAP allowed the identification and ranking of key clinical features influencing survival predictions. Zhang et al. proposed PMPred-AE, a deep learning model based on EfficientNetV2-L with an attention mechanism, for the automatic detection of pathological myopia. The model achieved high accuracy across training, validation, and test sets, and incorporated Grad-CAM for visual interpretability, allowing clinicians to see which retinal regions influenced the model's decisions, making it both an effective and explainable diagnostic tool.

3 Patient-facing applications

Studies have showcased how generative AI can directly support patients in managing their health (i.e., a patient-facing LLM application), particularly through accessible and personalized tools. Jin et al. evaluated a GPT-based recipe generation tool designed to improve the nutritional management of individuals undergoing peritoneal dialysis. The pilot study found significant improvements in serum prealbumin levels, suggesting that personalized dietary plans generated via LLMs can be both clinically effective and user-friendly. In another clinical application, Aydin et al. (a) offered a critical perspective on the use of LLMs for patient-centered medication guidance and self-decision support. While highlighting the promise of these tools in enhancing health literacy and supporting patients in remote or resource-limited settings, the authors caution against over-reliance on AI-generated information, particularly in high-stakes scenarios involving drug interactions or complex conditions.

4 Education and training

The integration of generative AI into health profession education is prompting both enthusiasm and caution, as emerging research examines its impact on learners' preparedness, skills, and ethical sensibilities. In a reflective opinion article, Sharifi Kelarijani et al. argue that, while tools like ChatGPT offer new opportunities for nursing education (e.g., rapid access to information and assistance with assignments), they also risk diminishing students' critical thinking, communication, and clinical reasoning skills if used without proper pedagogical oversight. Echoing these concerns from a student-centered perspective, Gualda-Gea et al. surveyed senior medical students and found limited prior exposure to AI tools but strong recognition of their future importance. Most students supported integrating AI into the curriculum, though many also expressed concern about ethical implications, potential biases, and over-reliance on chatbot-generated information.

Beyond student attitudes, the practical application of LLMs in health education is beginning to take shape. Aydin et al. (b) conducted a scoping review that selected 201 articles and mapped out how LLMs are currently being used to support patient education, identifying six key themes ranging from generating patient-friendly educational materials to enhancing doctor-patient communication. LLMs were found to demonstrate the ability to deliver accurate answers to patient questions, improve the quality of existing educational content, and rephrase medical information in a way that is easier for patients to understand. Nonetheless, issues related to readability, accuracy, and potential biases remain a concern.

5 Medical documentation and synthetic text creation

Two contributions to this Research Topic explored distinct applications of generative AI within the clinical domain. Lu et al. evaluated the use of GPT-4o for generating medical history records and found that its outputs were comparable in quality to those written by resident physicians. This points to a promising role for LLMs in supporting clinical documentation workflows. Differently, Ren et al. addressed challenges related to data access and privacy by proposing a method for generating synthetic clinical letters using pre-trained language models. Their framework enables the creation of de-identified, yet semantically rich, clinical texts that can be used for training and evaluating downstream NLP tasks such as named entity recognition.

6 Implications and future directions

LLMs demonstrate promise across varied healthcare contexts but require rigorous evaluation and safeguards. Their integration into medical and healthcare practice by clinicians, nurses, and pharmacists offers the potential to streamline clinical decision support, diagnostics, management, patient-facing applications, education/training, medical documentation, and synthetic text creation. However, realizing this potential demands more than technical advancement; it requires a concerted effort to ensure ethical, transparent, and accountable implementation.

Wang et al. reinforced these considerations through a comprehensive bibliometric analysis of ChatGPT's application in nursing. Their study highlights growing international interest, particularly in domains such as nursing education and clinical decision-making, while also pointing to the fragmented and early-stage nature of the research landscape. Despite increasing publication volume and global engagement, collaboration across author groups remains limited, and ethical concerns, including misinformation, over-reliance, and data security, are insufficiently addressed. These findings underscore the need for interdisciplinary cooperation, empirical evaluation, and a stronger emphasis on responsible innovation as LLMs become more integrated into healthcare practice.

Importantly, as Bélisle-Pipon cautions, we must also reevaluate how we conceptualize the shortcomings of LLMs. Framing their inaccuracies as mere "hallucinations" may obscure the deeper epistemic issue: that these models generate plausible text without any concern for truth. Recasting such failures as "bullshit", in the philosophical sense of conveying information without regard to accuracy, underscores the serious risks of over-reliance on LLMs in high-stakes clinical settings. This critique invites the medical and healthcare AI community to adopt a more skeptical and reflective posture, one that resists hype and prioritizes verification, contextual understanding, and human oversight.

This Research Topic highlights both the extraordinary potential, ethical aspects, and the current limitations of LLMs in healthcare applications by physicians, nurses, and pharmacists. It offers a foundation for critical inquiry as the field matures. Going forward, the responsible deployment of LLMs in healthcare must be guided not only by innovation but also by ethical foresight, transparency, and a deep commitment to patient well-being.

Author contributions

AT: Writing – original draft, Writing – review & editing. AA-a: Writing – review & editing. TH: Writing – original draft, Writing – review & editing. RD: Writing – review & editing. LR: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Ariel Soares Teles would like to thank the National Council for Scientific and Technological Development—CNPq [grant numbers 308059/2022-0 and 441817/2023-8]. Lívia Ruback would like to thank the Programa de Incentivo aos Novos Docentes (PIND) from Unicamp, grant number 519.287.

Acknowledgments

We thank all researchers who contributed to this Research Topic and the editorial team at Frontiers in Medicine for their support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors declared that they were editorial board members of Frontiers at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Medicine

TYPE Brief Research Report
PUBLISHED 20 June 2024
DOI 10.3389/fmed.2024.1380148

Check for updates

OPEN ACCESS

EDITED BY
Thomas F. Heston,
University of Washington, United States

REVIEWED BY
Abdallah Al-Ani,
King Hussein Cancer Center, Jordan
Xia Jing,
Clemson University, United States

*CORRESPONDENCE
Lajos Pusztai
✉ lajos.pusztai@yale.edu
Frederick M. Howard
✉ frederick.howard@uchospitals.edu

†These authors share first authorship

RECEIVED 01 February 2024
ACCEPTED 10 June 2024
PUBLISHED 20 June 2024

CITATION
Ríos-Hoyo A, Shan NL, Li A,
Pearson AT, Pusztai L and Howard FM (2024)
Evaluation of large language models as a diagnostic aid for complex medical cases.
Front. Med. 11:1380148.
doi: 10.3389/fmed.2024.1380148

Evaluation of large language models as a diagnostic aid for complex medical cases

Alejandro Ríos-Hoyo[1†], Naing Lin Shan[1†], Anran Li[2], Alexander T. Pearson[2], Lajos Pusztai[1*] and Frederick M. Howard[2*]

[1]Yale Cancer Center, Yale School of Medicine, New Haven, CT, United States, [2]Department of Medicine, University of Chicago, Chicago, IL, United States

Background: The use of large language models (LLM) has recently gained popularity in diverse areas, including answering questions posted by patients as well as medical professionals.

Objective: To evaluate the performance and limitations of LLMs in providing the correct diagnosis for a complex clinical case.

Design: Seventy-five consecutive clinical cases were selected from the Massachusetts General Hospital Case Records, and differential diagnoses were generated by OpenAI's GPT3.5 and 4 models.

Results: The mean number of diagnoses provided by the Massachusetts General Hospital case discussants was 16.77, by GPT3.5 30 and by GPT4 15.45 ($p < 0.0001$). GPT4 was more frequently able to list the correct diagnosis as first (22% versus 20% with GPT3.5, $p = 0.86$), provide the correct diagnosis among the top three generated diagnoses (42% versus 24%, $p = 0.075$). GPT4 was better at providing the correct diagnosis, when the different diagnoses were classified into groups according to the medical specialty and include the correct diagnosis at any point in the differential list (68% versus 48%, $p = 0.0063$). GPT4 provided a differential list that was more similar to the list provided by the case discussants than GPT3.5 (Jaccard Similarity Index 0.22 versus 0.12, $p = 0.001$). Inclusion of the correct diagnosis in the generated differential was correlated with PubMed articles matching the diagnosis (OR 1.40, 95% CI 1.25–1.56 for GPT3.5, OR 1.25, 95% CI 1.13–1.40 for GPT4), but not with disease incidence.

Conclusions and relevance: The GPT4 model was able to generate a differential diagnosis list with the correct diagnosis in approximately two thirds of cases, but the most likely diagnosis was often incorrect for both models. In its current state, this tool can at most be used as an aid to expand on potential diagnostic considerations for a case, and future LLMs should be trained which account for the discrepancy between disease incidence and availability in the literature.

KEYWORDS

large language model (LLM), ChatGPT, complex clinical cases, diagnosis, clinical case solving

10.3389/fmed.2024.1380148

1 Introduction

Large language models (LLMs) are complex, neural network-based models trained on vast amounts of text to accurately interpret human language. LLMs have been applied to a wide range of tasks within medical science, including simplifying radiology reports, accurately responding to questions posted by patients on an internet forum, generating realistic medical abstracts, and predicting in-hospital mortality (1–4). Although LLMs have shown passable accuracy in answering medical licensing exam questions in numerous studies (1–5), it is unclear if this performance can be leveraged to serve as a decision aid in real clinical practice, where cases have nuance beyond that of standardized testing. Given the widespread uptake of LLMs, they have been proposed as a diagnostic decision aid for students, and are likely in use despite the limited knowledge about specific model performance (6). Chat GPT (Generative Pre-trained Transformer) is a natural language processing model that became publicly available in November 2022, it provides outputs in response to inputs or prompts, learning its skills from internet data.

Different versions of GPT are currently available, GPT3.5 is a Chatbot based on the GPT3.5 model, whereas the GPT4 foundation features an approximately 1,000-fold increase in model parameters and an expanded context window length, resulting in an enhanced capability of solving complex tasks (7–9). GPT can be used to write computer code, analyze text, draft documents, create conversational agents, and has been shown to proficiently answer different standardized tests (7, 10) it has a considerable semantic medical knowledge and has been shown to be capable of medical reasoning (10). This has been reflected by its capabilities in answering medical questions (11), simplifying radiology reports, performing well at medical licensing exams, among others (1–4). It is currently considered an attractive tool in diverse settings of medicine, however these LLMs could potentially contribute to misinformation and exacerbate scientific misconduct in the setting of a lack of accountability and transparency.

This study aimed to characterize the performance and consistency of LLMs in diagnosing a series of challenging case records published from a single institution. In this study, we evaluated OpenAI's GPT-3.5 and GPT-4 models to establish a baseline for models trained on general (as opposed to medical-specific literature), as well as to identify patterns in misdiagnosis to inform fine-tuning of diagnostic decision aids. In this study we used cases from the Massachusetts General Hospital Case Records which have been published since 1923 in the New England Journal of Medicine. These cases have been used as teaching tools illustrating different clinical cases, and the workup of the differential diagnosis of frequently uncommon diseases or uncommon disease presentations (12). We introduced the case presentation of these clinical cases and asked GPT to provide a list of the most likely differential diagnosis.

2 Methods

Seventy-five sequential clinical cases were retrieved from the case records of the Massachusetts General Hospital, published in the New England Journal of Medicine, from January 2022 to November 2023 (12). This period was selected to ensure cases did not overlap with the training data for the LLMs. The case presentation was truncated prior to the discussant's review of the differential diagnosis, and text referencing figures or tables was removed. A uniform prompt requesting a differential diagnosis for the case presentation text was provided to OpenAI's GPT-3.5 (gpt-3.5-turbo) and GPT-4 (gpt-4) models. First, three prompts were tested on a subset of 10 cases for four replicates each. The prompts included (1) '*please read the following case, and provide a differential diagnosis for the underlying cause of this presentation*'; (2) as per (1) with the modification '*…provide a thorough and specific list of differential diagnosis…*'; and (3) as per (2) with the additional sentence *'please list the diagnosis that most explains all the features of the presentation first, and include rare diagnoses if they are the best explanation for the presentation.'* All prompts yielded similar lists, but the prompt (3) yielded diagnosis lists that most frequently listed the correct diagnosis first, and was chosen for all subsequent analysis. All clinical cases were queried with this prompt, with four replicates performed for each model (Supplementary Table 1).

The rank order of the correct diagnosis within the differential diagnosis list was established by consensus of study authors. The overlap between the full list of differential diagnoses provided by GPT and by the case discussant was similarly compared. Finally, accuracy of LLMs was correlated with disease incidence (estimated from literature review of PubMed as well as cdc.gov with references listed in Supplementary Table 1, as indexed by Google both with the search term 'diagnosis' incidence), with rare diseases without estimable incidence such as those only described in case reports assigned an incidence of 0.1 per 100,000, as well as representation of the diagnosis in medical literature as assessed by article count returned when searching for the diagnosis (or simplified surrogate term, as listed in Supplementary Table 1) in PubMed (conducted with an article cutoff of April 21st, 2023).

2.1 Statistical analysis

A Mann–Whitney U test was used to compare the number of diagnoses provided by case discussants and GPT models. A Fisher's exact test was used to compare whether the first diagnosis was the correct diagnosis, whether among the top three diagnosis was the correct diagnosis, whether the correct diagnosis was in the list of differential diagnosis from GPT3.5 and 4. To assess whether GPT was able to provide the correct diagnosis among different medical specialties, five groups were designated [Group 1: neurology and psychiatry; group 2: oncology and hematology; group 3: infectious diseases, internal medicine, endocrinology and toxicology; group 4 rheumatology, allergy and autoimmune diseases; group 5: others (cardiology, gastroenterology, genetic diseases, dermatology, nephrology and pediatrics)], A Fisher's exact test was used to compare results between GPT 3.5 and 4. A multivariable logistic regression model was used to determine the association between disease incidence and PubMed article count with these same three performance metrics. To assess the similarity between the differential diagnosis lists, the Jaccard similarity index was used (ranging from 0 to 1, 0 reflects no similarity, whereas 1 reflects a complete similarity between the analyzed sets), utilizing each case entry repeat, to test differences between GPT 3.5 and 4, a Mann–Whitney test was performed. To assess reproducibility across iterations of each model, intraclass correlation coefficients (ICC) were calculated using the two-way mixed effects, absolute agreement, multiple raters/measurements formulation (13), values of <0.5 and >0.9 reflect poor and excellent reliability, respectively. Statistical analyses and graphs were performed using GraphPad Prism 9.0 (GraphPad Software, Inc., San Diego, CA) and Python version 3.7.5 (Python Software Foundation) using statsmodels 0.13.2.

3 Results

3.1 Accuracy of GPT models in complex diagnostic challenges

Seventy-five cases from the Massachusetts General Hospital Case Records were introduced to the two GPT models. Compared to the case discussants, who provided a mean of 16.77 [interquartile range (IQR) (representing the distance between the first and the third quartile) 12] diagnoses, GPT4 produced a similar number (mean 15.45, IQR 11, $p = 0.302$) of unique diagnoses over four replicates, whereas GPT3.5 listed significantly more diagnoses (mean 30, IQR 10, $p = <0.0001$). GPT4 included the correct diagnosis in its differential list in two thirds (68%) of cases, with the correct diagnosis included in the top 3 items in the differential in 42% of cases, in contrast GPT3.5 included the correct diagnosis in its differential list in half (48%, $p = 0.006$) of the cases, and the correct diagnosis included in the top three differential diagnoses in 29% ($p = 0.075$) of the cases, thus observing that GPT4 outperforming GPT3.5 in both metrics (Figure 1). GPT4 was able to formulate more specific answers that better depicted the true diagnosis in many cases. For example, in Case 6–2022 (Immune checkpoint inhibitor-induced diabetes), GPT3.5 was only able to vaguely link the presentation to immunotherapy - "Side effects of cancer treatment: The patient's symptoms could be side effects of cancer treatment such as pembrolizumab…" - whereas GPT4 concisely answered "Pembrolizumab-induced diabetes mellitus."

3.2 Consistency of GPT model diagnostic lists

As the results of GPT models may differ across repetitions, it is important to understand how the prioritization of diagnoses might change if these tools are clinically implemented. Ranking of the correct diagnosis within a differential was more consistent across repetitions for GPT4 (ICC 0.65, 95% CI 0.42–0.80) than with GPT3.5 (ICC 0.37, 95% CI–0.25 – 0.71). The differential diagnosis list generated by GPT4 also had greater overlap with the discussant's list (Jaccard Similarity Index 0.22, IQR 0.12) than GPT3.5 (0.13, IQR 0.076, $p = <0.0001$, Figure 2) – although overlap was fair at best.

3.3 Associations of model accuracy with medical specialty and disease incidence

Each case was classified into medical specialties groups ($n = 5$), among these groups, GPT4 was numerically and statistically superior to GPT3.5 in all categories except in the Rheumatology, Allergy, and Autoimmune Diseases category (Table 1). We also assessed whether model accuracy was dependent on disease incidence or representation in the literature. PubMed article count

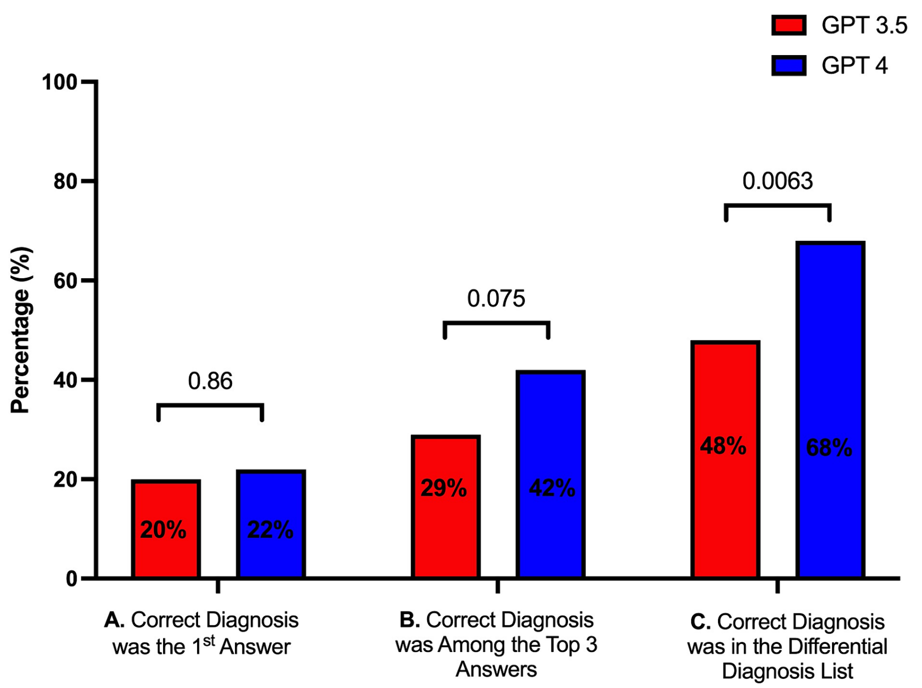


FIGURE 1
Performance of GPT3.5 and GPT4 in providing **(A)** the first diagnosis as the correct diagnosis, **(B)** the correct diagnosis among the top three diagnoses, and **(C)** the correct diagnosis among the entire list of diagnoses.

10.3389/fmed.2024.1380148

for the correct diagnosis was associated with a greater likelihood that the diagnosis would be included in the differential generated by GPT3.5 (Odds Ratio (OR) 1.40, 95% CI 1.25–1.56, $p < 0.001$) and GPT4 (OR 1.25, 95% CI 1.13–1.40, $p < 0.001$). Similar trends were seen for likelihood of a diagnosis being listed first or within the top 3 generated diagnoses (Table 2). Conversely, disease incidence was either a neutral or negative effect on the likelihood of a diagnosis being listed by either model.

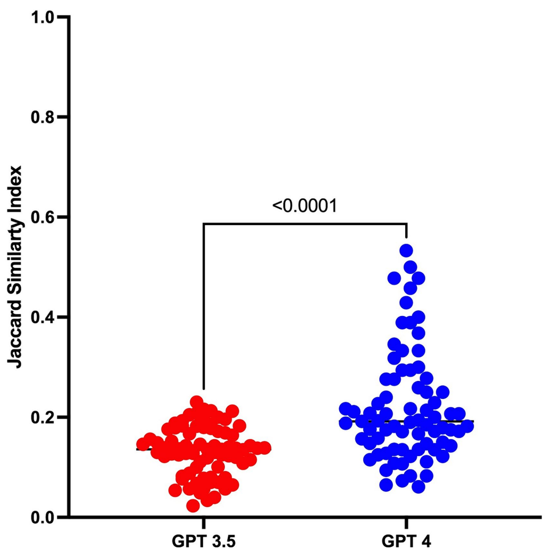


FIGURE 2
Jaccard Similarity Index indicating the overlap between GPT3.53/ GPT4 and the differential provided by the case discussant.

4 Discussion

We have demonstrated here a comprehensive characterization of the accuracy and reproducibility of two GPT models in solving complex clinical case scenarios. Whereas high accuracy was seen when evaluating GPT-3 in diagnosing common presentations such as upper respiratory tract infections (14), we found that in approximately one third of cases the best model failed to identify the correct diagnosis in complex cases. Thus, although current GPT models are insufficient to replace physician expertise, they may have some clinical utility as a diagnostic checklist (15) to reduce error when physicians are presented with a puzzling clinical scenario.

It is worth noting that although GPT3.5 was able to provide a longer list of differential diagnoses, these did not present a better concordance with the Massachusetts General Hospital case discussants diagnoses. Furthermore, GPT4 was not only better at providing the first diagnosis as the correct diagnosis, but it outperformed GPT3.5 in providing the correct diagnosis among the differential diagnosis lists.

A similar study by Zahir and collogues (16) used GPT and cases from the Massachusetts General Hospital case records to

TABLE 1 Performance of GPT 3.5 and 4 in providing the correct diagnosis, according to medical specialty.

	GPT 3.5 (%)	GPT 4 (%)	OR (95% CI)	*p*-value
Group 1 ($n = 9$)	41	72	5.2 (1.94–14.23)	0.0019
Group 2 ($n = 24$)	60	83	5.6 (2.95–10.73)	<0.0001
Group 3 ($n = 19$)	23	53	4.92 (2.39–9.77)	<0.0001
Group 4 ($n = 13$)	64	60	1.36 (0.62–3.04)	0.55
Group 5 ($n = 10$)	50	65	2.78 (1.10–6.86)	0.043

Odds ratios [OR] comparing GPT 4 vs. 3.5. Group 1: Neurology and Psychiatry, Group 2: Oncology and Hematology, Group 3: Infectious Diseases, Internal Medicine, Toxicology, Group 4: Rheumatology, Autoimmune Diseases, Group 5: Others (Cardiology, Genetic Diseases, Gastroenterology, Dermatology, Nephrology and Pediatrics).

TABLE 2 Performance of GPT 3.5 and 4 in providing the correct diagnosis, according to disease incidence and PubMed articles covering the disease.

	Top diagnosis correct		Correct diagnosis in top 3		Correct diagnosis in differential	
	OR (95% CI)	*p*-value	OR (95% CI)	*p*-value	OR (95% CI)	*p*-value
GPT 3.5						
Incidence (per 10-fold increase)	0.80 (0.67–0.95)	0.01	0.74 (0.64–0.87)	< 0.001	0.82 (0.74–0.92)	< 0.001
PubMed Articles (per 10-fold increase)	1.32 (1.12–1.56)	0.001	1.42 (1.23–1.64)	< 0.001	1.40 (1.25–1.56)	< 0.001
GPT 4						
Incidence (per 10-fold increase)	0.90 (0.80–1.02)	0.108	0.90 (0.81–0.99)	0.036	0.90 (0.82–0.99)	0.033
PubMed Articles (per 10-fold increase)	1.15 (1.01–1.30)	0.03	1.16 (1.04–1.28)	0.005	1.26 (1.13–1.40)	< 0.001

Odds ratios [OR] listed for a multivariate logistic regression including both incidence and article count.

assess whether the model's diagnoses matched the final case diagnosis, their results found an agreement between GPT4's top diagnosis and the final diagnosis in 39% of the cases, and in 64% of the cases the final diagnosis was included in the differential diagnosis list. These results contrast with ours, since we found that GPT4 was able to provide the correct diagnosis as the first answer in 22% of the cases, whereas it provided the correct diagnosis within the differential diagnosis list in 68% of the cases. In addition, Zhair's study found that GPT4 provided a mean of 9 differential diagnoses, similarly our study found a mean of 9.23 diagnoses.

Another study using a different, medicine-specific large language model called Med-PaLM, was able to provide accurate answers to different questions posted in a multiple-choice and long-form setting. Med-PaLM was superior in solving medical questions when compared to MultiMedQA (6 sets of open data that include similar questions to the United States Medical Licensing Examination (USMLE)), and HealthSearchQA (related to common consumer health related questions). MedPaLM was able to answer accurately different formats of questions, such as multiple choice and long form. In a second part of the study, clinicians from different countries were asked to solve 140 medical questions in long-form answers, the same task was performed by MedPaLM. The answers were assessed by clinicians with specialties in different medical fields, the answers provided by the LLM overall presented outstanding results, however MedPaLM's answers presented higher numbers of incorrect information, which most of the times was clinically significant (11).

When formulating a differential diagnosis, disease incidence as well as the severity/consequences of missed diagnosis are often considered (17). However, some common diseases are underrepresented in the literature, whereas some rare conditions are given particular emphasis in medical literature and educational materials. In an attempt to refine medical-domain performance, several models have been trained specifically on PubMed, which may be subject to this same bias (18). As LLMs are refined as diagnostic decision aids, strategies to align output with true disease prevalence are needed.

5 Limitations

One of the limitations of this study was the lack of publicly available diagnostic challenges with curated differential diagnosis lists, resulting in our use of a single source of cases which was only modest in size. The small sample size may lead to lower accuracy in precisely quantifying the difference in performance between the GPT models tested. Additionally, the Massachusetts General Hospital Case Records present complex cases that may not represent the most frequent case presentations – which may be more straightforward with higher diagnostic accuracy from AI models.

As the GPT models evaluated were trained on data collected on or before September 2021, and thus performance for certain diagnoses with changing epidemiology [such as monkeypox (19)] may be underestimated. We chose to evaluate OpenAI's GPT models in this study rather other LLMs due to their widespread uptake (20), as it is most likely to be in current use by physicians and trainees, and as such characterization of performance is most urgent. Furthermore, we used a single prompt to evaluate model performance in our primary analysis. Although preliminary analysis suggested that performance was similar across prompts, it is possible that modifications of the prompt may change the relative accuracy of GPT3.5 and 4 models.

Finally, although we found that disease incidence was either not associated or negatively associated with model accuracy, incidence is difficult to establish and these estimates represent our best efforts to define incidence through literature review. Incidence can vary widely depending on the population studied and across geographic regions, and these results may differ with alternate approaches to estimate incidence.

6 Conclusion

In this study we demonstrated that OpenAI's GPT-4 model outperformed GPT-3.5 in correctly diagnosing challenging clinical cases, but misdiagnosis was common, and at best such models might be used as decision aids in their current state. In training LLMs specifically as diagnostic aids, steps should be taken to account for the overrepresentation of some diagnoses in the medical literature. It is important to take into consideration certain aspect of using LLM in medicine, such as a negative impact in critical thinking, ethical considerations, as well as potentially detrimental consequences for the patient, thus the use of LLM in clinical medicine might not be ready for a global integration into clinical workflows.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

AR-H: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. NS: Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation. AL: Writing – review & editing, Software, Methodology, Investigation, Formal analysis. AP: Writing – review & editing, Software, Methodology, Investigation, Formal analysis, Data curation. LP: Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. FH: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. AR-H is supported by a grant from the Spanish Society of Medical Oncology (SEOM), LP is supported by the Susan Komen Leadership Award, and FH is supported by the NIH/NCI grant K08CA283261 and the Cancer Research Foundation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2024.1380148/full#supplementary-material

References

1. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digital Health*. (2023) 2:e0000198. doi: 10.1371/journal.pdig.0000198

2. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. (2023) 183:589–96. doi: 10.1001/jamainternmed.2023.1838

3. Yeo YH, Samaan JS, Ng WH, Ting P-S, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. (2023) 29:721–32. doi: 10.3350/cmh.2023.0089

4. Zheleiko I. Natural language processing in lifelong learning choices: a case of Finland. Lappeenranta. Lahti University of Technology LUT (2023), 12–26.

5. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digit Med*. (2023) 6:75. doi: 10.1038/s41746-023-00819-6

6. Tsang R. Practical applications of ChatGPT in undergraduate medical education. *J Med Educat Curri Develop*. (2023) 10:238212052311784. doi: 10.1177/23821205231178449

7. Open AI chat GPT. Accessed June 11, 2023.

8. Metz C. Open AI Plans to Up the Ante in Tech's A.I. Race The New York Times.

9. Koubaa A. A concise showdown. *TechRxiv*. (2023). doi: 10.36227/techrxiv.22312330.v1

10. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J, Laleh NG, et al. The future landscape of large language models in medicine. *Commun Med*. (2023) 3:141. doi: 10.1038/s43856-023-00370-1

11. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. (2022) 620:1–44. doi: 10.1038/s41586-023-06291-2

12. Harris NL. Case Records of the Massachusetts General Hospital — continuing to learn from the patient. *N Engl J Med*. (2003) 348:2252–4. doi: 10.1056/NEJMe030079

13. McGraw KO, Wong SP. "Forming inferences about some Intraclass correlations coefficients": correction. *Psychol Methods*. (1996) 1:390–0. doi: 10.1037/1082-989X.1.4.390

14. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative Pretrained transformer 3 Chatbot for clinical vignettes with common chief complaints: a pilot study. *Int J Environ Res Public Health*. (2023) 20:3378. doi: 10.3390/ijerph20043378

15. Kämmer JE, Schauber SK, Hautz SC, Stroben F, Hautz WE. Differential diagnosis checklists reduce diagnostic error differentially: a randomised experiment. *Med Educ*. (2021) 55:1172–82. doi: 10.1111/medu.14596

16. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*. (2023) 330:78–80. doi: 10.1001/jama.2023.8288

17. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: an evaluation of currently available computer programs. *J Gen Intern Med*. (2012) 27:213–9. doi: 10.1007/s11606-011-1804-8

18. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform*. (2022) 23:1–11. doi: 10.1093/bib/bbac409

19. Basgoz N, Brown CM, Smole SC, Madoff LC, Biddinger PD, Baugh JJ, et al. Case 24-2022: a 31-year-old man with perianal and penile ulcers, rectal pain, and rash. *N Engl J Med*. (2022) 387:547–56. doi: 10.1056/NEJMcpc2201244

20. Bhaimiya S. OpenAI cofounder Elon Musk said the non-profit he helped create is now focused on 'maximum-profit,' which is 'not what I intended at all'. Business Insider. (2023).

frontiers | Frontiers in Medicine

TYPE Opinion
PUBLISHED 24 July 2024
DOI 10.3389/fmed.2024.1415067

OPEN ACCESS

EDITED BY
Ariel Soares Teles,
Federal Institute of Education, Science and Technology of Maranhão, Brazil

REVIEWED BY
Silmar Teixeira,
Federal University of Piauí, Brazil

*CORRESPONDENCE
Mohamad Golitaleb
✉ mohamadgolitaleb@gmail.com

RECEIVED 20 April 2024
ACCEPTED 12 July 2024
PUBLISHED 24 July 2024

CITATION
Sharifi Kelarijani A, Safdari A and Golitaleb M (2024) Every coin has two sides: ChatGPT poses a potential threat to Nursing Students' Education. *Front. Med.* 11:1415067.
doi: 10.3389/fmed.2024.1415067

Every coin has two sides: ChatGPT poses a potential threat to Nursing Students' Education

Amirhossein Sharifi Kelarijani[1], Ali Safdari[1] and Mohamad Golitaleb[2,3]*

[1]Student Research Committee, Hamadan University of Medical Sciences, Hamedan, Iran, [2]Department of Critical Care Nursing, School of Nursing and Midwifery, Tehran University of Medical Sciences, Tehran, Iran, [3]Department of Nursing, School of Nursing, Arak University of Medical Sciences, Arak, Iran

KEYWORDS

nursing - education, ChatGPT, nursing, health profession education, nursing student

One recent example of artificial intelligence is the generative pre-trained transformer (ChatGPT) perceived as a conversational chatbot launched by the artificial intelligence company OpenAI® in 2022 and quickly gained widespread popularity (1). Recently, we have seen more research focusing on developing artificial intelligence technologies in nursing. However, most existing studies focus on the capabilities and functionalities of this robot, while due to the novelty of this technology application, there are extensive unknown dimensions associated with it. Therefore, identifying all dimensions of it requires time. These sensitivities are particularly more relevant in the education of nursing students. As nursing care continues to evolve, nursing education must also evolve. Today, the development of artificial intelligence technologies and technological advancements herald significant changes in the future of nursing (2).

Nurses constitute a vital part of the healthcare workforce worldwide and play a key role in promoting the health of communities. Therefore, health professionals' education are more important than other fields, Because the students of these fields are the future workforce and are supposed to protect human lives as their most valuable asset. Part of this importance can be attributed to the close connection between nurses' activities and patients' health and well-being. In many definitions of nursing, emphasis has been placed not only on its scientific aspect but also on its artistic aspect (3). Nursing as an art involves creatively using knowledge and science based on skill and expertise to convey emotions and concepts to others. Using chatGPT requires interpretation, sensitivity, and active participation. Skillful use of empirical knowledge to tailor to the unique needs of patients and cautious use of creativity is essential (4). Emphasizing ethical principles is another aspect of this concept.

Among the capabilities of ChatGPT in nursing education are providing nearly instant, comprehensive, and logical textual responses to instructors' academic questions, solving university assignments, and conducting research projects. Offering quick, accurate, and convincing responses can lead students to excessively trust ChatGPT as an information source and become dependent on it. Over-reliance on artificial intelligence technologies like ChatGPT may decrease direct interactions between nursing students and instructors. The experiences of specialized instructors and experienced nurses can be invaluable and help students better understand the course material. Personal experiences usually carry a higher value than pure scientific information, and ChatGPT does not provide this added value. This issue can lead to problems in developing essential skills in nursing students. These skills include critical thinking, clinical reasoning, the ability to design a nursing care plan, and problem-solving skills (5).

Furthermore, despite the ease of access to chatbots like ChatGPT, nursing students may be less inclined to find personal solutions and engage in critical thinking. This ultimately leads to the training of nurses who may need a greater understanding of patient care situations. Communication is essential in nursing, and nursing students, by developing their communication skills, enable better patient care. Suppose nursing students are constantly engaged with digital tools and artificial intelligence. In that case, their communication abilities may decrease in real-life situations, leading to insufficient and inappropriate communication with patients in clinical settings. Technological addiction, defined as excessive use of technology, is another potential threat posed by these robots to students, which can lead to psychological, social, and physical problems (6). Therefore, creating and maintaining a balance between artificial intelligence and human capabilities in nursing education seems essential.

While the features and capabilities of these chatbots can potentially revolutionize nursing education, if not used properly, they can be a double-edged sword! Certainly, the complete replacement of artificial intelligence for human intelligence is not possible, at least in nursing. We need nurses in clinical settings who, in addition to practical skills, possess clinical reasoning abilities, analytical power, and high problem-solving skills. Artificial intelligence technologies can threaten this and be perceived as a death knell for traditional learning. Therefore, preventing their potentially dangerous threats in nursing education is recommended, especially considering regulations and guidelines before their use becomes as uncontrollable as the COVID-19 pandemic. The presence of unknown issues surrounding artificial intelligence, like the dark side of the moon, emphasizes the need for further studies on artificial intelligence threats (beyond the dark side of artificial intelligence) concurrently with the development of existing knowledge about its capabilities.

Author contributions

ASh: Investigation, Writing – original draft, Writing – review & editing. ASa: Conceptualization, Writing – original draft, Writing – review & editing. MG: Conceptualization, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Tam W, Huynh T, Tang A, Luong S, Khatri Y, Zhou W. Nursing education in the age of artificial intelligence powered Chatbots (AI-Chatbots): are we ready yet? *Nurse Educ Today.* (2023) 129:105917. doi: 10.1016/j.nedt.2023.105917

2. Buchanan C, Howitt ML, Wilson R, Booth RG, Risling T, Bamford M. Predicted influences of artificial intelligence on nursing education: scoping review. *JMIR Nurs.* (2021) 4:e23933. doi: 10.2196/23933

3. DeLaune SC, Ladner PK, McTier L, Tollefson J. *Fundamentals of Nursing.* Southbank, VIC: Cengage AU (2023).

4. Shahsavari H, Salsali M, Mohammadpour A. Nursing as an Art. *Hayat.* (2010) 16:23—33.

5. Abujaber AA, Abd-Alrazaq A, Al-Qudimat AR, Nashwan AJ, AbuJaber A. A strengths, weaknesses, opportunities, and threats (SWOT) analysis of ChatGPT integration in nursing education: a narrative review. *Cureus.* (2023) 15:e48643. doi: 10.7759/cureus.48643

6. Gugliandolo M, Costa S, Kuss D, Cuzzocrea F, Verrastro V. Technological addiction in adolescents: the interplay between parenting and psychological basic needs. *Int J Ment Health Addict.* (2020) 18:1389–402. doi: 10.1007/s11469-019-00156-4

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 16 October 2024
DOI 10.3389/fmed.2024.1460553

Check for updates

OPEN ACCESS

EDITED BY
Rafat Damseh,
United Arab Emirates University,
United Arab Emirates

REVIEWED BY
Nataly Martini,
The University of Auckland, New Zealand
Omar El-Gayar,
Dakota State University, United States

*CORRESPONDENCE
Hongxing Kan
ffdkhx@ahtcm.edu.cn
Yingbao Geng
gengyingbao@ahtcm.edu.cn

RECEIVED 06 July 2024
ACCEPTED 03 October 2024
PUBLISHED 16 October 2024

CITATION
Liu W, Kan H, Jiang Y, Geng Y, Nie Y and Yang M (2024) MED-ChatGPT CoPilot: a ChatGPT medical assistant for case mining and adjunctive therapy.
Front. Med. 11:1460553.
doi: 10.3389/fmed.2024.1460553

MED-ChatGPT CoPilot: a ChatGPT medical assistant for case mining and adjunctive therapy

Wei Liu[1,2], Hongxing Kan[1,2]*, Yanfei Jiang[1], Yingbao Geng[1]*, Yiqi Nie[1,2] and Mingguang Yang[1]

[1]School of Medical Information Engineering, Anhui University of Traditional Chinese Medicine, Hefei, Anhui, China, [2]Anhui Computer Application Research Institute of Chinese Medicine, China Academy of Chinese Medical Sciences, Hefei, Anhui, China

Background: The large-scale language model, GPT-4-1106-preview, supports text of up to 128 k characters, which has enhanced the capability of processing vast quantities of text. This model can perform efficient and accurate text data mining without the need for retraining, aided by prompt engineering.

Method: The research approach includes prompt engineering and text vectorization processing. In this study, prompt engineering is applied to assist ChatGPT in text mining. Subsequently, the mined results are vectorized and incorporated into a local knowledge base. After cleansing 306 medical papers, data extraction was performed using ChatGPT. Following a validation and filtering process, 241 medical case data entries were obtained, leading to the construction of a local medical knowledge base. Additionally, drawing upon the Langchain framework and utilizing the local knowledge base in conjunction with ChatGPT, we successfully developed a fast and reliable chatbot. This chatbot is capable of providing recommended diagnostic and treatment information for various diseases.

Results: The performance of the designed ChatGPT model, which was enhanced by data from the local knowledge base, exceeded that of the original model by 7.90% on a set of medical questions.

Conclusion: ChatGPT, assisted by prompt engineering, demonstrates effective data mining capabilities for large-scale medical texts. In the future, we plan to incorporate a richer array of medical case data, expand the scale of the knowledge base, and enhance ChatGPT's performance in the medical field.

KEYWORDS

ChatGPT, large language model, data mining, prompt engineering, local knowledge base

1 Introduction

In November 2022, OpenAI released ChatGPT, a GPT-3.5 Turbo-powered intelligent chatbot, marking a significant advancement in knowledge management and artificial intelligence content creation (1). This technology has significantly enhanced the ability to comprehend and generate text, assisting researchers and medical experts in writing literature reviews and abstracts, as well as suggesting structures and references, and it can even be used to write draft papers (2). Leveraging its efficient semantic comprehension, language generation, and logical reasoning capabilities, ChatGPT has been widely applied in various fields such as media, education, healthcare, customer service, law, and data processing, garnering considerable attention (3, 4).

In the medical field, various studies (5–8) have explored the potential of ChatGPT, indicating that it can alleviate the burden on physicians without disrupting existing workflows. These studies emphasize the role of ChatGPT in enhancing the humane aspect of caregiving, including empathy and sensitivity to patients' emotional needs, which current large language models (LLMs) struggle to fully achieve. Faced with the unequal distribution of global medical resources, where developed countries have access to advanced medical technology and well-trained healthcare professionals while developing countries often struggle with limited medical infrastructure and insufficient healthcare providers (9, 10), it becomes necessary to develop tools that can efficiently address common health issues. LLMs possess the capability to utilize knowledge from various disciplines to efficiently handle labor-intensive and time-consuming tasks, such as literature search, medical entity recognition, and data analysis (11–13). Recent studies have shown that trained models are capable of reformulating texts to convey empathy more effectively, which can enhance mental health treatments (14). For example, these models can rephrase clinical advice or supportive messages to make them more comforting and understanding for patients. Additionally, LLMs have demonstrated potential in various medical fields such as radiology, ophthalmology, medical insurance, and urology (15–18).

Despite this, extracting medical information from the plethora of literature and storing it in a local knowledge base remains a challenge. Traditional methods are labor-intensive, weak in generalization capabilities, and often require familiarity with specific software tools or computer programming skills (19). Additionally, the high complexity of medical information (20, 21) and issues such as bottlenecks in data annotation (22) represent significant challenges that cannot be overlooked. However, with the emergence of high-performance, LLMs that support up to 128 k context, such as GPT-4-1106-preview (23), this process is expected to be revolutionized. As a data processing assistant, ChatGPT can work collaboratively with human researchers to advance text mining and data analysis (24, 25). ChatGPT's capabilities stem from its vast pre-trained text corpus, which is a large collection of diverse texts used during the model's training (26). This enables ChatGPT to naturally excel in language comprehension and named-entity recognition, identifying professional terms such as disease and drug names without additional training (27). Moreover, ChatGPT is adept at identifying and associating abbreviations with their full forms in medical data mining, such as "RA" (Rheumatoid Arthritis), "LFTs" (Liver Function Tests), and "PTT" (Partial Thromboplastin Time). This ability is crucial for reducing the quantity of duplicate "unique entities" in datasets that arise due to the use of various abbreviations, helping to avoid redundant data without new information. In contrast, traditional natural language processing methods often fail to recognize abbreviations and full names without a manually compiled medical abbreviation dictionary (28).

With the continuous advancement of artificial intelligence technology, AI has played a key role in various subfields of medicine, such as pancreatic cancer, liver cancer, digestive system diseases, and retinal pathologies (29–32). Deep learning and advanced algorithms aid physicians in more accurate diagnosis and prediction of diseases (33, 34). Research has also explored the use of AI in radiopathomics (the integration of radiology and pathology data to improve diagnostic accuracy) and diabetology (35, 36).

Although LLMs such as ChatGPT have made significant advancements in the medical field, enhancing the efficiency and accuracy of healthcare services, current research often focuses on the intelligent diagnosis of specific diseases. This leaves a gap in the development of comprehensive intelligent systems that are broadly applicable to multiple conditions. Additionally, the complexity, diversity, and challenges associated with data annotation in medical information processing remain significant hurdles. Studies indicate that well-designed prompts and contextual settings can substantially reduce the likelihood of ChatGPT generating erroneous information (24, 37). This insight has guided our approach in designing text data mining strategies, ensuring maximum efficiency and accuracy in GPT outputs. To address these issues, our research aims to utilize the latest ChatGPT model, which supports extensive context, to deeply mine medical information and assist in data annotation. Subsequently, we will construct a medical knowledge base derived from the deeply mined data, enhancing the application performance of LLMs in the medical domain. Finally, we propose a system named "MED-ChatGPT CoPilot," which integrates case mining with auxiliary treatment suggestions. This system is designed to provide medical professionals with an efficient and rapid method for case mining and data annotation, while also serving as a convenient medical knowledge advisor for patients.

2 Methods and experimental design

The workflow of this study is divided into five steps: (1) Data preprocessing; (2) Design of text mining using ChatGPT; (3) Building a local knowledge base and calculating similarity vectors; (4) Developing an auxiliary inquiry system based on the ChatGPT API; and (5) Utilizing ChatGPT to write script codes for assistance with tasks (Figure 1).

2.1 Data preprocessing

In the data preprocessing stage of this study, special attention was given to the security and reliability of the selected medical data. We conducted an extensive literature search using academic literature platforms such as ScienceDirect, Web of Science, and PubMed, with keywords including "clinical guidelines," "treatment guidelines," and "treatment strategies." Building upon this, we established a set of stringent screening criteria. Initially, we excluded all non-English literature, case reports, reviews, forum articles, brief communications, and expert opinion pieces, as well as studies that were not peer-reviewed or had a sample size of less than 100, to ensure the scientific

Abbreviations: GPT-4-1106-preview, Generative Pre-trained Transformer 4-1106-preview version; API, Application Programming Interface; LLMs, Large Language Models; RA, Rheumatoid Arthritis; LFTs, Liver Function Tests; PTT, Partial Thromboplastin Time; AI, Artificial Intelligence; RESTful, Representational State Transfer; gRPC, gRPC Remote Procedure Calls; HERs, Electronic Health Records; GDPR, General Data Protection Regulation.

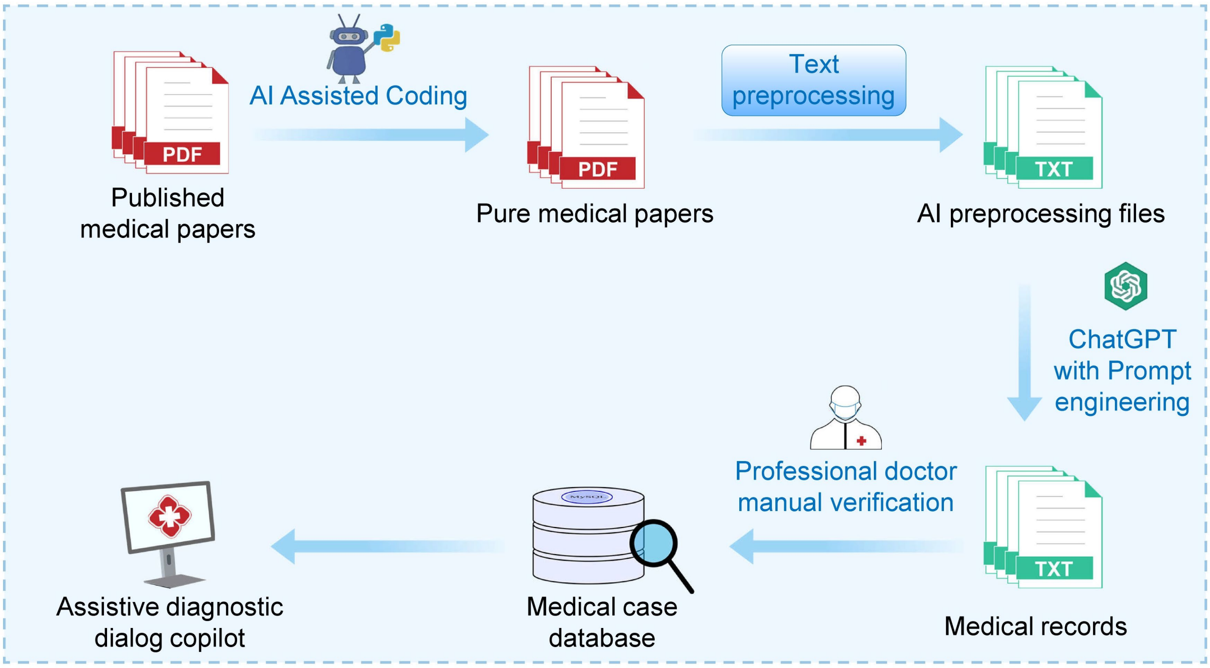


FIGURE 1
Schematic diagram of the MED-ChatGPT CoPilot workflow. The diagram illustrates a MED-ChatGPT CoPilot-assisted workflow that begins with "Published medical papers" being preprocessed using a Python script generated with ChatGPT's assistance. The preprocessing step removes extraneous information, such as references, acknowledgments, and other non-essential sections, resulting in "Pure medical papers." These pure medical papers focus solely on relevant medical content and are then converted into text files to ensure compatibility and ease of processing by ChatGPT. Subsequently, through collaborative efforts between ChatGPT and Prompt Engineering, "Medical records" are generated. These records undergo professional medical review, and 241 cases are vectorized and indexed to construct the "Medical Case Database," a local repository of medical knowledge. Finally, the integration of this local vector knowledge database with ChatGPT fulfills the objective of this study: to create a medical assistant for case mining and aiding diagnosis and treatment. It is important to note that the fifth step of the workflow involves using ChatGPT to compose script code to assist with tasks, a process which is interwoven throughout the entire workflow and elaborated on in Section 2.5. The reason for converting the papers into text files is that ChatGPT processes text more efficiently and accurately than PDFs. Text files are easier to parse and manipulate programmatically, ensuring that the relevant medical information is seamlessly integrated into the workflow.

quality and statistical power of the selected papers. Furthermore, we eliminated studies that lacked a clear mechanism of disease treatment or verification of treatment efficacy. The literature that passed this meticulous screening process needed to provide detailed descriptions of treatment strategies for specific diseases, be based on large clinical trials or multicenter studies and have clear data collection and analysis methodologies, which ensures both the quality of data and the broad applicability of the research.

To further ensure the high quality of the selected studies, we conducted an additional quality assessment by evaluating the average citation count of each paper. Only those papers with an average citation count of five or more were included, indicating that these studies have been recognized and validated by the scientific community. After this two-round screening process, we ultimately selected 306 high-quality medical research papers published within the past 5 years that covered various disease domains. Through this rigorous selection process, we ensured the scientific validity and practical utility of the chosen papers, providing a solid data foundation for this study.

To address the issue of text length limitations inherent in LLMs for text processing, we have devised an innovative technical approach. This approach involves filtering paragraphs directly related to disease diagnosis and treatment from the complexities of the literature. We utilize meticulously crafted regular expressions to identify and automatically exclude non-essential information, such as reference citations and acknowledgments, based on their formatting patterns. To ensure the accuracy and relevance of the filtering process, a professional doctor reviewed the criteria and results to confirm that only non-essential information was removed. This step was crucial to maintain the integrity and context of the medical information in the filtered data. Additionally, the detailed process and specific implementation methods have been compiled in the supplementary materials, available in the Data Availability section along with the corresponding code and datasets. Considering that excessive text length can impact the efficiency and effectiveness of the model (38), we have eliminated overly verbose and medically irrelevant text segments post-filtering. By retaining only the essential medical content, we aim to enhance the model's performance without compromising data reliability. Through this series of preprocessing steps, we have ensured that each selected medical literature includes at least one disease and its corresponding treatment method and that the token size and format comply with the input requirements of the ChatGPT-4-1106-preview model, namely, keeping the context length under 128 thousand tokens.

Considering both the context length limitations of ChatGPT and the need to maintain the text's quality and relevance, we adopted a balanced approach to preprocessing. While we acknowledge that extensive preprocessing could potentially limit the generalizability of our findings, we also recognize the critical need for ChatGPT to process information that is accurate and dense, enabling more effective case mining and support for diagnosis and treatment. Thus, our preprocessing was carefully calibrated to eliminate clearly non-essential content while preserving the vital details essential for understanding disease diagnostics and therapeutic strategies. By doing so, we achieved a compromise between the need for preprocessing and preserving the integrity and wide applicability of our study to diverse medical literature.

2.2 Text data mining design based on ChatGPT

Through repeated experimentation and fine-tuning, we determined the optimal prompt design and background setting (also known as task description) and appended cues after each question to enhance the efficiency and accuracy of GPT outputs and ensure the content's standardization. Within the prompts, we required the large model to divide each disease-related information into six parts: (1) Disease name; (2) Clinical manifestations; (3) Recommended treatment protocols; (4) Recommended medications; (5) Precautions and side effects; and (6) Additional recommendations (such as lifestyle advice). The workflow diagram for this study's ChatGPT-assisted medical text mining task using prompt engineering is presented in Figure 2. Specific prompts, background information, as well as input and output examples, can be found in Table 1.

Moreover, OpenAI's API allows researchers to precisely control key parameters of the GPT model, including "temperature" and "history." The "temperature" parameter dictates the degree of conservatism or innovation in the model's output, with values ranging from 0 (very conservative) to 1 (highly innovative). In this study, we set eleven different "temperature" values at intervals of 0.1, from 0 to 1, to test their effect on data mining efficacy. To ensure consistent experimental conditions, we conducted three independent rounds of ChatGPT tests for each "temperature" setting and recorded the outputs. Moreover, to eliminate the influence of "history" on the model's outputs, we reset ChatGPT's "history" parameter after each interaction, ensuring that each invocation started from a state with no memory, to avoid any impact from previous interactions on subsequent outputs. By comparing model performance under various settings, we found that the GPT-4-1106-preview model performed best with the "temperature" set to 0.1. Hence, we selected this parameter configuration as the standard setup for extracting medical text information.

2.3 Building a local knowledge base and vectorized similarity computation

Since the advent of ChatGPT, Large LLMs have been warmly received globally, and tool frameworks like LangChain (39) have been developed. LangChain is a framework designed specifically for developing applications based on LLMs, offering developers a range of modules and tools to simplify the integration process with LLMs. Using LangChain, developers can easily leverage language models to perform complex tasks, including text-to-image conversion, document question answering, and chatbot construction. In this research, we adopt the LangChain framework to deeply integrate pre-trained language models with a local knowledge base. Using "disease names" and "clinical manifestations" from medical case information to build an index, we then employ the text2vec-large-Chinese vectorization model (40) to vectorize the medical case data. This model not only demonstrates excellent vectorization capabilities for Chinese and English texts but is also fully open-source and free, aligning with our future plans to incorporate Chinese medical cases into the database.

While exploring different retrieval enhancement methods, we also considered other forms, such as the direct embedding of source files (41). However, we opted for vectorized similarity calculation and the construction of a local knowledge base because this approach provides different benefits for researchers and patients. For researchers, it offers a scalable and interpretable representation of medical case information, enabling complex semantic retrieval and deeper insights,

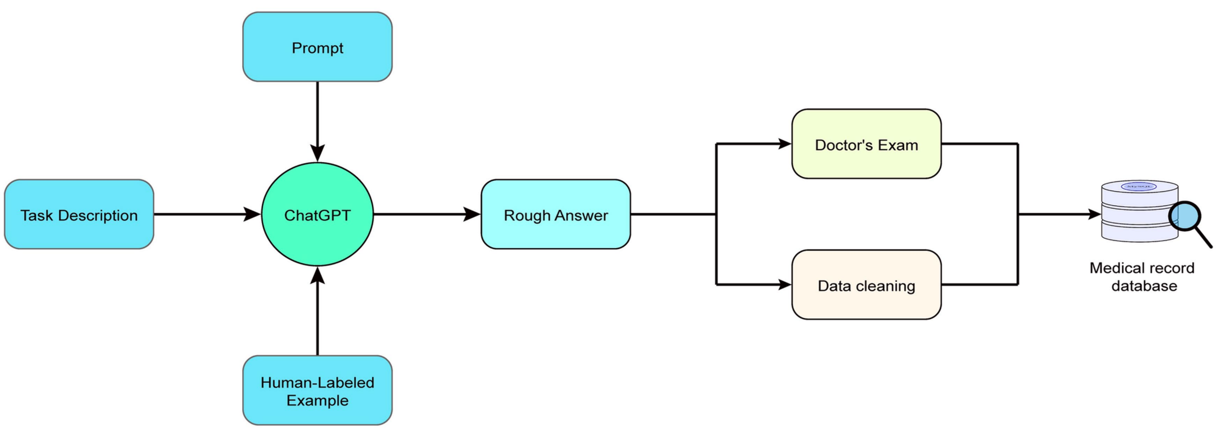


FIGURE 2
An overview of the workflow for medical text mining with ChatGPT assisted by prompt engineering. After the large model extracts preliminary data, the data undergo cleansing and verification by professional physicians before proceeding to the next step of building a local knowledge base, which will be detailed in Section 2.3.

TABLE 1 Detailed prompt and example documentation for study.

Task description	"context": "You are a medical assistant specializing in answering common disease questions, answering questions based on the papers you have read. Use the provided context to answer the question. If the question is irrelevant in the context, or the context does not mention anything related to the disease, answer 'Based on the information in the medical literature I have read so far, I cannot provide a reliable answer to this question. Please modify your question."
Prompt	Please extract the following information from the provided text: Disease name, Clinical manifestations, Recommended treatment plan, Recommended medication, Precautions and side effects, Other suggestions (such as lifestyle advice). If the text contains information regarding other diseases, please also present it in the required format. If certain attributes of the disease are not mentioned, then there is no need to provide me with information about that disease. Then, refine the extracted information in the following format: Disease name:\n Clinical manifestations:\n Recommended treatment plan:\n Recommended medication:\n Precautions and side effects:\n Other suggestions:\n Please note that the recommended medication and its corresponding precautions and side effects should be matched correctly。If the requested information is not extracted, let me know so. Below is the text context to be extracted:
Input	"158. INFECTIONS DUE TO CAMPYLOBACTER AND RELATED SPECIES – Martin J. Blaser DEFINITION Bacteria of the genus Campylobacter and of the related genera Arcobacter and Helicobacter (Chap. 154) cause a variety of pyogenic infections. Although acute diarrheal illnesses are most common, these organisms may cause infections in virtually all parts of the body, especially in compromised hosts, and these infections may have late nonsuppurative sequelae. The designation Campylobacter comes from the Greek for "curved rod" and refers to the organism's vibrio-like morphology." … omission of the following text, which in total contains 2,276 words.
Output	Disease name: Campylobacter and related species infections Clinical manifestations: Diarrhea, abdominal pain, fever, Diarrheal illness may be grossly bloody, Cramping, pseudoappendicitis, Convulsions in febrile young children, Intermittent diarrhea or nonspecific abdominal pain without localizing signs (*C. fetus*),Prolonged relapsing systemic illness with fever, chills, and myalgias (*C. fetus*),Recurrent cellulitis with fever and bacteremia in immunocompromised hosts (*H. cinaedi*) Recommended treatment plan: Fluid and electrolyte replacement for diarrheal illnesses, Antimicrobial therapy for high fever, bloody diarrhea, severe diarrhea, persistence for more than 1 week, and worsening of symptoms, For systemic infections, empirical treatment with gentamicin, imipenem, or chloramphenicol, followed by susceptibility testing, Prolonged therapy for immunocompromised patients or endovascular infections Recommended medication: Erythromycin (250 mg orally four times daily or 30 to 50 mg/kg per day for children),Clarithromycin and azithromycin as alternatives, Ciprofloxacin (500 mg orally twice daily) or another fluoroquinolone for adults, Tetracycline and furazolidone as other alternatives, Gentamicin, imipenem, or chloramphenicol for systemic infections, Ciprofloxacin and amoxicillin/clavulanate for susceptible strains Precautions and side effects: Resistance to fluoroquinolones is increasing, Antimotility agents may prolong symptoms and are not recommended, Side effects specific to each medication should be considered and monitored Other suggestions: Cook poultry sufficiently to prevent infection, Avoid raw (unpasteurized) milk or untreated water, Practice good hygiene when in contact with household pets, Seek medical attention if symptoms persist or worsen

This table documents the specific prompts, background information, and examples of input and output used in detail in this study.

which surpass simple text matching. On the other hand, for patients, our method emphasizes an intuitive and easy-to-manage user interface that makes medical information more accessible and understandable. Compared to file embedding, our approach focuses on the specific needs of each user group, ensuring that researchers can access detailed and nuanced data while patients receive clear and straightforward information. Furthermore, given the sensitivity and complexity of medical information, we believe that providing a clear, interpretable knowledge representation is more important than relying solely on automated, potentially difficult-to-explain embedding vectors.

After vectorizing the medical case data, the storage and retrieval of vector content rely on the Qdrant vector search database (42). Vector databases are an emerging means of data interaction that combine with abstract data representations produced by machine learning models, such as deep learning. Vector databases exhibit exceptional performance in applications such as semantic search and recommendation systems (43). Qdrant is an open-source vector database designed for the new generation of AI applications, using a cloud-native architecture and offering RESTful and gRPC APIs for embedding vector management. It supports search functions for images, voice, and video and can be integrated with AI engines, enhancing the breadth and depth of its applications. Additionally, Qdrant uses the Cosine Similarity algorithm to improve retrieval accuracy, the formula of which is as follows:

$$Similarity\left(A,B\right)=\frac{A\cdot B}{A\times B}=\frac{\sum_{i+1}^{n}\left(A_i\times B_i\right)}{\sqrt{\sum_{i+1}^{n}A_i^2}\times\sqrt{\sum_{i=1}^{n}B_i^2}}\#$$

2.4 Establishing a ChatGPT-based auxiliary diagnostic chatbot

After completing the vectorization embedding of medical case texts and constructing a corresponding local medical case vector database, this research utilized Python's Streamlit package to build a front-end interface, enabling auxiliary diagnostic dialogues with users. When a user initiates a query, the system first processes the content of the query into a vectorized form. By calculating the cosine similarity between the user's query vector and each vector in the medical case vector database, the system can precisely retrieve the text units that best match the user's needs. After the optimal matching text unit is selected, it is applied in conjunction with the prompt engineering to the ChatGPT model. Benefiting from ChatGPT's powerful capabilities in text generation and logical reasoning, this method combines the data-driven characteristics of the vector database, effectively enhancing the accuracy of the consultation and significantly reducing the tendency for "hallucination" phenomena that large-scale language models are prone to (44, 45). Ultimately, the system presents the processed results on the user interface. A schematic of the integrated workflow for the medical dialogue robot is presented in Figure 3.

To ensure the safety and accuracy of medical advice, this system's chatbot processes queries solely from the database, thereby reducing the production of misleading information. Furthermore, the chatbot is capable of "remembering" previous conversations, including recognized contexts and pertinent medical case information, to maintain the coherence and data-driven nature of the responses.

2.5 ChatGPT-assisted script code generation work

In this study, we have fully leveraged the automation capabilities of ChatGPT to generate Python scripts that are used for parsing medical literature, generating prompts, processing text, and conducting data mining, with the results being output in a predefined format. Traditionally, this process required complex programming skills and a significant investment of time; however, with the efficient code generation capabilities of ChatGPT, this process has been significantly simplified and accelerated. For instance, in less than a second, the GPT model is capable of generating a data preprocessing script, which can automatically convert PDF-formatted papers into clean medical text files devoid of irrelevant information, markedly reducing the workload and time costs. Researchers simply need to describe their requirements and the desired output format in natural language, and ChatGPT rapidly generates the corresponding Python code. This code can be directly copied, pasted, modified, and executed, greatly enhancing research efficiency. Should any errors arise during the execution of the code, this large-scale model also provides instant assistance to help researchers quickly identify and correct these errors.

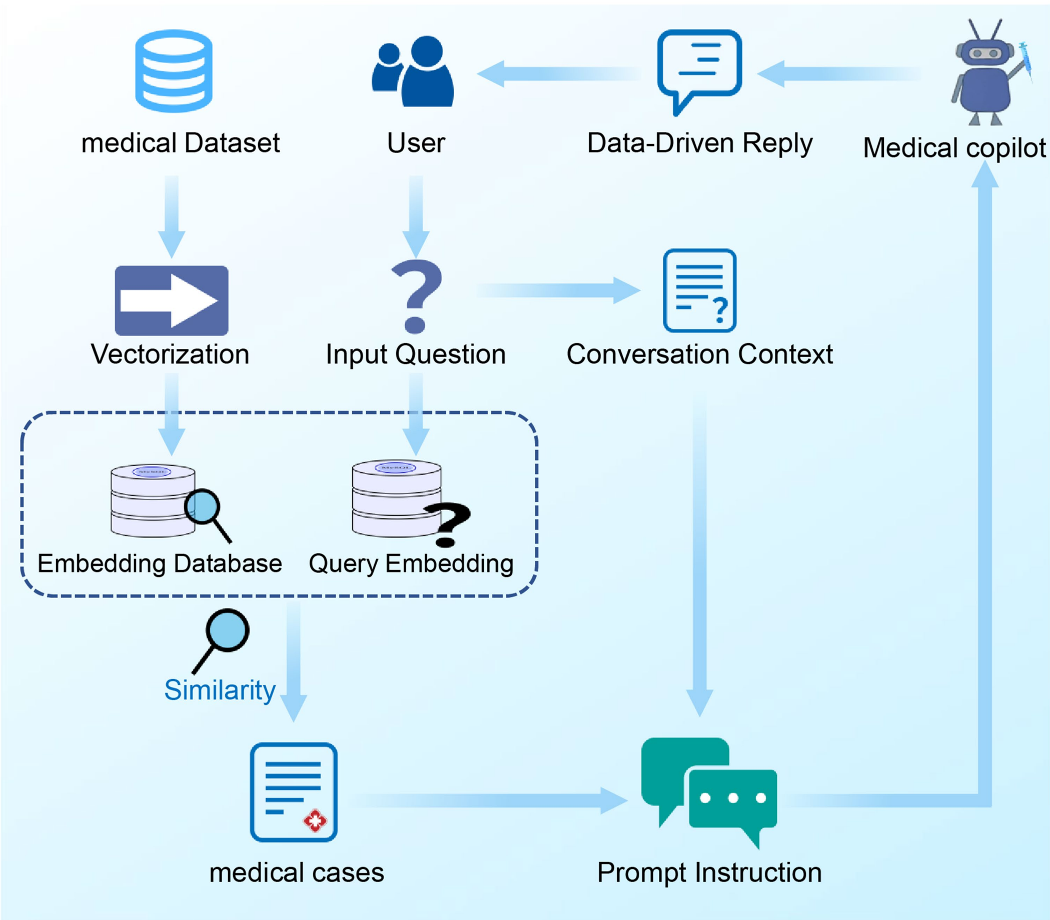


FIGURE 3
The integrated workflow of the medical chatbot. The integrated workflow of the medical chatbot. This workflow begins with the user initiating a query, which is then vectorized using a pre-trained language model. The system calculates the cosine similarity between the user's query vector and vectors in the local medical case database. The top three matching medical cases are retrieved, and further matched with the user's dialogue context to select the most relevant case. This selected case, combined with prompt engineering, is then input into the ChatGPT model. The ChatGPT model generates a response that is presented to the user through the interface.

We have thoroughly validated and tested the generated scripts to ensure their accuracy and reliability in parsing and processing medical texts. Furthermore, to promote reusability, transparency, and assessment of this study, these scripts (including pdf_to_pool.py, remove_ref.py, txt_to_csv.py, etc.) have been shared in the supplementary materials. This section has been expressly added to highlight the possibility that even non-programmers, with the assistance of GPT, can complete and fine-tune automated workflows.

2.6 Evaluation of experimental design

In order to investigate the impact of the MED-ChatGPT medical assistant on the time efficiency of medical case mining and its quality of responding to a specific set of medical questions after being augmented with a medical knowledge base, this study designed a series of evaluative experiments.

To assess the efficiency of ChatGPT in mining medical texts, we devised a repeated comparative validation experiment. Initially, we compiled a pool of 306 high-quality medical research papers published within the last 5 years. From this pool, we randomly selected 100 papers to ensure a representative sample for our study. The random selection was performed using Python's `random` module, specifically the `sample` function, which ensures each paper has an equal probability of being chosen. For each experimental round, 20 papers were chosen for comparative validation, with a total of five rounds conducted. In each round, we maintained constant other variables (such as API interface parameters, prompts, and temperature) and assigned both the ChatGPT-4-1106-preview model and three Chinese medical professionals with master's degrees and over 5 years of clinical experience to process these papers. We recorded the time taken by each to complete the tasks in order to calculate the average time. To account for potential fatigue phenomena human participants might experience when processing large volumes of text, we spaced the experiment over 5 days, with 1 day between each round.

Subsequently, to evaluate the performance of MED-ChatGPT CoPilot in medical question-answering tasks, we designed a series of experiments aimed at comparing the question-answering quality of the ChatGPT model, enhanced with a medical knowledge base, against the original baseline model when addressing a specific set of medical questions. The experiments utilized the MultiMedQA medical benchmark dataset (46), which comprises questions that include multiple-choice and long-form responses. From its subset, the MedQA dataset (47), we randomly selected 300 questions as our test samples, ensuring the generalizability and randomness of the test results. The reason for using only a subset of the MedQA dataset, rather than the complete benchmark dataset, is that these questions all followed the style guidelines of the United States Medical Licensing Examination, consisting solely of single-choice questions. This approach was chosen to reduce subjectivity in the evaluation and ensure the definitiveness of the answers. We compared the performance of the original baseline model (ChatGPT-4-1106-preview) with that of the model enhanced with a medical knowledge base, using the same set of questions for both models, and recorded all question-answer pairs (see Supplementary Data). The evaluation criterion was based on the proportion of questions correctly answered by the models relative to the total number of test questions, thereby measuring and comparing the performance of the two models.

3 Results

3.1 Text mining time performance evaluation

In the domain of medical text mining efficiency, experimental results (Figure 4) demonstrate that the efficiency of information extraction from medical cases by large-scale models significantly surpasses that of a human with substantial medical knowledge. In the comparative study of medical text mining efficiency, we calculated the mean, standard deviation, and 95% confidence interval for the time taken by the GPT model and a human with a master's degree in medicine to process literature. The GPT model requires an average of only 11.89 s per document (with a standard deviation of 1.76 s and a 95% confidence interval ranging from 10.96 to 12.82 s), totaling less than 4 min to process 20 documents. In contrast, the average time taken by the medical master's degree holder to process the same number of documents ranges from 157.8 to 379.2 s per document (with standard deviations of 21.58 s and 46.28 s, and 95% confidence intervals of 150.34–165.26 s and 359.14–399.26 s, respectively), and this duration already takes into account the assistance of English translation software. This further highlights the advantages of GPT in multilingual processing.

3.2 Evaluation of the ChatGPT model enhanced with a local knowledge base for medical knowledge performance

In the assessment of medical knowledge performance, this study compared two models utilizing identical internal interface parameters:

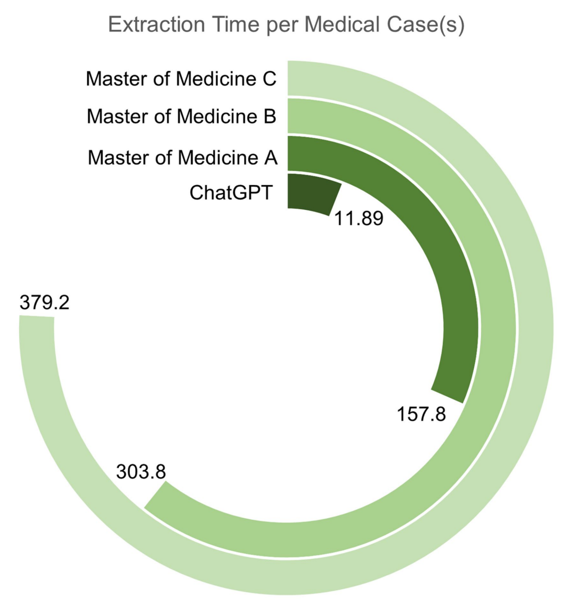


FIGURE 4
The timeline comparison chart for medical text information mining in this study. It records the average time required for three medical masters and the medical assistant from this study to extract information from the same medical case, clearly demonstrating their relative efficiency in completing this task.

the original ChatGPT-4-1106-preview base model and MED-ChatGPT CoPilot, the model developed in this study augmented with a medical knowledge base. The evaluation results are presented in Table 2.

After analyzing 300 randomly selected sample questions, the initial ChatGPT-4-1106-preview baseline model exhibited an accuracy rate of 71.67%, whereas MED-ChatGPT CoPilot demonstrated an accuracy rate of 77.33%. This comparative result clearly indicates that, on a specific set of medical questions, MED-ChatGPT CoPilot surpassed the original baseline model, with an increased accuracy of 7.90%.

However, it must be noted that the performance improvement is, to some extent, constrained by the current scale of the knowledge base. Given that the knowledge base presently contains only 241 medical case entries, the observed enhancement in performance is already significant. It is reasonable to anticipate that with further expansion of the knowledge base, the model's performance could see even greater improvements.

3.3 Consulting diagnostic and therapeutic recommendations and medication suggestions through an auxiliary robotic system

The developed system maximizes interactivity and user experience by converting database information into a conversational format, as detailed in Figure 3. This medical dialogue bot facilitates the acceleration of research findings to clinical application. Built upon ChatGPT's medical assistant architecture, it supports medical case mining and diagnostic support. All conversational data are derived from detailed analyses of medical papers, ensuring safety and reliability. Figure 5 presents a schematic representation of the MED-ChatGPT copilot interacting with a user for medical consultation. In this dialogue display, the user can inquire about detailed information regarding a disease, encompassing key information such as clinical manifestations and medication recommendations.

In summary, the results indicate that the MED-ChatGPT CoPilot model, enhanced with a specifically curated local medical knowledge base, shows notable improvements in medical text mining and diagnostic assistance. This model provides high-quality, peer-reviewed medical recommendations, ensuring higher relevance and accuracy in the medical field. The MED-ChatGPT CoPilot-supported workflow excels in mining time efficiency and annotation efficiency, significantly simplifying medical research, case mining, and text annotation tasks.

4 Discussion

4.1 Background of existing research

Medical text mining aims to extract valuable information from complex medical data to aid in diagnosis, treatment, and disease prediction. Previous studies have applied various machine learning algorithms, including k-nearest neighbors, decision trees, logistic regression, naive Bayes, and support vector machines, to this task (48–50). However, these efforts face significant challenges. International variations in medical information and the scarcity of annotated databases hinder the effectiveness of medical data mining.

Traditional methods rely heavily on keyword matching, which often leads to suboptimal outcomes. The advent of Electronic Health Records (EHRs) has improved data standardization, yet it introduces new concerns regarding patient privacy and the legal acquisition of EHR data (51, 52). Despite advancements, these systems still grapple with non-standardized terminology and fragmented information distribution (53).

Additionally, the COVID-19 pandemic has sparked interest in integrating artificial intelligence into medical assistance systems. Numerous projects have attempted to incorporate AI into disease diagnosis and treatment (54–57), but they mostly still rely on traditional methods, such as keyword matching approaches. Recent innovations include multimodal approaches, such as combining neuroimaging and voice analysis to diagnose Parkinson's disease (58), highlighting the potential for more complex systems. However, these advancements are still in progress, and there is an urgent need for more comprehensive, accurate, and personalized AI-driven medical tools.

4.2 Key findings and innovations

Our study demonstrates that ChatGPT can significantly enhance efficiency and accuracy in medical text mining and diagnostic support. By integrating a local medical knowledge base with vectorized similarity computations, we improved the precision of retrieving relevant medical cases and ensured user-friendly data presentation. The MED-ChatGPT CoPilot model, combining ChatGPT with curated medical data, notably increased accuracy in medical question-answering tasks from 71.67 to 77.33%. Additionally, the use of ChatGPT for automated script generation streamlined the research process, making advanced medical text processing more accessible. These innovations collectively advance the field of medical text mining and diagnostic assistance.

4.3 Identified limitations and challenges

Despite the significant advantages demonstrated by the MED-ChatGPT CoPilot model in medical text mining and diagnostic assistance, several limitations and challenges were identified in this study. One major challenge is the inherent dependency on the quality and comprehensiveness of the local medical knowledge base. Although the curated database of 241 medical cases provided a solid foundation for enhancing the model's accuracy, its relatively limited scope means it may not cover all possible medical conditions or the

TABLE 2 Performance statistics of MED-ChatGPT CoPilot versus the original model on a medical knowledge question set.

Model name	Correct answers	Incorrect answers	Total accuracy (%)	Improvement ratio (%)
ChatGPT-4-1106-preview	215	85	71.67	7.90
MED-ChatGPT CoPilot	232	68	77.33	

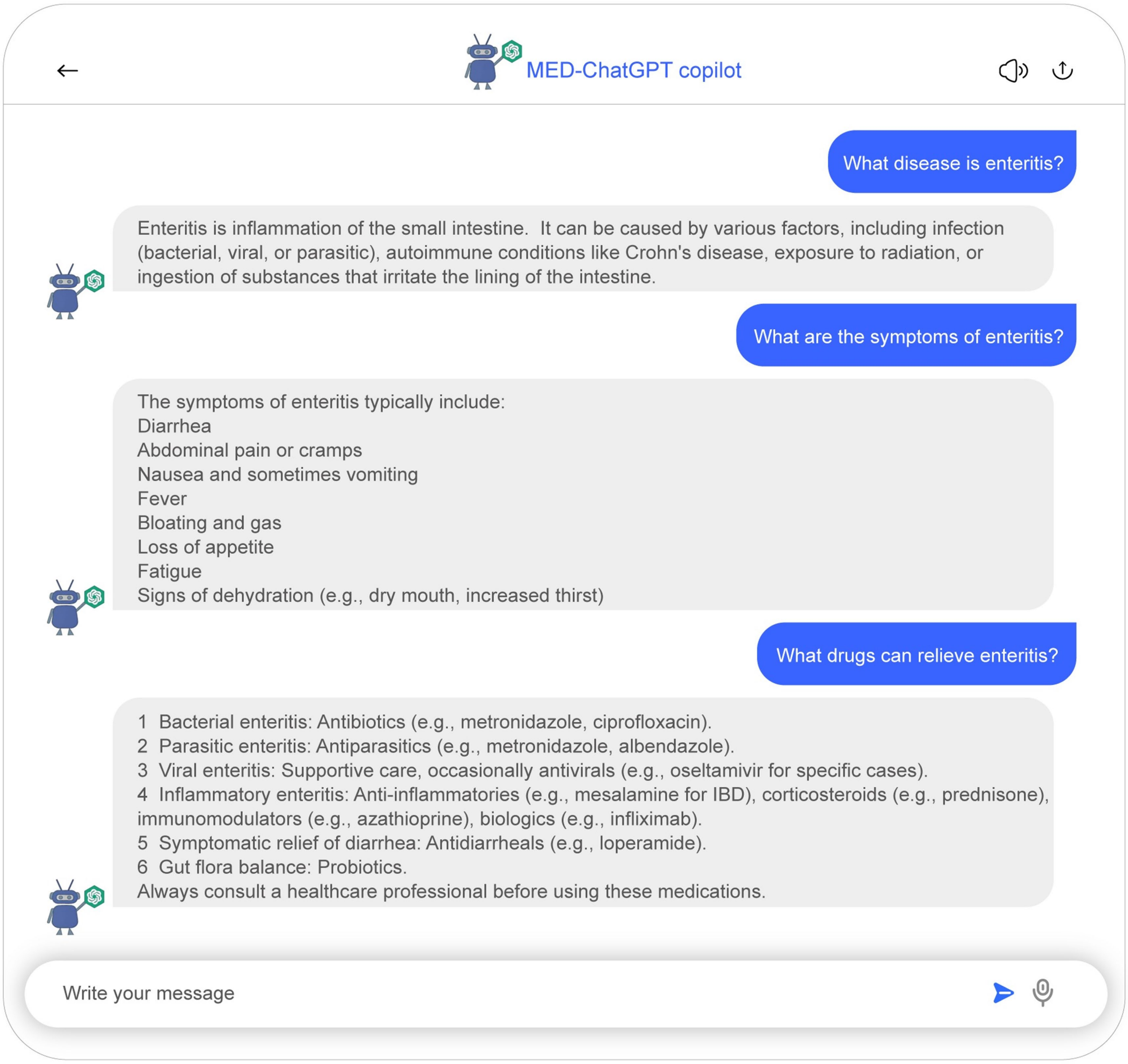


FIGURE 5
Demonstrative sample of the medical chatbot interaction. Users are able to inquire about detailed information regarding diseases, which encompasses an introduction to the disease, clinical manifestations, as well as recommendations for medication.

latest research developments comprehensively. This constraint could lead to gaps in the model's diagnostic capabilities, especially for rare or newly emerging diseases.

Another area warranting careful consideration is the model's ability to provide medical advice. Although MED-ChatGPT CoPilot has shown to outperform standard search engines such as Google by providing more structured and comprehensive diagnostic suggestions, there are instances where the recommendations from the model appear similar to those generated by general search engines. For example, while a Google search for symptoms of enteritis might list potential causes such as bacterial, viral, or parasitic infections, our model goes further by suggesting specific tests based on patient symptoms and history to differentiate these causes. However, this enhancement is sometimes subtle, and the perceived similarity in the output can undermine the perceived value of using our specialized system over a general search engine.

The model's performance enhancement observed in this study is also influenced by the specific configuration and tuning parameters, such as the temperature setting, which were determined through iterative experimentation. While these settings provided optimal results for our dataset, they might not be universally applicable across different medical domains or datasets, necessitating further fine-tuning and validation for broader applicability.

Moreover, while this study did not utilize any personal or private patient data, incorporating detailed individual patient data could further enhance the system's capability in assisting medical diagnoses and treatments. In the future, we aim to integrate legally compliant and secure patient records to enrich the system. Ensuring the reliability and legality of data handling will be paramount throughout this process to safeguard patient privacy and comply with data protection regulations.

4.4 Future research directions and hypotheses

In light of the findings and challenges identified in this study, several future research directions and hypotheses have emerged. One critical area for future work is the expansion and enrichment of the local medical knowledge base. Increasing the number of medical cases and integrating the latest research developments will likely enhance the model's diagnostic capabilities, particularly for rare and emerging diseases. This expansion could be achieved through continuous updates and collaboration with medical institutions to incorporate new clinical data and treatment protocols.

Another promising direction is the incorporation of personalized patient data into the system, which could significantly improve the relevance and accuracy of diagnostic and therapeutic recommendations. However, this approach necessitates stringent measures to ensure data privacy and compliance with legal regulations, such as the General Data Protection Regulation (GDPR). Developing robust data anonymization techniques and secure data handling protocols will be crucial in this endeavor.

The integration of multimodal data sources, such as imaging and genomic data, with text-based information presents another exciting prospect. Combining these diverse data types could provide a more comprehensive understanding of complex diseases, leading to more accurate diagnoses and personalized treatment plans. Exploring advanced machine learning techniques, such as multimodal learning, could facilitate this integration.

By pursuing these research directions, we aim to further refine and expand the capabilities of our MED-ChatGPT CoPilot, ultimately contributing to more effective and personalized medical care.

5 Conclusion

The MED-ChatGPT copilot effectively utilizes prompt engineering techniques and a local knowledge base to construct a high-precision, reliable medical assistant, providing an innovative and efficient solution for medical text mining and adjunctive diagnosis.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Author contributions

WL: Conceptualization, Formal analysis, Investigation, Methodology, Resources, Software, Writing – original draft, Writing – review & editing. HK: Funding acquisition, Project administration, Supervision, Validation, Writing – review & editing. YJ: Methodology, Writing – review & editing. YG: Visualization, Writing – review & editing. YN: Funding acquisition, Writing – review & editing. MY: Data curation, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. We received support from the University Synergy Innovation Program of Anhui Province, project number: GXXT-2023-071. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

I express my deep gratitude to my supervisor and faculty members for their guidance and knowledge. Special thanks to my senior colleague for her invaluable assistance with my research and data organization. Lastly, I thank my family for their unwavering love and encouragement.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. OpenAI. Available at: https://openai.com/ (accessed November 25, 2023).

2. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care*. (2023) 27:75. doi: 10.1186/s13054-023-04380-2

3. Cheng K, Guo Q, He Y, Lu Y, Gu S, Wu H. Exploring the potential of GPT-4 in biomedical engineering: the Dawn of a new era. *Ann Biomed Eng*. (2023) 51:1645–53. doi: 10.1007/s10439-023-03221-1

4. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med*. (2023) 388:1233–9. doi: 10.1056/NEJMsr2214184

5. Hou Y, Yeung J, Xu H, Su C, Wang F, Zhang R. From Answers to Insights: Unveiling the Strengths and Limitations of ChatGPT and Biomedical Knowledge Graphs. *Research Square*. (2023). doi: 10.21203/rs.3.rs-3185632/v1

6. Seth I, Sinkjær Kenney P, Bulloch G, Hunter-Smith DJ, Bo Thomsen J, Rozen WM. Artificial or augmented authorship? A conversation with a Chatbot on base of thumb arthritis. *Plast Reconstr Surg Glob Open*. (2023) 11:e4999. doi: 10.1097/GOX.0000000000004999

7. Kim H-W, Shin D-H, Kim J, Lee G-H, Cho JW. Assessing the performance of ChatGPT's responses to questions related to epilepsy: a cross-sectional study on natural language processing and medical information retrieval. *Seizure*. (2024) 114:1–8. doi: 10.1016/j.seizure.2023.11.013

8. Zhou Q, Liu C, Duan Y, Sun K, Li Y, Kan H, et al. GastroBot: a Chinese gastrointestinal disease chatbot based on the retrieval-augmented generation. *Front Med*. (2024) 11:1392555. doi: 10.3389/fmed.2024.1392555

9. Huang W-T, Chen P-S, Liu JJ, Chen Y-R, Chen Y-H. Dynamic configuration scheduling problem for stochastic medical resources. *J Biomed Inform*. (2018) 80:96–105. doi: 10.1016/j.jbi.2018.03.005

10. Ye Y, Huang L, Wang J, Chuang Y-C, Pan L. Patient allocation method in major epidemics under the situation of hierarchical diagnosis and treatment. *BMC Med Inform Decis Mak*. (2022) 22:331. doi: 10.1186/s12911-022-02074-3

11. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA*. (2023) 330:866–9. doi: 10.1001/jama.2023.14217

12. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data Mining of Free-Text CT reports on lung Cancer. *Radiology*. (2023) 308:e231362. doi: 10.1148/radiol.231362

13. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. (2023) 9:e45312. doi: 10.2196/45312

14. Pande SD, Hasane Ahammad SK, Gurav MN, Faragallah OS, Eid MMA, Rashed ANZ. Depression detection based on social networking sites using data mining. *Multimed Tools Appl*. (2024) 83:25951–67. doi: 10.1007/s11042-023-16564-7

15. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. *Diagn Interv Imaging*. (2023) 104:269–74. doi: 10.1016/j.diii.2023.02.003

16. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence Chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol*. (2023) 141:589–97. doi: 10.1001/jamaophthalmol.2023.1144

17. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. (2023) 11:887. doi: 10.3390/healthcare11060887

18. Eppler M, Ganjavi C, Ramacciotti LS, Piazza P, Rodler S, Checcucci E, et al. Awareness and use of ChatGPT and large language models: a prospective cross-sectional global survey in urology. *Eur Urol*. (2023) 85:146–53. doi: 10.1016/j.eururo.2023.10.014

19. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat Mach Intell*. (2023) 5:46–57. doi: 10.1038/s42256-022-00593-2

20. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. *Biomed Inform Insights*. (2016) 8:BII.S31559–10. doi: 10.4137/BII.S31559

21. Wu W-T, Li Y-J, Feng A-Z, Li L, Huang T, Xu A-D, et al. Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Mil Med Res*. (2021) 8:44. doi: 10.1186/s40779-021-00338-z

22. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. *JMIR Med Inform*. (2020) 8:e17984. doi: 10.2196/17984

23. Introducing GPTs. Available at: https://openai.com/blog/introducing-gpts (accessed November 25, 2023).

24. Zheng Z, Zhang O, Borgs C, Chayes JT, Yaghi OM. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *J Am Chem Soc*. (2023) 145:18048–62. doi: 10.1021/jacs.3c05819

25. Lynch CJ, Jensen EJ, Zamponi V, O'Brien K, Frydenlund E, Gore R. A structured narrative prompt for prompting narratives from large language models: sentiment assessment of ChatGPT-generated narratives and real tweets. *Future Internet*. (2023) 15:375. doi: 10.3390/fi15120375

26. Luo X, Deng Z, Yang B, Luo MY. Pre-trained language models in medicine: a survey. *Artif Intell Med*. (2024) 154:102904. doi: 10.1016/j.artmed.2024.102904

27. Wei WI, Leung CLK, Tang A, Mcneil EB, Wong SYS, Kwok KO. Extracting symptoms from free-text responses using ChatGPT among COVID-19 cases in Hong Kong. *Clin Microbiol Infect*. (2023) 30:142.e1–3. doi: 10.1016/j.cmi.2023.11.002

28. Hoang L, Guan Y, Kilicoglu H. Methodological information extraction from randomized controlled trial publications: a pilot study. AMIA Annu Symp Proc (2023) 2022:542–551. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10148349/ (accessed December 4, 2023).

29. Jiang J, Chao W-L, Culp S, Krishna SG. Artificial intelligence in the diagnosis and treatment of pancreatic cystic lesions and adenocarcinoma. *Cancers*. (2023) 15:2410. doi: 10.3390/cancers15092410

30. Pellat A, Barat M, Coriat R, Soyer P, Dohan A. Artificial intelligence: a review of current applications in hepatocellular carcinoma imaging. *Diagn Interv Imaging*. (2023) 104:24–36. doi: 10.1016/j.diii.2022.10.001

31. Chen H-Y, Ge P, Liu J-Y, Qu J-L, Bao F, Xu C-M, et al. Artificial intelligence: emerging player in the diagnosis and treatment of digestive disease. *World J Gastroenterol*. (2022) 28:2152–62. doi: 10.3748/wjg.v28.i20.2152

32. Shiihara H, Sonoda S, Terasaki H, Fujiwara K, Funatsu R, Shiba Y, et al. Wayfinding artificial intelligence to detect clinically meaningful spots of retinal diseases: artificial intelligence to help retina specialists in real world practice. *PLoS One*. (2023) 18:e0283214. doi: 10.1371/journal.pone.0283214

33. Lai W, Kuang M, Wang X, Ghafariasl P, Sabzalian MH, Lee S. Skin cancer diagnosis (SCD) using artificial neural network (ANN) and improved gray wolf optimization (IGWO). *Sci Rep*. (2023) 13:19377. doi: 10.1038/s41598-023-45039-w

34. Lynch CJ, Gore R. Short-range forecasting of COVID-19 during early onset at county, Health District, and state geographic levels using seven methods: comparative forecasting study. *J Med Internet Res*. (2021) 23:e24925. doi: 10.2196/24925

35. Feng L, Liu Z, Li C, Li Z, Lou X, Shao L, et al. Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a multicentre observational study. *Lancet Digit Health*. (2022) 4:e8–e17. doi: 10.1016/S2589-7500(21)00215-6

36. Hulman A, Dollerup OL, Mortensen JF, Fenech ME, Norman K, Støvring H, et al. ChatGPT- versus human-generated answers to frequently asked questions about diabetes: a Turing test-inspired survey among employees of a Danish diabetes center. *PLoS One*. (2023) 18:e0290773. doi: 10.1371/journal.pone.0290773

37. Cai X, Lai H, Wang X, Wang L, Liu W, Wang Y, et al. Comprehensive evaluation of molecule property prediction with ChatGPT. *Methods*. (2024) 222:133–41. doi: 10.1016/j.ymeth.2024.01.004

38. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. A comparative study of pretrained language models for long clinical text. *J Am Med Inform Assoc*. (2023) 30:340–7. doi: 10.1093/jamia/ocac225

39. Introduction | Langchain. Available at: https://python.langchain.com/docs/get_started/introduction (accessed December 4, 2023).

40. GanymedeNil/text2vec-large-chinese Hugging Face. Available at: https://huggingface.co/GanymedeNil/text2vec-large-chinese (accessed November 26, 2023).

41. Verma A, Goyal N, Bansal P, Gambhir P. Comparing the performance of various encoder models and vectorization techniques on text classification. In: 2023 14th international conference on computing communication and networking technologies (ICCCNT). Delhi, India: IEEE (2023). 1–7.

42. Qdrant – Vector Database. Available at: https://qdrant.tech/ (accessed December 3, 2023).

43. Qdrant Documentation – Qdrant. Available at: https://qdrant.tech/documentation/ (accessed December 5, 2023).

44. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. (2023) 55:1–38. doi: 10.1145/3571730

45. Cai LZ, Shaheen A, Jin A, Fukui R, Yi JS, Yannuzzi N, et al. Performance of generative large language models on ophthalmology board–style questions. *Am J Ophthalmol*. (2023) 254:141–9. doi: 10.1016/j.ajo.2023.05.024

46. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. (2023) 620:172–80. doi: 10.1038/s41586-023-06291-2

47. Jin D, Pan E, Oufattole N, Weng W-H, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl Sci*. (2021) 11:6421. doi: 10.3390/app11146421

48. Abdollahi M, Gao X, Mei Y, Ghosh S, Li J, Narag M. Substituting clinical features using synthetic medical phrases: medical text data augmentation techniques. *Artif Intell Med*. (2021) 120:102167. doi: 10.1016/j.artmed.2021.102167

49. Wrenn JO, Westerman D, Reeves RM, Ward MJ. 221EMF development and validation of a text rendering and data retrieval system for extracting clinical information from paper medical records. *Ann Emerg Med*. (2020) 76:S86. doi: 10.1016/j.annemergmed.2020.09.234

50. Chen N, Ren J. An EHR data quality evaluation approach based on medical knowledge and text matching. *IRBM*. (2023) 44:100782. doi: 10.1016/j.irbm.2023.100782

51. Evans RS. Electronic health records: then, now, and in the future. *Yearb Med Inform*. (2018) 25:S48–61. doi: 10.15265/IYS-2016-s006

52. AlMarzooqi FM, Moonesar IA, AlQutob R. Healthcare professional and user perceptions of eHealth data and record privacy in Dubai. *Information*. (2020) 11:415. doi: 10.3390/info11090415

53. Mollart L, Irwin P, Noble D, Kinsman L. Promoting patient safety using electronic medical records in nursing/midwifery undergraduate curricula: discussion paper. *Nurse Educ Pract*. (2023) 70:103653. doi: 10.1016/j.nepr.2023.103653

54. Khanna R, Shah E. Robotics in screening, diagnosis and treatment of breast Cancer: a perspective view. *Clin Breast Cancer*. (2024) 24:17–26. doi: 10.1016/j.clbc.2023.09.016

55. Hu X, Ruan M, Zou S, Huang M, Lin L, Zheng W, et al. Initial experience of robotic-assisted laparoendoscopic single site intraligamental myomectomy ambulatory surgery—report of two cases. *Int Surg*. (2023) 6:42–9. doi: 10.1016/j.isurg.2023.07.002

56. Sforza S, Marco BB, Haid B, Baydilli N, Donmez MI, Spinoit A-F, et al. A multi-institutional European comparative study of open versus robotic-assisted laparoscopic ureteral reimplantation in children with high grade (IV–V) vesicoureteral reflux. *J Pediatr Urol*. (2024) 20:283–91. doi: 10.1016/j.jpurol.2023.11.006

57. Zhang Y, Bai R, Han J, Chen Q, Gao X. Research on TCM Diabetes Assisted Diagnosis and Treatment Plan Integrating Association Mining and Quantitative Calculation. *Procedia Computer Science*. (2021). 188:52–60. Elsevier.

58. Gupta R, Kumari S, Senapati A, Ambasta RK, Kumar P. New era of artificial intelligence and machine learning-based detection, diagnosis, and therapeutics in Parkinson's disease. *Ageing Res Rev*. (2023) 90:102013. doi: 10.1016/j.arr.2023.102013

TYPE Mini Review
PUBLISHED 29 October 2024
DOI 10.3389/fmed.2024.1477898

OPEN ACCESS

EDITED BY
Rafat Damseh,
United Arab Emirates University,
United Arab Emirates

REVIEWED BY
Johanna Mora,
Bristol Myers Squibb, United States
Beenish Chaudhry,
University of Louisiana at Lafayette,
United States

*CORRESPONDENCE
Konstantinos Margetis
✉ Konstantinos.Margetis@mountsinai.org

RECEIVED 08 August 2024
ACCEPTED 03 October 2024
PUBLISHED 29 October 2024

CITATION
Aydin S, Karabacak M, Vlachos V and Margetis K (2024) Large language models in patient education: a scoping review of applications in medicine.
Front. Med. 11:1477898.
doi: 10.3389/fmed.2024.1477898

Large language models in patient education: a scoping review of applications in medicine

Serhat Aydin[1], Mert Karabacak[2], Victoria Vlachos[3] and Konstantinos Margetis[2]*

[1]School of Medicine, Koç University, Istanbul, Türkiye, [2]Department of Neurosurgery, Mount Sinai Health System, New York, NY, United States, [3]College of Human Ecology, Cornell University, Ithaca, NY, United States

Introduction: Large Language Models (LLMs) are sophisticated algorithms that analyze and generate vast amounts of textual data, mimicking human communication. Notable LLMs include GPT-4o by Open AI, Claude 3.5 Sonnet by Anthropic, and Gemini by Google. This scoping review aims to synthesize the current applications and potential uses of LLMs in patient education and engagement.

Materials and methods: Following the PRISMA-ScR checklist and methodologies by Arksey, O'Malley, and Levac, we conducted a scoping review. We searched PubMed in June 2024, using keywords and MeSH terms related to LLMs and patient education. Two authors conducted the initial screening, and discrepancies were resolved by consensus. We employed thematic analysis to address our primary research question.

Results: The review identified 201 studies, predominantly from the United States (58.2%). Six themes emerged: generating patient education materials, interpreting medical information, providing lifestyle recommendations, supporting customized medication use, offering perioperative care instructions, and optimizing doctor-patient interaction. LLMs were found to provide accurate responses to patient queries, enhance existing educational materials, and translate medical information into patient-friendly language. However, challenges such as readability, accuracy, and potential biases were noted.

Discussion: LLMs demonstrate significant potential in patient education and engagement by creating accessible educational materials, interpreting complex medical information, and enhancing communication between patients and healthcare providers. Nonetheless, issues related to the accuracy and readability of LLM-generated content, as well as ethical concerns, require further research and development. Future studies should focus on improving LLMs and ensuring content reliability while addressing ethical considerations.

KEYWORDS

large language models, ChatGPT, patient education, artificial intelligence, machine learning, deep learning

1 Introduction

Large Language Models (LLMs) are sophisticated algorithms that analyze and generate extensive textual data (1). These models leverage vast corpora of unlabeled text and incorporate reinforcement learning from human feedback to discern syntactical patterns and contextual nuances within languages. Consequently, LLMs can produce responses that closely mimic human communication when presented with diverse, open-ended queries (2–4). Several notable LLMs have emerged recently, including GPT-4o by Open AI (5), Claude 3.5 Sonnet by Anthropic (6), and Gemini by Google (7).

LLMs have demonstrated significant potential in medicine, with transformative applications across various domains, including clinical settings. These AI-powered systems can streamline clinical workflows, help with clinical decision-making, and ultimately improve patient outcomes. Recent studies highlight the utility of LLMs in clinical decision support, providing valuable insights that enable healthcare teams to make more informed treatment decisions (8–10). LLMs also show promise as educational tools by enhancing the quality and accessibility of materials. However, from a patient's perspective, they present both opportunities and risks. The varying levels of medical knowledge among patients may impede their ability to critically assess the information provided by LLMs, unlike clinicians who are trained to do so.

As of July 2024, there was limited synthesis of knowledge regarding the evidence base, applications, and evaluation methods of LLMs in patient education and engagement. This scoping review aims to address this gap by mapping the available literature on potential applications of LLMs in patient education and identifying future research directions. Our primary research question is: "What are the current and potential uses of LLMs in patient education and engagement as described in the literature?" This review seeks to enhance future discussions on using LLMs for patient care, including education, engagement, workload reduction, patient-centered health customization, and communication.

2 Materials and methods

This study employed a scoping review methodology, adhering to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) checklist (11). The review process was based on the methodological framework developed by Arksey and O'Malley (12), with further refinements as proposed by Levac et al. (13).

2.1 Literature search

A literature search was conducted in June 2024 using the PubMed database. The search strategy, detailed in Supplementary Methods S1, combined relevant keywords and Medical Subject Headings (MeSH) terms related to LLMs and patient education.

2.2 Study selection

Citation management was facilitated by Covidence software (Veritas Health Innovation). The inclusion criteria encompassed studies addressing the use, accuracy, relevance, or effectiveness of LLMs in patient education, patient engagement, answering patient-specific questions, or generating patient education materials. Studies were excluded if they did not primarily focus on LLMs for patient education, engagement, or answering patient questions; did not assess LLMs in healthcare settings or had only indirect relations to patients; or focused solely on technical aspects or architecture of LLMs without considering their application in patient education or engagement. A detailed description of the inclusion and exclusion criteria is provided in Supplementary Methods S2.

The selection process involved two stages. In the initial screening, two authors (SA and VV) independently reviewed the titles and abstracts of retrieved articles. Studies passing the initial screening were then read in full by both authors. Studies deemed eligible by both reviewers were included in the analysis. In cases of disagreement, a third author (MK) was consulted to resolve discrepancies.

2.3 Thematic analysis

We employed thematic analysis, following the methodology proposed by Braun and Clarke (14), to address our primary research question. The process began with an author (SA) reading and coding 25 randomly selected articles, focusing on content related to the potential uses of LLMs in patient education and engagement. Subsequently, two authors (SA and MK) examined the remaining manuscripts, seeking additional themes or data that could either reinforce or challenge the established themes. This iterative process facilitated further refinement of the themes through group discussions centered on patient education and engagement.

3 Results

3.1 Literature search

The initial search strategy yielded 661 papers. After removing one duplicate, 660 papers remained for screening. Based on title and abstract screening, 365 papers (55.3%) were excluded. Full-text review was conducted for 295 papers (44.7% of the initial pool), resulting in 201 papers (30% of the initial pool) meeting the study inclusion criteria (Supplementary Figure S1). Supplementary Data S1 presents all of the included papers.

3.2 Descriptive analysis

The geographical distribution of the studies revealed a predominance from the United States, accounting for 58.2% (117/201) of the articles. Turkey and China followed, each

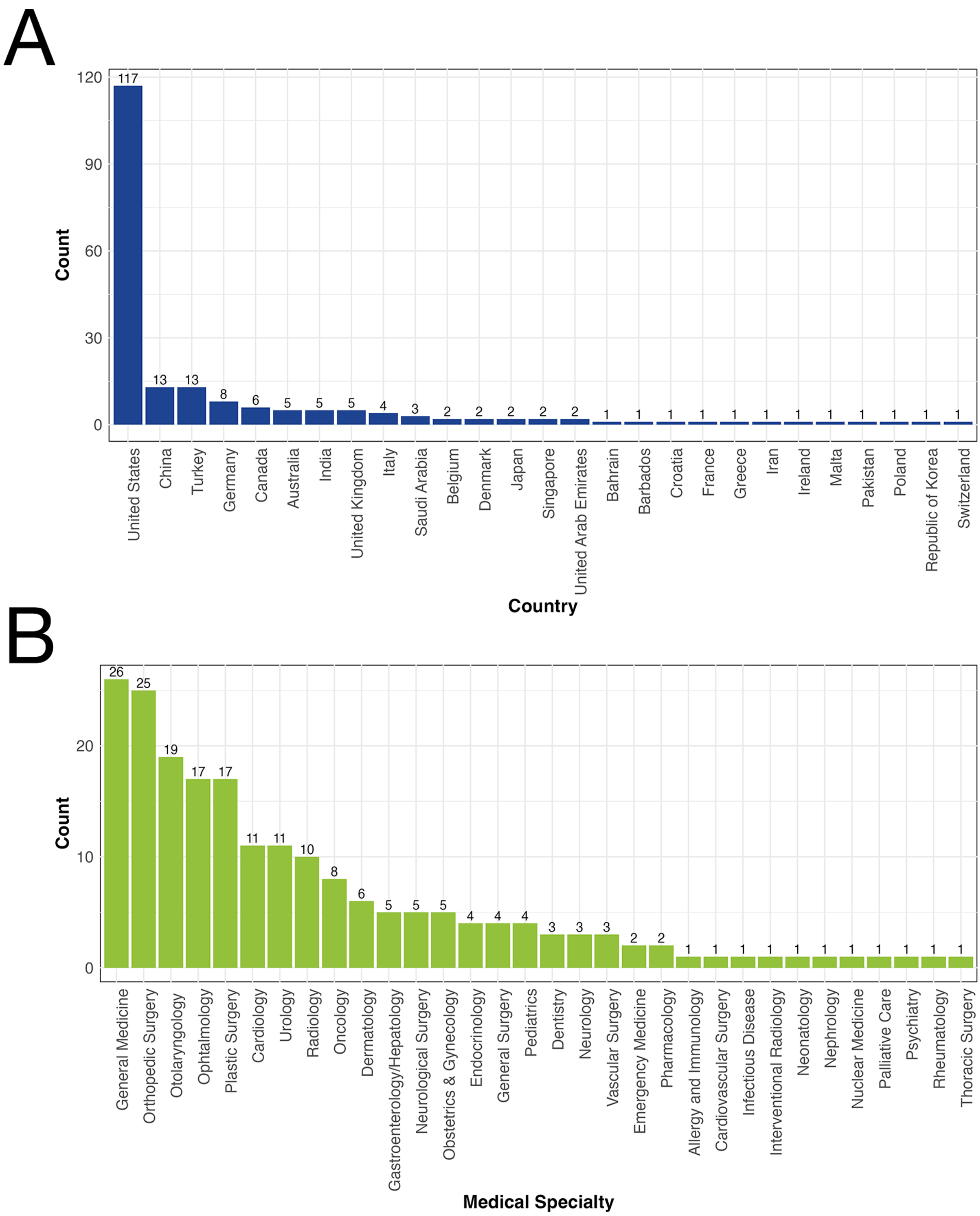


FIGURE 1
(A) Geographical distribution of studies on large language models (LLMs) in patient education. **(B)** Specialty distribution of studies on large language models (LLMs) in patient education.

contributing 6.4% (13/201) of the articles (Figure 1A). The studies spanned 35 medical specialties, with general medicine representing the largest proportion at 12.9% (26/201), closely followed by orthopedic surgery at 12.4% (25/201), and otolaryngology at 9.4% (19/201) (Figure 1B).

3.3 Thematic analysis

Our analysis identified six main themes with associated subthemes regarding the use of LLMs in patient education and engagement:

1 Generating Patient Education Materials
   a Answering Patient Questions
   b Enhancing Existing Patient Education Materials
   c Translation of Patient Education Materials
2 Interpreting Medical Information from a Patient Perspective
3 Providing Lifestyle Recommendations and Improving Health Literacy
4 Customized Medication Use and Self-Decision
5 Providing Pre-, Peri-, and Post-Operative Care Instructions
6 Optimizing Doctor-Patient Interaction

   a Facilitating Understanding of Consent Forms
   b Enhancing Communication Establishment

Table 1 presents these six themes as represented across the analyzed articles, along with illustrative quotes. Supplementary Data S2 indicates the theme to which each paper belongs.

The theme "Generating Patient Education Materials" was predominant, encompassing 80.5% (162/201) of the articles across its three subthemes. Within this theme, "Answering Patient Questions" was the most prevalent subtheme, representing 71.6% (144/201) of all articles. The remaining themes were distributed as follows: "Interpreting Medical Information from a Patient Perspective" and "Providing Lifestyle Recommendations and Improving Health Literacy" each accounted for 4.5% (9/201) of the articles. "Providing Pre-, Peri-, and Post-Operative Care Instructions" was represented in 6.9% (14/201) of the articles, while "Optimizing Doctor-Patient Interaction" appeared in 2.5% (5/201) of the articles. The least represented theme was "Customized Medication Use and Self-Decision," accounting for 1% (2/201) of the articles.

3.3.1 Theme 1: generating patient education materials

The generation of patient education materials emerged as a prominent theme, with three key subthemes: answering patient questions, enhancing existing materials, and translating medical content. Answering patient questions was the most significant subtheme, representing 71.6% of the articles (8, 15–157). In these studies, LLMs created educational content by responding to common questions, direct patient inquiries, and expert-formulated queries, demonstrating their potential to address diverse patient information needs.

Most studies found LLMs provided accurate responses to patient queries. Almagazzachi et al. reported 92.5% accuracy for ChatGPT's answers to hypertension questions (18). However, accuracy varied by specialty. In a study on pediatric in-toeing, Amaral et al. found 46% of responses were excellent, and 44% were satisfactory with minimal clarification needed (19). These findings suggest LLMs' potential in patient education, while highlighting performance differences across medical fields.

The readability of LLM-generated content varied considerably across studies. ChatGPT's responses often required a higher reading level, potentially limiting accessibility for some patients. Campbell et al. demonstrated that ChatGPT's unprompted answers on obstructive sleep apnea had a mean Flesch–Kincaid grade level of 14.15, which decreased to 12.45 when prompted (32). This indicates that even with specific instructions, the content remained at a college reading level. In contrast, other LLMs showed better readability in some cases. Chervonski et al. reported that Google BARD produced more accessible content, with responses on vascular surgery diseases achieving a mean Flesch Reading Ease score of 58.9, indicating improved readability (40). When compared to traditional search engines, LLMs revealed a trade-off between comprehensiveness and readability. Cohen et al. found that while ChatGPT provided more detailed and higher-quality responses to cataract surgery FAQs compared to Google, these responses were at a higher reading level (42). These findings suggest that while LLMs may offer more comprehensive information, they do not always improve accessibility for the average patient.

LLMs show promise in transforming existing materials into more readable, patient-centered formats (158–174). Numerous studies demonstrate their ability to enhance readability across various medical education materials (158–161, 163–165, 168, 170–172, 174). Fanning et al. found comparable performance between ChatGPT-3.5 and ChatGPT-4 in improving plastic surgery material readability (166). Moons et al. reported Google BARD surpassed GPT in readability improvement but tended to omit information (169). Some studies, however, found no improvement or decreased readability (162, 167), indicating variability in LLM effectiveness. Interestingly, Sudharshan et al. noted LLMs were more accurate in creating readable Spanish materials (173), suggesting potential for addressing language-specific challenges.

Research on LLMs for translating patient education materials remains limited. However, a significant study by Grimm et al. showed ChatGPT-4's ability to produce accurate, understandable, and actionable translations of otorhinolaryngology content in English, Spanish, and Mandarin (175). This finding suggests LLMs' potential in overcoming language barriers in patient education.

3.3.2 Theme 2: interpreting medical information from a patient perspective

Nine studies investigated LLMs' capacity to interpret complex medical information, evaluating their feasibility, accuracy, readability, and effectiveness in translating medical jargon. He et al. found ChatGPT-4 outperformed other LLMs and human responses from Q&A websites in accuracy, helpfulness, relevance, and safety when answering laboratory test result questions (176). However, Meyer et al. reported that ChatGPT, Gemini, and Le Chat were less accurate and more generalized than certified physicians in interpreting laboratory results (177), highlighting the variability in LLM performance across different contexts.

LLMs demonstrate potential in improving radiological information interpretation and communication. Kuckelman et al. found ChatGPT-4 produced generally accurate summaries of musculoskeletal radiology reports, noting some variability in human interpretation (82). Lyu et al. showed ChatGPT-4 enhanced translated radiology report quality and accessibility, despite occasional oversimplifications (178). Sarangi et al. reported ChatGPT-3.5 effectively simplified radiological reports while maintaining essential diagnostic information, though performance varied across conditions and imaging modalities (179). Several other studies support these findings, suggesting LLMs' promising role in radiology communication (180–182).

Zaretsky et al. evaluated ChatGPT-4's ability to convert discharge summaries into patient-friendly formats. The transformed summaries

TABLE 1 Representative quotes illustrating key themes identified in studies on the use of large language models (LLMs) in patient education.

Theme	Representative quotes
1. Generating Patient Education Materials a. Answering Patient Questions b. Enhancing Existing Patient Education Materials c. Translation of Patient Education Materials	**New Frontiers in Health Literacy: Using ChatGPT to Simplify Health Information for People in the Community** [Ayre et al. (159)] Ayre et al. evaluated ChatGPT-3.5's ability to simplify health information for individuals with low literacy. The study found that ChatGPT effectively reduced text complexity by lowering the reading level, using simpler language, and decreasing passive voice usage. It retained about 80% of key messages, with more complex texts seeing greater improvements. However, most simplified texts still did not meet recommended health literacy targets. The researchers concluded that ChatGPT could provide a useful "first draft" of plain language health information, which could then be refined through human review. **Enhancing Readability of Online Patient-Facing Content: The Role of AI Chatbots in Improving Cancer Information Accessibility** [Abreu et al. (158)] Abreu et al. assessed ChatGPT-4's effectiveness in improving the readability of cancer-related content from NCCN Member Institutions. The AI-generated outputs significantly reduced the reading level from university freshman to high school freshman level. This improvement in accessibility did not compromise content accuracy or quality. The simplified text featured shorter sentences and simpler words, earning a "good" quality rating on the DISCERN instrument. This study demonstrates AI's potential to make complex medical information more accessible to patients. **Leveraging large language models for generating responses to patient messages – a subjective analysis** [Liu et al. (94)] Liu et al. compared fine-tuned LLaMA-based models (CLAIR-Short and CLAIR-Long) with ChatGPT in generating responses to patient messages. CLAIR-Long, fine-tuned with a mix of local patient messages and open-source data, performed comparably to ChatGPT-4 in empathy, responsiveness, and accuracy. CLAIR-Short, fine-tuned only with local data, produced concise responses similar to healthcare providers but less detailed. While ChatGPT-4 generally ranked highest, the study showed that fine-tuned models, especially CLAIR-Long, could be effective for patient education and empathetic communication. **Assessing the Accuracy and Reliability of AI-Generated Responses to Patient Questions Regarding Spine Surgery** [Kasthuri et al. (76)] Kasthuri et al. evaluated the GPT-4-enhanced Bing search engine's responses to common spine surgery questions. Spine surgeons found the responses generally accurate and complete, with re-querying improving initially inaccurate answers. The study highlighted GPT-4-based models' ability to provide useful summaries from web sources, but noted concerns about response quality variability. Most information came from commercial websites, with no significant correlation between response accuracy and source type. This research underscores the need for ongoing evaluation and refinement of LLMs for clinical use. **Easing the Burden on Caregivers-Applications of Artificial Intelligence for Physicians and Caregivers of Children with Cleft Lip and Palate** [Chaker et al. (199)] Chaker et al. tested ChatGPT-3.5's ability to assist caregivers of children with cleft lip and palate. The AI achieved a 69% accuracy rate compared to senior pediatric plastic surgeons when answering common postoperative questions. While information-related errors were the AI's main weakness, the study emphasized AI's potential to ease caregiver burden by generating educational materials and offering perioperative support. This research highlights both the promise and current limitations of AI in specialized medical fields. **The utility of ChatGPT as a generative medical translator** [Grimm et al. (175)] Grimm et al. explored GPT-4's utility in translating otolaryngology-related medical content into English, Spanish, and Mandarin. Using the Patient Education Materials Assessment Tool (PEMAT), they found that GPT-4 produced translations with comparable accuracy, understandability, and actionability across all three languages. This study suggests that LLMs like GPT-4 could play a valuable role in bridging language barriers in healthcare, potentially improving access to medical information for diverse patient populations.
2. Interpreting Medical Information from a Patient Perspective	**Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study** [He et al. (176)] He et al. conducted a comprehensive evaluation of several LLMs in interpreting laboratory test results for lay patients. The study compared GPT-4, GPT-3.5, LLaMA 2, MedAlpaca, and ORCA_mini across multiple metrics including accuracy, relevance, helpfulness, and safety. GPT-4 emerged as the top performer in all categories, followed closely by GPT-3.5. LLaMA 2, while providing detailed explanations, ranked third. MedAlpaca and ORCA_mini were less effective, with MedAlpaca showing the poorest performance. This study highlights the current superiority of GPT-4 and GPT-3.5 in translating complex medical information for patient understanding, suggesting their potential utility in healthcare communication. **Translating musculoskeletal radiology reports into patient-friendly summaries using ChatGPT-4** [Kuckelman et al. (82)] Kuckelman et al. explored GPT-4's capability in simplifying musculoskeletal radiology reports for patients. The AI successfully generated summaries that were both readable and concise, with independent readers generally rating them as accurate and complete. GPT-4 demonstrated proficiency in simplifying medical jargon, making reports more accessible to patients. While there was some variation in accuracy and completeness ratings among readers, indicating a degree of subjectivity, the overall results were positive. The study suggests that GPT-4 could be a valuable tool in enhancing patient comprehension of radiology results, potentially reducing the immediate need for physician explanation. **Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format** [Zaretsky et al. (183)] Zaretsky et al. investigated GPT-4's ability to transform complex inpatient discharge summaries into more patient-friendly formats. The AI-transformed summaries showed marked improvements in readability, with the Flesch–Kincaid Grade Level decreasing from 11.0 to 6.2. Understandability scores, measured by PEMAT, increased significantly from 13 to 81%. However, the study revealed mixed results in terms of accuracy and completeness. While 54% of reviews gave the highest accuracy rating, 18% identified safety concerns due to omissions or incorrect information (hallucinations). These findings indicate that while GPT-4 can greatly enhance the accessibility of discharge information, further refinement is necessary to ensure consistent accuracy and safety for practical use in healthcare settings.

(Continued)

TABLE 1 (Continued)

Theme	Representative quotes
3. Providing Lifestyle Recommendations and Improving Health Literacy	**Examining the role of ChatGPT in promoting health behaviors and lifestyle changes among cancer patients** [Alanezi et al. (184)] Alanezi et al. explored ChatGPT-3.5's potential in promoting health behavior changes among cancer patients. The study found that the AI significantly improved health literacy, enhanced self-management practices, and provided valuable emotional and motivational support. Patients appreciated the AI's ability to address their concerns, offer personalized suggestions, and connect them with relevant resources. However, the research also identified challenges, including privacy concerns, limitations in deep personalization, and occasional reliability issues. Despite these drawbacks, ChatGPT-3.5 proved effective in facilitating positive health behaviors and lifestyle changes, particularly in helping patients better understand and manage their conditions. **Assessing the Accuracy of Generative Conversational Artificial Intelligence in Debunking Sleep Health Myths: Mixed Methods Comparative Study With Expert Analysis** [Bragazzi et al. (185)] Bragazzi et al. assessed GPT-4's accuracy in debunking common sleep-related myths. The AI correctly identified 85% of the presented myths as either "false" or "generally false," demonstrating a sensitivity of 85% and a positive predictive value of 100%. GPT-4's performance in identifying false statements was comparable to that of sleep experts, with high interrater agreement (ICC = 0.83). However, the AI sometimes struggled with nuanced scenarios, particularly myths containing partial truths or complex scientific concepts. The study concluded that while GPT-4 is a reliable tool for addressing sleep-related misinformation, it should not replace expert opinion in more nuanced areas. **Is ChatGPT an Effective Tool for Providing Dietary Advice?** [Ponzo et al. (190)] Ponzo et al. evaluated ChatGPT-3.5's ability to provide accurate and appropriate dietary advice for various non-communicable diseases (NCDs). The AI's advice was generally appropriate, with correctness rates ranging from 55.5% for sarcopenia to 73.3% for non-alcoholic fatty liver disease (NAFLD). However, the study revealed limitations in complex scenarios involving multiple overlapping conditions, where ChatGPT-3.5 sometimes provided contradictory or inappropriate recommendations. The researchers concluded that while ChatGPT-3.5 shows promise as a supplementary tool for dietary advice, it cannot yet replace personalized guidance from healthcare professionals, especially in managing complex cases.
4. Customized Medication Use and Self-Decision	**Snakebite Advice and Counseling From Artificial Intelligence: An Acute Venomous Snakebite Consultation With ChatGPT** [Altamimi et al. (192)] Altamimi et al. evaluated ChatGPT-3.5's performance in providing information for managing venomous snakebites. The AI offered clear, evidence-based advice on initial first aid, the importance of seeking urgent medical attention, potential symptoms, and the role of antivenom. However, the study identified several limitations in the AI's capabilities. These included a lack of personalization, outdated information, and an inability to account for regional variations in snake species and venom characteristics. While ChatGPT-3.5 proved effective in delivering general advice and preliminary guidance, the researchers emphasized that it should not replace professional medical consultations, especially in critical situations like snakebites. The study concluded by recommending future developments focus on addressing these limitations to enhance the AI's utility in such scenarios. **Automating untruths: ChatGPT, self-managed medication abortion, and the threat of misinformation in a post-Roe world** [McMahon et al. (193)] McMahon et al. investigated the accuracy of ChatGPT-3.5's responses regarding self-managed medication abortion (SMMA). The study revealed a concerning discrepancy in the AI's information provision. While ChatGPT-3.5 correctly described clinician-managed medication abortion as safe and effective, it inaccurately portrayed SMMA as significantly more dangerous, exaggerating the risks of complications. This misrepresentation contradicts substantial evidence supporting SMMA's safety and effectiveness. The researchers highlighted the potential dangers of such misinformation, noting it could increase stigma and deter individuals from seeking safe abortion methods, thereby posing a threat to public health. These findings emphasize the critical need for improving AI models to ensure they provide accurate and reliable health information, particularly on sensitive topics with significant public health implications.

(Continued)

10.3389/fmed.2024.1477898

TABLE 1 (Continued)

Theme	Representative quotes
5. Providing Pre-, Peri-, and Post-Operative Care Instructions	**Enhancing Postoperative Cochlear Implant Care With ChatGPT-4: A Study on Artificial Intelligence (AI)-Assisted Patient Education and Support** [Aliyeva et al. (194)] Aliyeva et al. evaluated ChatGPT-4's effectiveness in providing postoperative care information for cochlear implant patients. The AI demonstrated high accuracy, clarity, and relevance in answering common postoperative questions. Its responses aligned well with current medical guidelines, ensuring patients received accurate and comprehensible information. The study found ChatGPT-4 to be a valuable supplementary resource, especially when access to healthcare professionals is limited. While emphasizing that ChatGPT-4 cannot replace professional medical advice, the researchers noted its potential to support patient education and reduce anxiety by providing timely information in resource-constrained settings. **Evaluation of large language model responses to Mohs surgery preoperative questions** [Breneman et al. (206)] Breneman et al. compared the performance of three large language models (ChatGPT-3.5, Google Bard, and Microsoft CoPilot) in answering preoperative questions about Mohs surgery. ChatGPT-3.5 outperformed the other models in accuracy (80%) and completeness (100%) of responses. However, its higher reading level (12.7) potentially made the information less accessible to some patients. Google Bard and Microsoft CoPilot, while less accurate and complete, provided more readable responses. The study highlighted the potential of LLMs like ChatGPT-3.5 in offering valuable preoperative information but cautioned about possible inaccuracies or irrelevant details, emphasizing the need for careful implementation in patient education. **Feasibility of GPT-3 and GPT-4 for in-Depth Patient Education Prior to Interventional Radiological Procedures: A Comparative Analysis** [Scheschenja et al. (195)] Scheschenja et al. conducted a comparative analysis of GPT-3 and GPT-4 in providing patient education for interventional radiology procedures. GPT-4 showed superior performance, with 35.3% of its responses rated as "completely correct" compared to GPT-3's 30.8%. GPT-4 also had fewer "mostly incorrect" responses (2.3% vs. GPT-3's 5.3%). Despite these differences, both models were considered safe and effective for patient education, with GPT-4 having a slight edge. The researchers concluded that while these AI tools can enhance patient understanding of complex procedures, they should be used cautiously due to the potential for inaccuracies or incomplete information.
6. Optimizing Doctor-Patient Interaction a. Facilitating Understanding of Consent Forms b. Enhancing Communication Establishment	**Bridging the literacy gap for surgical consents: an AI-human expert collaborative approach** [Ali et al. (208)] Ali et al. investigated the use of GPT-4 to simplify surgical consent forms, aiming to make them more accessible to patients with varying health literacy levels. The study found that GPT-4 significantly improved the readability of consent forms from 15 academic medical centers, reducing the average reading level from college freshman to 8th-grade level. Moreover, GPT-4 generated procedure-specific consent forms that maintained medical and legal sufficiency, scoring perfectly on a validated rubric and passing expert review without changes. This research demonstrates the potential of AI-human collaboration in enhancing the clarity and comprehensibility of consent forms, ensuring patients receive clear, detailed information about their surgical procedures. **Generating Informed Consent Documents Related to Blepharoplasty Using ChatGPT** [Shiraishi et al. (209)] Shiraishi et al. evaluated ChatGPT's performance in generating informed consent (IC) documents for blepharoplasty. While the study showed promise for LLMs in enhancing patient communication, it also highlighted areas needing improvement. Board-certified plastic surgeons rated AI-generated documents lower than original IC documents in accuracy, informativeness, and accessibility. Even after revisions, the AI-generated documents still scored lower in accuracy and accessibility. Interestingly, nonmedical staff found no significant difference between AI-generated and original documents. The study concluded that while ChatGPT has potential, it currently cannot replace human-generated IC documents due to issues with professional terminology and content accuracy, emphasizing the need for further refinement. **Putting ChatGPT's Medical Advice to the (Turing) Test: Survey Study** [Nov et al. (110)] Nov et al. assessed laypeople's ability to distinguish between medical advice from ChatGPT-3.5 and human healthcare providers. Participants could only weakly differentiate between the sources, correctly identifying them about 65% of the time. Trust in ChatGPT-3.5's responses decreased with increasing medical complexity of the questions, with higher trust in logistical responses and lower trust in diagnostic and treatment-related responses. The study concluded that while ChatGPT-3.5 can provide credible advice for low-risk queries, it may not be reliable for more complex health issues, suggesting the need for further research to optimize its use in patient-provider communications. **Can Large Language Models Generate Outpatient Clinic Letters at First Consultation That Incorporate Complication Profiles From UK and USA Aesthetic Plastic Surgery Associations?** [Roberts et al. (211)] Roberts et al. compared ChatGPT-4, ChatGPT-3.5, and Google Bard in generating outpatient clinic letters incorporating complication profiles from aesthetic plastic surgery associations. ChatGPT-4 showed the highest overall compliance, scoring 0.92 for BAAPS and 0.99 for ASPS compliance. However, its performance dropped to 0.52 for ASPS gold-standard profiles, indicating challenges with paywalled content. ChatGPT-3.5 and Google Bard demonstrated lower compliance overall. This study highlights the potential of advanced LLMs in generating compliant medical documentation, while also revealing limitations in accessing and integrating specialized, restricted information.

10.3389/fmed.2024.1477898

showed significant improvements in readability and understandability. However, the study raised concerns about accuracy and completeness, noting instances of omissions and hallucinations (183).

3.3.3 Theme 3: providing lifestyle recommendations and improving health literacy

Nine studies explored LLMs' potential in offering lifestyle recommendations and enhancing health literacy. Alanezi et al. found ChatGPT effective in promoting health behavior changes among cancer patients, boosting health literacy and self-management (184). Bragazzi et al. showed ChatGPT's capability to debunk sleep-related myths and provide accessible advice (185). In a follow-up study, they found Google BARD slightly outperformed ChatGPT-4 in identifying false statements and offering practical sleep-related advice (186). These findings suggest LLMs' promising role in health education and lifestyle guidance.

Gray et al. demonstrated ChatGPT's ability to generate realistic prenatal counseling dialogues (187). Minutolo et al. proposed a conversational agent to enhance health literacy by making Patient Information Leaflets queryable (188). Mondal et al. found ChatGPT provided reasonably accurate responses to lifestyle-related disease queries (189). Ponzo et al. reported ChatGPT offered general dietary guidance for NCDs but struggled with complex, multi-condition cases (190). Willms et al. explored ChatGPT's potential in creating physical activity app content, emphasizing the need for expert review (1). Zaleski et al. found AI-generated exercise recommendations generally accurate but lacking comprehensiveness and at a college reading level (191). These studies highlight LLMs' diverse applications in health education while noting their limitations.

3.3.4 Theme 4: customized medication use and self-decision

Two studies explored LLMs' potential in medication guidance and self-decision support. Altamimi et al. found ChatGPT provided accurate advice on acute venomous snakebite management, while emphasizing the importance of professional care (192). In contrast, McMahon et al. observed ChatGPT accurately described clinician-managed abortion as safe but incorrectly portrayed self-managed abortion as dangerous, highlighting potential misinformation risks (193). These findings underscore both the promise and pitfalls of using LLMs for sensitive medical information.

3.3.5 Theme 5: providing pre-/peri-/post-operative care instructions

Studies investigated LLMs' use in surgical patient education. Aliyeva et al. found ChatGPT-4 excelled in providing postoperative care instructions for cochlear implant patients, especially in remote settings (194). LLMs showed proficiency in offering postoperative guidance across various surgical specialties (180, 195–202). Dhar et al. noted ChatGPT's accuracy in answering tonsillectomy questions, with some pain management inaccuracies (203). Patil et al. reported ChatGPT provided quality preoperative information for ophthalmic surgeries, though occasionally overlooking adverse events (204). Meyer et al. found ChatGPT reliable for postoperative gynecological surgery instructions (205). Breneman et al. and Kienzle et al. evaluated ChatGPT for preoperative counseling in Mohs surgery and knee arthroplasty, finding it potentially useful but cautioning about non-existing references (206, 207).

3.3.6 Theme 6: optimizing doctor-patient interaction

This theme explores LLMs' potential to enhance doctor-patient communication, particularly in simplifying consent forms and improving general medical communication. Ali et al. found ChatGPT-4 successfully simplified surgical consent forms to an 8th-grade reading level while maintaining accuracy (208). Shiraishi et al. reported that revised ChatGPT-prepared informed consent documents for blepharoplasty were more desirable than originals (209).

LLMs also showed promise in broader doctor-patient communication. An et al. introduced an LLM-based education model that improved patients' understanding of their conditions and treatments (210). Roberts et al. demonstrated LLMs could generate comprehensible outpatient clinic letters for cosmetic surgery, potentially saving clinicians' time (211). Xue et al. found ChatGPT performed well in logical reasoning and medical knowledge education during remote orthopedic consultations (212). These studies highlight LLMs' potential to enhance various aspects of medical communication.

4 Discussion

This scoping review synthesizes current applications and potential uses of LLMs in patient education and engagement, offering insights into their transformative potential and integration challenges in healthcare settings. LLMs demonstrate significant promise in creating patient education materials, with studies reporting that health-related questions were accurately answered over 90% of the time by systems like ChatGPT, covering a broad range of topics from hypertension to pediatric conditions (18, 31). The depth of these responses potentially offers substantial value to patients seeking detailed understanding of their ailments. However, readability remains a notable concern, potentially limiting accessibility for some patient populations.

LLMs have demonstrated competence in interpreting complex medical information from laboratory reports, radiology results, and discharge summaries. ChatGPT-4, for instance, generated informative summaries of radiology reports, making them more accessible to non-medical professionals (82, 178). However, concerns about the quality and comprehensiveness of LLM-generated information persist. Issues such as hallucinations, omissions, or plausible but incorrect information have been noted. Zaretsky et al. observed that while ChatGPT-4 could transform discharge summaries into more patient-friendly formats, occasional inaccuracies, and omissions could potentially mislead patients (183). These findings underscore the necessity for professional oversight in deploying LLMs in healthcare settings to ensure the reliability and accuracy of AI-generated content.

LLMs show promise as lifestyle recommendations and health literacy tools, effectively encouraging healthy behaviors and dispelling health myths. Alanezi et al. found that ChatGPT provided significant support in developing health literacy among cancer patients, motivating self-management through emotional, informational, and motivational assistance (184). Bragazzi and Garbarino demonstrated ChatGPT's effectiveness in debunking sleep-related misconceptions, accurately distinguishing between false and genuine health information (185). However, personalization and accuracy remain challenging. While AI can offer useful preliminary advice, it requires further development to provide relevant, situation-specific suggestions tailored to individual patients. This customization is crucial for ensuring that patients can trust and adhere to the recommendations provided.

LLMs play a significant role in providing information on self-medication and personalized drug utilization, offering detailed insights on drug interactions, correct usage, and potential side effects. Altamimi et al. found ChatGPT's information helpful and accurate in guiding acute venomous snakebite management, though it appropriately emphasized the need for professional medical care (192). LLMs also show potential in patient triage, quickly analyzing symptoms and medical history to prioritize cases based on severity (10). However, the quality of LLM-provided information varies considerably. McMahon et al. reported that ChatGPT gave inaccurate and misleading information about self-managed medication abortion, incorrectly portraying it as dangerous despite evidence of its safety and efficacy (193). This inconsistency highlights the risks of relying on AI without professional oversight and underscores the need for LLMs to provide accurate, up-to-date, and context-sensitive information to support safe self-medication practices.

4.1 Implications and future research

The integration of LLMs into patient education and engagement shows significant potential for improving health literacy and healthcare delivery efficiency. However, this review highlights the need for continued improvement in the accuracy and personalization of AI-generated content. Future research should focus on developing more accurate LLM algorithms to enhance reliability as medical information sources, exploring multimodal LLMs, and establishing robust validation frameworks for their ethical use. Ensuring AI-based information aligns with the latest medical guidelines and is tailored for diverse patient populations is crucial. Conducting longitudinal studies to assess the long-term effects of LLMs on patient outcomes and satisfaction will provide valuable insights. Additionally, addressing ethical concerns, including data privacy and potential biases in LLM-generated content, is essential. These research directions are crucial for the responsible and effective integration of LLMs in healthcare settings. Finally, LLMs may carry biases from their training data, potentially propagating misinformation or reinforcing healthcare disparities. Future research should address these limitations by ensuring LLM tools are accurate, reliable, and equitable across diverse patient populations, while also exploring their long-term effects and ethical implications.

4.2 Limitations

This scoping review has several limitations. The quality of included studies varied, with some using small sample sizes or subjective assessments, potentially limiting result generalizability. Most studies were conducted in high-income countries, raising questions about their relevance to low-and middle-income settings with different healthcare needs and infrastructure. The evaluation of various LLMs and versions complicates drawing overarching conclusions. Inconsistent evaluation metrics across studies hindered result comparison and synthesis.

5 Conclusion

LLMs demonstrate transformative potential in patient education and engagement across various levels of medical care. Their ability to provide accurate, detailed, and timely information can significantly enhance patients' understanding of their healthcare and promote active involvement. However, current limitations in accuracy and readability highlight the need for further refinement to ensure reliable integration with healthcare systems. Extensive research and development of AI tools are necessary to fully harness their potential for improving patient outcomes and healthcare efficiency. A critical priority for medical applications is to ensure the ethical and responsible use of these tools, necessitating robust supervision and validation processes.

Author contributions

SA: Conceptualization, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing, Data curation. MK: Conceptualization, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing, Project administration, Supervision. VV: Conceptualization, Data curation, Writing – original draft, Writing – review & editing. KM: Conceptualization, Investigation, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2024.1477898/full#supplementary-material

References

1. Willms A, Liu S. Exploring the feasibility of using Chatgpt to create just-in-time adaptive physical activity mHealth intervention content: case study. *JMIR Med Educ*. (2024) 10:e51426. doi: 10.2196/51426

2. Park YJ, Pillai A, Deng J, Guo E, Gupta M, Paget M, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Med Inform Decis Mak*. (2024) 24:72. doi: 10.1186/s12911-024-02459-6

3. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. *iScience*. (2024) 27:109713. doi: 10.1016/j.isci.2024.109713

4. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of Chatgpt and other large language models. *JAMA*. (2023) 330:315–6. doi: 10.1001/jama.2023.9651

5. Open AI. (2024). Available at: https://openai.com/index/hello-gpt-4o (accessed 2024).

6. Anthropic. (2024). Available at: https://www.anthropic.com/news/claude-3-5-sonnet (accessed 2024).

7. Google. (2023). Available at: https://gemini.google.com/ (accessed 2024).

8. Peng W, Feng Y, Yao C, Zhang S, Zhuo H, Qiu T, et al. Evaluating Ai in medicine: a comparative analysis of expert and Chatgpt responses to colorectal Cancer questions. *Sci Rep*. (2024) 14:2840. doi: 10.1038/s41598-024-52853-3

9. Sallam M. Healthcare Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. (2023) 11:887. doi: 10.3390/healthcare11060887

10. Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, et al. The Role of large language models in transforming emergency medicine: scoping review. *JMIR Med Inform*. (2024) 12:e53787:e53787. doi: 10.2196/53787

11. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. Prisma extension for scoping reviews (Prisma-Scr): checklist and explanation. *Ann Intern Med*. (2018) 169:467–73. doi: 10.7326/M18-0850

12. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol*. (2005) 8:19–32. doi: 10.1080/1364557032000119616

13. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. *Implement Sci*. (2010) 5:1–9. doi: 10.1186/1748-5908-5-69/TABLES/3

14. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. (2006) 3:77–101. doi: 10.1191/1478088706qp063oa

15. Al-Sharif EM, Penteado RC, Dib El Jalbout N, Topilow NJ, Shoji MK, Kikkawa DO, et al. Evaluating the accuracy of Chatgpt and Google Bard in fielding oculoplastic patient queries: a comparative study on artificial versus human intelligence. Ophthalmic. *Plast Reconstr Surg*. (2024) 40:303–11. doi: 10.1097/IOP.0000000000002567

16. Alapati R, Campbell D, Molin N, Creighton E, Wei Z, Boon M, et al. Evaluating insomnia queries from an artificial intelligence Chatbot for patient education. *J Clin Sleep Med*. (2024) 20:583–94. doi: 10.5664/jcsm.10948

17. Alessandri-Bonetti M, Liu HY, Palmesano M, Nguyen VT, Egro FM. Online patient education in body contouring: a comparison between Google and Chatgpt. *J Plast Reconstr Aesthet Surg*. (2023) 87:390–402. doi: 10.1016/j.bjps.2023.10.091

18. Almagazzachi A, Mustafa A, Eighaei Sedeh A, Vazquez Gonzalez AE, Polianovskaia A, Abood M, et al. Generative artificial intelligence in patient education: Chatgpt takes on hypertension questions. *Cureus*. (2024) 16:e53441. doi: 10.7759/cureus.53441

19. Amaral JZ, Schultz RJ, Martin BM, Taylor T, Touban B, McGraw-Heinrich J, et al. Evaluating chat generative pre-trained transformer responses to common pediatric in-toeing questions. *J Pediatr Orthop*. (2024) 44:e592–7. doi: 10.1097/BPO.0000000000002695

20. Amin KS, Mayes LC, Khosla P, Doshi RH. Assessing the efficacy of large language models in health literacy: a comprehensive cross-sectional study. *Yale J Biol Med*. (2024) 97:17–27. doi: 10.59249/ZTOZ1966

21. Anastasio AT, FBT M, Karavan MPJr, Adams SBJr. Evaluating the quality and usability of artificial intelligence-generated responses to common patient questions in foot and ankle surgery. *Foot Ankle Orthop*. (2023) 8:24730114231209919. doi: 10.1177/24730114231209919

22. Atarere J, Naqvi H, Haas C, Adewunmi C, Bandaru S, Allamneni R, et al. Applicability of online chat-based artificial intelligence models to colorectal Cancer screening. *Dig Dis Sci*. (2024) 69:791–7. doi: 10.1007/s10620-024-08274-3

23. Athavale A, Baier J, Ross E, Fukaya E. The potential of Chatbots in chronic venous disease patient management. *JVS Vasc Insights*. (2023) 1:1. doi: 10.1016/j.jvsvi.2023.100019

24. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. (2023) 183:589–96. doi: 10.1001/jamainternmed.2023.1838

25. Ayoub NF, Lee YJ, Grimm D, Balakrishnan K. Comparison between Chatgpt and Google search as sources of postoperative patient instructions. *JAMA Otolaryngol Head Neck Surg*. (2023) 149:556–8. doi: 10.1001/jamaoto.2023.0704

26. Ayoub NF, Lee YJ, Grimm D, Divi V. Head-to-head comparison of Chatgpt versus Google search for medical knowledge acquisition. *Otolaryngol Head Neck Surg*. (2024) 170:1484–91. doi: 10.1002/ohn.465

27. Balel Y. Can Chatgpt be used in Oral and maxillofacial surgery? *J Stomatol Oral Maxillofac Surg*. (2023) 124:101471. doi: 10.1016/j.jormas.2023.101471

28. Bellinger JR, De La Chapa JS, Kwak MW, Ramos GA, Morrison D, Kesser BW. Bppv information on Google versus ai (Chatgpt). *Otolaryngol Head Neck Surg*. (2024) 170:1504–11. doi: 10.1002/ohn.506

29. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model Chatbot responses to online patient eye care questions. *JAMA Netw Open*. (2023) 6:e2330320. doi: 10.1001/jamanetworkopen.2023.30320

30. Brozović J, Mikulić B, Tomas M, Juzbašić M, Blašković M. Assessing the performance of Bing chat artificial intelligence: dental exams, clinical guidelines, and Patients' frequent questions. *J Dent*. (2024) 144:104927. doi: 10.1016/j.jdent.2024.104927

31. Caglar U, Yildiz O, Meric A, Ayranci A, Gelmis M, Sarilar O, et al. Evaluating the performance of Chatgpt in answering questions related to pediatric urology. *J Pediatr Urol*. (2024) 20:26.e1–5. doi: 10.1016/j.jpurol.2023.08.003

32. Campbell DJ, Estephan LE, Mastrolonardo EV, Amin DR, Huntley CT, Boon MS. Evaluating Chatgpt responses on obstructive sleep apnea for patient education. *J Clin Sleep Med*. (2023) 19:1989–95. doi: 10.5664/jcsm.10728

33. Campbell DJ, Estephan LE, Sina EM, Mastrolonardo EV, Alapati R, Amin DR, et al. Evaluating Chatgpt responses on thyroid nodules for patient education. *Thyroid*. (2024) 34:371–7. doi: 10.1089/thy.2023.0491

34. Cappellani F, Card KR, Shields CL, Pulido JS, Haller JA. Reliability and accuracy of artificial intelligence Chatgpt in providing information on ophthalmic diseases and management to patients. *Eye (Lond)*. (2024) 38:1368–73. doi: 10.1038/s41433-023-02906-0

35. Carnino JM, Pellegrini WR, Willis M, Cohen MB, Paz-Lansberg M, Davis EM, et al. Assessing Chatgpt's responses to otolaryngology patient questions. *Ann Otol Rhinol Laryngol*. (2024) 133:658–64. doi: 10.1177/00034894241249621

36. Chen D, Parsa R, Hope A, Hannon B, Mak E, Eng L, et al. Physician and artificial intelligence Chatbot responses to Cancer questions from social media. *JAMA Oncol*. (2024) 10:956–60. doi: 10.1001/jamaoncol.2024.0836

37. Chen X, Zhang W, Zhao Z, Xu P, Zheng Y, Shi D, et al. Icga-Gpt: report generation and question answering for Indocyanine green angiography images. *Br J Ophthalmol*. (2024) 108:1450–6. doi: 10.1136/bjo-2023-324446

38. Cheong KX, Zhang C, Tan TE, Fenner BJ, Wong WM, Teo KY, et al. Comparing generative and retrieval-based Chatbots in answering patient questions regarding age-related macular degeneration and diabetic retinopathy. *Br J Ophthalmol*. (2024) 108:1443–9. doi: 10.1136/bjo-2023-324533

39. Cheong RCT, Unadkat S, McNeillis V, Williamson A, Joseph J, Randhawa P, et al. Artificial intelligence Chatbots as sources of patient education material for obstructive sleep Apnoea: Chatgpt versus Google Bard. *Eur Arch Otorrinolaringol*. (2024) 281:985–93. doi: 10.1007/s00405-023-08319-9

40. Chervonski E, Harish KB, Rockman CB, Sadek M, Teter KA, Jacobowitz GR, et al. Generative artificial intelligence Chatbots may provide appropriate informational responses to common vascular surgery questions by patients. *Vascular*. (2024):17085381241240550. doi: 10.1177/17085381241240550

41. Christy M, Morris MT, Goldfarb CA, Dy CJ. Appropriateness and reliability of an online artificial intelligence Platform's responses to common questions regarding distal radius fractures. *J Hand Surg Am*. (2024) 49:91–8. doi: 10.1016/j.jhsa.2023.10.019

42. Cohen SA, Brant A, Fisher AC, Pershing S, Do D, Pan C. Dr. Google vs. Dr. Chatgpt: exploring the use of artificial intelligence in ophthalmology by comparing the accuracy, safety, and readability of responses to frequently asked patient questions regarding cataracts and cataract surgery. *Semin Ophthalmol*. (2024) 39:472–9. doi: 10.1080/08820538.2024.2326058

43. Connors C, Gupta K, Khusid JA, Khargi R, Yaghoubian AJ, Levy M, et al. Evaluation of the current status of artificial intelligence for Endourology patient education: a blind comparison of Chatgpt and Google Bard against traditional information resources. *J Endourol*. (2024) 38:843–51. doi: 10.1089/end.2023.0696

44. Cornelison BR, Erstad BL, Edwards C. Accuracy of a Chatbot in answering questions that patients should ask before taking a new medication. *J Am Pharm Assoc*. (2003) 64:102110. doi: 10.1016/j.japh.2024.102110

45. Croen BJ, Abdullah MS, Berns E, Rapaport S, Hahn AK, Barrett CC, et al. Evaluation of patient education materials from large-language artificial intelligence models on carpal tunnel release. *Hand*. (2024) N Y:15589447241247332. doi: 10.1177/15589447241247332

46. Crook BS, Park CN, Hurley ET, Richard MJ, Pidgeon TS. Evaluation of online artificial intelligence-generated information on common hand procedures. *J Hand Surg Am*. (2023) 48:1122–7. doi: 10.1016/j.jhsa.2023.08.003

47. Cung M, Sosa B, Yang HS, McDonald MM, Matthews BG, Vlug AG, et al. The performance of artificial intelligence Chatbot large language models to address skeletal

biology and bone health queries. *J Bone Miner Res*. (2024) 39:106–15. doi: 10.1093/jbmr/zjad007

48. Davis R, Eppler M, Ayo-Ajibola O, Loh-Doyle JC, Nabhani J, Samplaski M, et al. Evaluating the effectiveness of artificial intelligence-powered large language models application in disseminating appropriate and readable health information in urology. *J Urol*. (2023) 210:688–94. doi: 10.1097/JU.0000000000003615

49. Dimitriadis F, Alkagiet S, Tsigkriki L, Kleitsioti P, Sidiropoulos G, Efstratiou D, et al. Chatgpt and patients with heart failure. *Angiology*. (2024):33197241238403. doi: 10.1177/00033197241238403

50. Doğan L, Özçakmakcı GB, Yılmaz İE. The performance of Chatbots and the Aapos website as a tool for amblyopia education. *J Pediatr Ophthalmol Strabismus*. (2024) 61:325–31. doi: 10.3928/01913913-20240409-01

51. Dubin JA, Bains SS, DeRogatis MJ, Moore MC, Hameed D, Mont MA, et al. Appropriateness of frequently asked patient questions following Total hip arthroplasty from Chatgpt compared to arthroplasty-trained nurses. *J Arthroplast*. (2024) 39:S306–11. doi: 10.1016/j.arth.2024.04.020

52. Durairaj KK, Baker O, Bertossi D, Dayan S, Karimi K, Kim R, et al. Artificial intelligence versus expert plastic surgeon: comparative study shows Chatgpt "wins" Rhinoplasty consultations: should we be worried? *Facial Plast Surg Aesthet Med*. (2023) 26:270–5. doi: 10.1089/fpsam.2023.0224

53. Fahy S, Niemann M, Böhm P, Winkler T, Oehme S. Assessment of the quality and readability of information provided by Chatgpt in relation to the use of platelet-rich plasma therapy for osteoarthritis. *J Pers Med*. (2024) 14:495. doi: 10.3390/jpm14050495

54. Fahy S, Oehme S, Milinkovic D, Jung T, Bartek B. Assessment of quality and readability of information provided by Chatgpt in relation to anterior cruciate ligament injury. *J Pers Med*. (2024) 14:104. doi: 10.3390/jpm14010104

55. Gabriel J, Shafik L, Alanbuki A, Larner T. The utility of the Chatgpt artificial intelligence tool for patient education and enquiry in robotic radical prostatectomy. *Int Urol Nephrol*. (2023) 55:2717–32. doi: 10.1007/s11255-023-03729-4

56. Gajjar AA, Kumar RP, Paliwoda ED, Kuo CC, Adida S, Legarreta AD, et al. Usefulness and accuracy of artificial intelligence Chatbot responses to patient questions for neurosurgical procedures. *Neurosurgery*. (2024) 95:171–178. doi: 10.1227/neu.0000000000002856

57. Garcia Valencia OA, Thongprayoon C, Miao J, Suppadungsuk S, Krisanapan P, Craici IM, et al. Empowering inclusivity: improving readability of living kidney donation information with Chatgpt. *Front Digit Health*. (2024) 6:1366967. doi: 10.3389/fdgth.2024.1366967

58. Ghanem D, Shu H, Bergstein V, Marrache M, Love A, Hughes A, et al. Educating patients on osteoporosis and bone health: can "Chatgpt" provide high-quality content? *Eur J Orthop Surg Traumatol*. (2024) 34:2757–65. doi: 10.1007/s00590-024-03990-y

59. Ghanem YK, Rouhi AD, Al-Houssan A, Saleh Z, Moccia MC, Joshi H, et al. Dr. Google to Dr. Chatgpt: assessing the content and quality of artificial intelligence-generated medical information on appendicitis. *Surg Endosc*. (2024) 38:2887–93. doi: 10.1007/s00464-024-10739-5

60. Gordon EB, Towbin AJ, Wingrove P, Shafique U, Haas B, Kitts AB, et al. Enhancing patient communication with chat-Gpt in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. *J Am Coll Radiol*. (2024) 21:353–9. doi: 10.1016/j.jacr.2023.09.011

61. Gül Ş, Erdemir İ, Hanci V, Aydoğmuş E, Erkoç YS. How artificial intelligence can provide information about subdural hematoma: assessment of readability, reliability, and quality of Chatgpt, Bard, and perplexity responses. *Medicine (Baltimore)*. (2024) 103:e38009. doi: 10.1097/MD.0000000000038009

62. Günay S, Yiğit Y, Halhalli HC, Tulgar S, Alkahlout BH, Azad AM. Ai in patient education: assessing the impact of Chatgpt-4 on conveying comprehensive information about chest pain. *Am J Emerg Med*. (2024) 77:220–1. doi: 10.1016/j.ajem.2023.12.047

63. Haidar O, Jaques A, McCaughran PW, Metcalfe MJ. Ai-generated information for vascular patients: assessing the standard of procedure-specific information provided by the Chatgpt Ai-language model. *Cureus*. (2023) 15:e49764. doi: 10.7759/cureus.49764

64. Halawani A, Mitchell A, Saffarzadeh M, Wong V, Chew BH, Forbes CM. Accuracy and readability of kidney stone patient information materials generated by a large language model compared to official urologic organizations. *Urology*. (2024) 186:107–13. doi: 10.1016/j.urology.2023.11.042

65. Hernandez CA, Vazquez Gonzalez AE, Polianovskaia A, Amoro Sanchez R, Muyolema Arce V, Mustafa A, et al. The future of patient education: Ai-driven guide for type 2 diabetes. *Cureus*. (2023) 15:e48919. doi: 10.7759/cureus.48919

66. Hershenhouse JS, Mokhtar D, Eppler MB, Rodler S, Storino Ramacciotti L, Ganjavi C, et al. Accuracy, readability, and understandability of large language models for prostate Cancer information to the public. *Prostate Cancer Prostatic Dis*. (2024). doi: 10.1038/s41391-024-00826-y

67. Hillmann HAK, Angelini E, Karfoul N, Feickert S, Mueller-Leisse J, Duncker D. Accuracy and comprehensibility of chat-based artificial intelligence for patient information on atrial fibrillation and cardiac implantable electronic devices. *Europace*. (2023) 26:euad369. doi: 10.1093/europace/euad369

68. Hirpara MM, Amin L, Aloyan T, Shilleh N, Lewis P. Does the internet provide quality information on metoidioplasty? Using the modified ensuring quality information for patients tool to evaluate artificial intelligence-generated and online information on metoidioplasty. *Ann Plast Surg*. (2024) 92:S361–5. doi: 10.1097/SAP.0000000000003797

69. Høj S, Thomsen SF, Meteran H, Sigsgaard T. Artificial intelligence and allergic rhinitis: does Chatgpt increase or impair the knowledge? *J Public Health (Oxf)*. (2024) 46:123–6. doi: 10.1093/pubmed/fdad219

70. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. Chatgpt vs Google for queries related to dementia and other cognitive decline: comparison of results. *J Med Internet Res*. (2023) 25:e48966. doi: 10.2196/48966

71. Ibrahim MT, Khaskheli SA, Shahzad H, Noordin S. Language-adaptive artificial intelligence: assessing Chatgpt's answer to frequently asked questions on Total hip arthroplasty questions. *J Pak Med Assoc*. (2024) 74:S161–4. doi: 10.47391/JPMA.AKU-9S-25

72. Jazi AHD, Mahjoubi M, Shahabi S, Alqahtani AR, Haddad A, Pazouki A, et al. Bariatric evaluation through Ai: a survey of expert opinions versus Chatgpt-4 (Beta-Seov). *Obes Surg*. (2023) 33:3971–80. doi: 10.1007/s11695-023-06903-w

73. Johns WL, Kellish A, Farronato D, Ciccotti MG, Hammoud S. Chatgpt can offer satisfactory responses to common patient questions regarding elbow ulnar collateral ligament reconstruction. *Arthrosc Sports Med Rehabil*. (2024) 6:100893. doi: 10.1016/j.asmr.2024.100893

74. Johnson CM, Bradley CS, Kenne KA, Rabice S, Takacs E, Vollstedt A, et al. Evaluation of Chatgpt for pelvic floor surgery counseling. *Urogynecology (Phila)*. (2024) 30:245–50. doi: 10.1097/SPV.0000000000001459

75. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The capability of Chatgpt in predicting and explaining common drug-drug interactions. *Cureus*. (2023) 15:e36272. doi: 10.7759/cureus.36272

76. Kasthuri VS, Glueck J, Pham H, Daher M, Balmaceno-Criss M, McDonald CL, et al. Assessing the accuracy and reliability of Ai-generated responses to patient questions regarding spine surgery. *J Bone Joint Surg Am*. (2024) 106:1136–42. doi: 10.2106/JBJS.23.00914

77. Kim MJ, Admane S, Chang YK, Shih KK, Reddy A, Tang M, et al. Chatbot performance in defining and differentiating palliative care, supportive care, hospice care. *J Pain Symptom Manage*. (2024) 67:e381–91. doi: 10.1016/j.jpainsymman.2024.01.008

78. King RC, Samaan JS, Yeo YH, Mody B, Lombardo DM, Ghashghaei R. Appropriateness of Chatgpt in answering heart failure related questions. *Heart Lung Circ*. (2024) 33:1314–8. doi: 10.1016/j.hlc.2024.03.005

79. King RC, Samaan JS, Yeo YH, Peng Y, Kunkel DC, Habib AA, et al. A multidisciplinary assessment of Chatgpt's knowledge of amyloidosis: observational study. *JMIR Cardio*. (2024) 8:e53421. doi: 10.2196/53421

80. Köroğlu EY, Fakı S, Beştepe N, Tam AA, Çuhacı Seyrek N, Topaloglu O, et al. A novel approach: evaluating Chatgpt's utility for the management of thyroid nodules. *Cureus*. (2023) 15:e47576. doi: 10.7759/cureus.47576

81. Kozaily E, Geagea M, Akdogan ER, Atkins J, Elshazly MB, Guglin M, et al. Accuracy and consistency of online large language model-based artificial intelligence chat platforms in answering Patients' questions about heart failure. *Int J Cardiol*. (2024) 408:132115. doi: 10.1016/j.ijcard.2024.132115

82. Kuckelman IJ, Wetley K, Yi PH, Ross AB. Translating musculoskeletal radiology reports into patient-friendly summaries using Chatgpt-4. *Skeletal Radiol*. (2024) 53:1621–4. doi: 10.1007/s00256-024-04599-2

83. Kuckelman IJ, Yi PH, Bui M, Onuh I, Anderson JA, Ross AB. Assessing Ai-powered patient education: a case study in radiology. *Acad Radiol*. (2024) 31:338–42. doi: 10.1016/j.acra.2023.08.020

84. Kuşcu O, Pamuk AE, Sütay Süslü N, Hosal S. Is Chatgpt accurate and reliable in answering questions regarding head and neck Cancer? *Front Oncol*. (2023) 13:1256459. doi: 10.3389/fonc.2023.1256459

85. Lambert R, Choo ZY, Gradwohl K, Schroedl L, Ruiz De Luzuriaga A. Assessing the application of large language models in generating dermatologic patient education materials according to Reading level: qualitative study. *JMIR Dermatol*. (2024) 7:e55898. doi: 10.2196/55898

86. Lang S, Vitale J, Fekete TF, Haschtmann D, Reitmeir R, Ropelato M, et al. Are large language models valid tools for patient information on lumbar disc herniation? The spine surgeons' perspective. *Brain Spine*. (2024) 4:102804. doi: 10.1016/j.bas.2024.102804

87. Lechien JR, Carroll TL, Huston MN, Naunheim MR. Chatgpt-4 accuracy for patient education in laryngopharyngeal reflux. *Eur Arch Otorrinolaringol*. (2024) 281:2547–52. doi: 10.1007/s00405-024-08560-w

88. Lee TJ, Campbell DJ, Patel S, Hossain A, Radfar N, Siddiqui E, et al. Unlocking health literacy: the ultimate guide to hypertension education from Chatgpt versus Google Gemini. *Cureus*. (2024) 16:e59898. doi: 10.7759/cureus.59898

89. Lee TJ, Rao AK, Campbell DJ, Radfar N, Dayal M, Khrais A. Evaluating Chatgpt-3.5 and Chatgpt-4.0 responses on hyperlipidemia for patient education. *Cureus*. (2024) 16:e61067. doi: 10.7759/cureus.61067

90. Li L, Li P, Wang K, Zhang L, Zhao H, Ji H. Benchmarking state-of-the-art large language models for migraine patient education: a comparison of performances on the responses to common queries. *J Med Internet Res*. (2024) 26:e55927. doi: 10.2196/55927

91. Li W, Chen J, Chen F, Liang J, Yu H. Exploring the potential of Chatgpt-4 in responding to common questions about Abdominoplasty: An Ai-based case study of a plastic surgery consultation. *Aesth Plast Surg*. (2024) 48:1571–83. doi: 10.1007/s00266-023-03660-0

92. Lim B, Seth I, Kah S, Sofiadellis F, Ross RJ, Rozen WM, et al. Using generative artificial intelligence tools in cosmetic surgery: a study on Rhinoplasty, facelifts, and blepharoplasty procedures. *J Clin Med*. (2023) 12:6524. doi: 10.3390/jcm12206524

93. Liu HY, Alessandri Bonetti M, De Lorenzi F, Gimbel ML, Nguyen VT, Egro FM. Consulting the digital doctor: Google versus Chatgpt as sources of information on breast implant-associated anaplastic large cell lymphoma and breast implant illness. *Aesth Plast Surg*. (2024) 48:590–607. doi: 10.1007/s00266-023-03713-4

94. Liu S, McCoy AB, Wright AP, Carew B, Genkins JZ, Huang SS, et al. Leveraging large language models for generating responses to patient messages-a subjective analysis. *J Am Med Inform Assoc*. (2024) 31:1367–79. doi: 10.1093/jamia/ocae052

95. Lv X, Zhang X, Li Y, Ding X, Lai H, Shi J. Leveraging large language models for improved patient access and self-management: Assessor-blinded comparison between expert-and Ai-generated content. *J Med Internet Res*. (2024) 26:e55847. doi: 10.2196/55847

96. Mashatian S, Armstrong DG, Ritter A, Robbins J, Aziz S, Alenabi I, et al. Building trustworthy generative artificial intelligence for diabetes care and limb preservation: a medical knowledge extraction case. *J Diabetes Sci Technol*. (2024):19322968241253568. doi: 10.1177/19322968241253568

97. Mastrokostas PG, Mastrokostas LE, Emara AK, Wellington IJ, Ginalis E, Houten JK, et al. Gpt-4 as a source of patient information for anterior cervical discectomy and fusion: a comparative analysis against Google web search. Global. *Spine J*. (2024):21925682241241241. doi: 10.1177/21925682241241241

98. McCarthy CJ, Berkowitz S, Ramalingam V, Ahmed M. Evaluation of an artificial intelligence Chatbot for delivery of Ir patient education material: a comparison with societal website content. *J Vasc Interv Radiol*. (2023) 34:1760–8.e32. doi: 10.1016/j.jvir.2023.05.037

99. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing Chatgpt responses to common patient questions regarding Total hip arthroplasty. *J Bone Joint Surg Am*. (2023) 105:1519–26. doi: 10.2106/JBJS.23.00209

100. Mika AP, Mulvey HE, Engstrom SM, Polkowski GG, Martin JR, Wilson JM. Can Chatgpt answer patient questions regarding Total knee arthroplasty? *J Knee Surg*. (2024) 37:664–73. doi: 10.1055/s-0044-1782233

101. Mishra V, Sarraju A, Kalwani NM, Dexter JP. Evaluation of prompts to simplify cardiovascular disease information generated using a large language model: cross-sectional study. *J Med Internet Res*. (2024) 26:e55388. doi: 10.2196/55388

102. Moazzam Z, Lima HA, Endo Y, Noria S, Needleman B, Pawlik TM. A paradigm shift: online artificial intelligence platforms as an informational resource in bariatric surgery. *Obes Surg*. (2023) 33:2611–4. doi: 10.1007/s11695-023-06675-3

103. Moise A, Centomo-Bozzo A, Orishchak O, Alnoury MK, Daniel SJ. Can Chatgpt guide parents on Tympanostomy tube insertion? *Children (Basel)*. (2023) 10:1634. doi: 10.3390/children10101634

104. Mondal H, Mondal S, Podder I. Using Chatgpt for writing articles for Patients' education for dermatological diseases: a pilot study. *Indian Dermatol Online J*. (2023) 14:482–6. doi: 10.4103/idoj.idoj_72_23

105. Mondal H, Panigrahi M, Mishra B, Behera JK, Mondal S. A pilot study on the capability of artificial intelligence in preparation of Patients' educational materials for Indian public health issues. *J Family Med Prim Care*. (2023) 12:1659–62. doi: 10.4103/jfmpc.jfmpc_262_23

106. Monroe CL, Abdelhafez YG, Atsina K, Aman E, Nardo L, Madani MH. Evaluation of responses to cardiac imaging questions by the artificial intelligence large language model Chatgpt. *Clin Imaging*. (2024) 112:110193. doi: 10.1016/j.clinimag.2024.110193

107. Mootz AA, Carvalho B, Sultan P, Nguyen TP, Reale SC. The accuracy of Chatgpt-generated responses in answering commonly asked patient questions about labor epidurals: a survey-based study. *Anesth Analg*. (2024) 138:1142–4. doi: 10.1213/ANE.0000000000006801

108. Munir MM, Endo Y, Ejaz A, Dillhoff M, Cloyd JM, Pawlik TM. Online artificial intelligence platforms and their applicability to gastrointestinal surgical operations. *J Gastrointest Surg*. (2024) 28:64–9. doi: 10.1016/j.gassur.2023.11.019

109. Musheyev D, Pan A, Loeb S, Kabarriti AE. How well Do artificial intelligence Chatbots respond to the top search queries about urological malignancies? *Eur Urol*. (2024) 85:13–6. doi: 10.1016/j.eururo.2023.07.004

110. Nov O, Singh N, Mann D. Putting Chatgpt's medical advice to the (Turing) test: survey study. *JMIR Med Educ*. (2023) 9:e46939. doi: 10.2196/46939

111. O'Hagan R, Kim RH, Abittan BJ, Caldas S, Ungar J, Ungar B. Trends in accuracy and appropriateness of alopecia Areata information obtained from a popular online large language model, Chatgpt. *Dermatology*. (2023) 239:952–7. doi: 10.1159/000534005

112. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence Chatbot responses to top searched queries about Cancer. *JAMA Oncol*. (2023) 9:1437–40. doi: 10.1001/jamaoncol.2023.2947

113. Parekh AS, McCahon JAS, Nghe A, Pedowitz DI, Daniel JN, Parekh SG. Foot and ankle patient education materials and artificial intelligence Chatbots: a comparative analysis. *Foot Ankle Spec*. (2024):19386400241235834. doi: 10.1177/19386400241235834

114. Pohl NB, Derector E, Rivlin M, Bachoura A, Tosti R, Kachooei AR, et al. A quality and readability comparison of artificial intelligence and popular health website education materials for common hand surgery procedures. *Hand Surg Rehabil*. (2024) 43:101723. doi: 10.1016/j.hansur.2024.101723

115. Potapenko I, Malmqvist L, Subhi Y, Hamann S. Artificial intelligence-based Chatgpt responses for patient questions on optic disc Drusen. *Ophthalmol Ther*. (2023) 12:3109–19. doi: 10.1007/s40123-023-00800-2

116. Pradhan F, Fiedler A, Samson K, Olivera-Martinez M, Manatsathit W, Peeraphatdit T. Artificial intelligence compared with human-derived patient educational materials on cirrhosis. *Hepatol Commun*. (2024) 8:e0367. doi: 10.1097/HC9.0000000000000367

117. Rahimli Ocakoglu S, Coskun B. The emerging role of Ai in patient education: a comparative analysis of Llm accuracy for pelvic organ prolapse. *Med Princ Pract*. (2024) 33:330–7. doi: 10.1159/000538538

118. Razdan S, Siegal AR, Brewer Y, Sljivich M, Valenzuela RJ. Assessing Chatgpt's ability to answer questions pertaining to erectile dysfunction: can our patients trust it? *Int J Impot Res*. (2023). doi: 10.1038/s41443-023-00797-z

119. Reichenpfader D, Rösslhuemer P, Denecke K. Large language model-based evaluation of medical question answering systems: algorithm development and case study. *Stud Health Technol Inform*. (2024) 313:22–7. doi: 10.3233/SHTI240006

120. Roster K, Kann RB, Farabi B, Gronbeck C, Brownstone N, Lipner SR. Readability and health literacy scores for Chatgpt-generated dermatology public education materials: cross-sectional analysis of sunscreen and melanoma questions. *JMIR Dermatol*. (2024) 7:e50163. doi: 10.2196/50163

121. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the accuracy of responses by the language model Chatgpt to questions regarding bariatric surgery. *Obes Surg*. (2023) 33:1790–6. doi: 10.1007/s11695-023-06603-5

122. Şan H, Bayrakçi Ö, Çağdaş B, Serdengeçti M, Alagöz E. Reliability and readability analysis of Gpt-4 and Google Bard as a patient information source for the Most commonly applied radionuclide treatments in Cancer patients. *Rev Esp Med Nucl Imagen Mol (Engl Ed)*. (2024):500021. doi: 10.1016/j.remnie.2024.500021

123. Sciberras M, Farrugia Y, Gordon H, Furfaro F, Allocca M, Torres J, et al. Accuracy of information given by Chatgpt for patients with inflammatory bowel disease in relation to Ecco guidelines. *J Crohns Colitis*. (2024) 18:1215–21. doi: 10.1093/ecco-jcc/jjae040

124. Şenoymak MC, Erbatur NH, Şenoymak İ, Fırat SN. The role of artificial intelligence in endocrine management: assessing Chatgpt's responses to Prolactinoma queries. *J Pers Med*. (2024) 14:330. doi: 10.3390/jpm14040330

125. Seth I, Cox A, Xie Y, Bulloch G, Hunter-Smith DJ, Rozen WM, et al. Evaluating Chatbot efficacy for answering frequently asked questions in plastic surgery: a Chatgpt case study focused on breast augmentation. *Aesthet Surg J*. (2023) 43:1126–35. doi: 10.1093/asj/sjad140

126. Shah YB, Ghosh A, Hochberg AR, Rapoport E, Lallas CD, Shah MS, et al. Comparison of Chatgpt and traditional patient education materials for Men's health. *Urol Pract*. (2024) 11:87–94. doi: 10.1097/UPJ.0000000000000490

127. Shen SA, Perez-Heydrich CA, Xie DX, Nellis JC. Chatgpt vs. web search for patient questions: what does Chatgpt Do better? *Eur Arch Otorrinolaringol*. (2024) 281:3219–25. doi: 10.1007/s00405-024-08524-0

128. Shiraishi M, Lee H, Kanayama K, Moriwaki Y, Okazaki M. Appropriateness of artificial intelligence Chatbots in diabetic foot ulcer management. *Int J Low Extrem Wounds*. (2024):15347346241236811. doi: 10.1177/15347346241236811

129. Song H, Xia Y, Luo Z, Liu H, Song Y, Zeng X, et al. Evaluating the performance of different large language models on health consultation and patient education in Urolithiasis. *J Med Syst*. (2023) 47:125. doi: 10.1007/s10916-023-02021-3

130. Spallek S, Birrell L, Kershaw S, Devine EK, Thornton L. Can we use Chatgpt for mental health and substance use education? Examining its quality and potential harms. *JMIR Med Educ*. (2023) 9:e51243. doi: 10.2196/51243

131. Srinivasan N, Samaan JS, Rajeev ND, Kanu MU, Yeo YH, Samakar K. Large language models and bariatric surgery patient education: a comparative readability analysis of Gpt-3.5, Gpt-4, Bard, and online institutional resources. *Surg Endosc*. (2024) 38:2522–32. doi: 10.1007/s00464-024-10720-2

132. Subramanian T, Araghi K, Amen TB, Kaidi A, Sosa B, Shahi P, et al. Chat generative Pretraining transformer answers patient-focused questions in cervical spine surgery. *Clin Spine Surg*. (2024) 37:E278–81. doi: 10.1097/BSD.0000000000001600

133. Tailor PD, Dalvin LA, Chen JJ, Iezzi R, Olsen TW, Scruggs BA, et al. A comparative study of responses to retina questions from either experts, expert-edited large language models, or expert-edited large language models alone. *Ophthalmol Sci*. (2024) 4:100485. doi: 10.1016/j.xops.2024.100485

134. Tailor PD, Xu TT, Fortes BH, Iezzi R, Olsen TW, Starr MR, et al. Appropriateness of ophthalmology recommendations from an online chat-based artificial intelligence model. *Mayo Clin Proc Digit Health*. (2024) 2:119–28. doi: 10.1016/j.mcpdig.2024.01.003

135. Tao BK, Handzic A, Hua NJ, Vosoughi AR, Margolin EA, Micieli JA. Utility of Chatgpt for automated creation of patient education handouts: An application in neuro-ophthalmology. *J Neuroophthalmol*. (2024) 44:119–24. doi: 10.1097/WNO.0000000000002074

136. WLT T, Cheng R, Weinblatt AI, Bergstein V, Long WJ. An artificial intelligence Chatbot is an accurate and useful online patient resource prior to Total knee arthroplasty. *J Arthroplast*. (2024) 39:S358–62. doi: 10.1016/j.arth.2024.02.005

137. Tepe M, Emekli E. Assessing the responses of large language models (Chatgpt-4, Gemini, and Microsoft Copilot) to frequently asked questions in breast imaging: a study on readability and accuracy. *Cureus*. (2024) 16:e59960. doi: 10.7759/cureus.59960

138. Tharakan S, Klein B, Bartlett L, Atlas A, Parada SA, Cohn RM. Do Chatgpt and Google differ in answers to commonly asked patient questions regarding Total shoulder and Total elbow arthroplasty? *J Shoulder Elb Surg*. (2024) 33:e429–37. doi: 10.1016/j.jse.2023.11.014

139. Thia I, Saluja M. Chatgpt: is this patient education tool for urological malignancies readable for the general population? *Res Rep Urol*. (2024) 16:31–7. doi: 10.2147/RRU.S440633

140. Van Bulck L, Moons P. What if your patient switches from Dr. Google to Dr. Chatgpt? A vignette-based survey of the trustworthiness, value, and danger of Chatgpt-generated responses to health questions. *Eur J Cardiovasc Nurs*. (2024) 23:95–8. doi: 10.1093/eurjcn/zvad038

141. Washington CJ, Abouyared M, Karanth S, Braithwaite D, Birkeland A, Silverman DA, et al. The use of Chatbots in head and neck mucosal malignancy treatment recommendations. *Otolaryngol Head Neck Surg*. (2024) 171:1062–8. doi: 10.1002/ohn.818

142. Wei K, Fritz C, Rajasekaran K. Answering head and neck Cancer questions: An assessment of Chatgpt responses. *Am J Otolaryngol*. (2024) 45:104085. doi: 10.1016/j.amjoto.2023.104085

143. Wrenn SP, Mika AP, Ponce RB, Mitchell PM. Evaluating Chatgpt's ability to answer common patient questions regarding hip fracture. *J Am Acad Orthop Surg*. (2024) 32:656–9. doi: 10.5435/JAAOS-D-23-00877

144. Wright BM, Bodnar MS, Moore AD, Maseda MC, Kucharik MP, Diaz CC, et al. Is Chatgpt a trusted source of information for Total hip and knee arthroplasty patients? *Bone Jt Open*. (2024) 5:139–46. doi: 10.1302/2633-1462.52.BJO-2023-0113.R1

145. Wu G, Zhao W, Wong A, Lee DA. Patients with floaters: answers from virtual assistants and large language models. *Digit Health*. (2024) 10:20552076241229933. doi: 10.1177/20552076241229933

146. Wu Y, Zhang Z, Dong X, Hong S, Hu Y, Liang P, et al. Evaluating the performance of the language model Chatgpt in responding to common questions of people with epilepsy. *Epilepsy Behav*. (2024) 151:109645. doi: 10.1016/j.yebeh.2024.109645

147. Yalla GR, Hyman N, Hock LE, Zhang Q, Shukla AG, Kolomeyer NN. Performance of artificial intelligence Chatbots on Glaucoma questions adapted from patient brochures. *Cureus*. (2024) 16:e56766. doi: 10.7759/cureus.56766

148. Yan S, Du D, Liu X, Dai Y, Kim MK, Zhou X, et al. Assessment of the reliability and clinical applicability of Chatgpt's responses to Patients' common queries about Rosacea. *Patient Prefer Adherence*. (2024) 18:249–53. doi: 10.2147/PPA.S444928

149. Yan SY, Liu YF, Ma L, Xiao LL, Hu X, Guo R, et al. Walking forward or on hold: could the Chatgpt be applied for seeking health information in neurosurgical settings? *Ibrain*. (2024) 10:111–5. doi: 10.1002/ibra.12149

150. Ye C, Zweck E, Ma Z, Smith J, Katz S. Doctor versus artificial intelligence: patient and physician evaluation of large language model responses to rheumatology patient questions in a cross-sectional study. *Arthritis Rheumatol*. (2024) 76:479–84. doi: 10.1002/art.42737

151. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of Chatgpt in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. (2023) 29:721–32. doi: 10.3350/cmh.2023.0089

152. Yılmaz IBE, Doğan L. Talking technology: exploring Chatbots as a tool for cataract patient education. *Clin Exp Optom*. (2024):1–9. doi: 10.1080/08164622.2023.2298812

153. Yüce A, Erkurt N, Yerli M, Misir A. The potential of Chatgpt for high-quality information in patient education for sports surgery. *Cureus*. (2024) 16:e58874. doi: 10.7759/cureus.58874

154. Yun JY, Kim DJ, Lee N, Kim EK. A comprehensive evaluation of Chatgpt consultation quality for augmentation mammoplasty: a comparative analysis between plastic surgeons and laypersons. *Int J Med Inform*. (2023) 179:105219. doi: 10.1016/j.ijmedinf.2023.105219

155. Zalzal HG, Abraham A, Cheng J, Shah RK. Can ChatGPT help patients answer their otolaryngology questions? *Laryngoscope Investig Otolaryngol*. (2024) 9:e1193. doi: 10.1002/lio2.1193

156. Zhang S, Liau ZQG, Tan KLM, Chua WL. Evaluating the accuracy and relevance of Chatgpt responses to frequently asked questions regarding total knee replacement. *Knee Surg Relat Res*. (2024) 36:15. doi: 10.1186/s43019-024-00218-5

157. Zhang Y, Dong Y, Mei Z, Hou Y, Wei M, Yeung YH, et al. Performance of large language models on benign prostatic hyperplasia frequently asked questions. *Prostate*. (2024) 84:807–13. doi: 10.1002/pros.24699

158. Abreu AA, Murimwa GZ, Farah E, Stewart JW, Zhang L, Rodriguez J, et al. Enhancing readability of online patient-facing content: the role of Ai Chatbots in improving Cancer information accessibility. *J Natl Compr Cancer Netw*. (2024) 22:e237334. doi: 10.6004/jnccn.2023.7334

159. Ayre J, Mac O, McCaffery K, McKay BR, Liu M, Shi Y, et al. New Frontiers in health literacy: using Chatgpt to simplify health information for people in the community. *J Gen Intern Med*. (2024) 39:573–7. doi: 10.1007/s11606-023-08469-w

160. Baldwin AJ. An artificial intelligence language model improves readability of burns first aid information. *Burns*. (2024) 50:1122–7. doi: 10.1016/j.burns.2024.03.005

161. Browne R, Gull K, Hurley CM, Sugrue RM, O'Sullivan JB. Chatgpt-4 can help hand surgeons communicate better with patients. *J Hand Surg Glob Online*. (2024) 6:436–8. doi: 10.1016/j.jhsg.2024.03.008

162. Covington EW, Watts Alexander CS, Sewell J, Hutchison AM, Kay J, Tocco L, et al. Unlocking the future of patient education: Chatgpt vs. Lexicomp® as sources of patient education materials. *J Am Pharm Assoc (2003)*. (2024):102119. doi: 10.1016/j.japh.2024.102119

163. Dihan Q, Chauhan MZ, Eleiwa TK, Hassan AK, Sallam AB, Khouri AS, et al. Using large language models to generate educational materials on childhood Glaucoma. *Am J Ophthalmol*. (2024) 265:28–38. doi: 10.1016/j.ajo.2024.04.004

164. Eid K, Eid A, Wang D, Raiker RS, Chen S, Nguyen J. Optimizing ophthalmology patient education via Chatbot-generated materials: readability analysis of Ai-generated patient education materials and the American Society of Ophthalmic Plastic and Reconstructive Surgery Patient Brochures. *Ophthalmic Plast Reconstr Surg*. (2024) 40:212–6. doi: 10.1097/IOP.0000000000002549

165. Eppler MB, Ganjavi C, Knudsen JE, Davis RJ, Ayo-Ajibola O, Desai A, et al. Bridging the gap between urological research and patient understanding: the role of large language models in automated generation of Layperson's summaries. *Urol Pract*. (2023) 10:436–43. doi: 10.1097/UPJ.0000000000000428

166. Fanning JE, Escobar-Domingo MJ, Foppiani J, Lee D, Miller AS, Janis JE, et al. Improving readability and automating content analysis of plastic surgery webpages with Chatgpt. *J Surg Res*. (2024) 299:103–11. doi: 10.1016/j.jss.2024.04.006

167. Hung YC, Chaker SC, Sigel M, Saad M, Slater ED. Comparison of patient education materials generated by chat generative pre-trained transformer versus experts: An innovative way to increase readability of patient education materials. *Ann Plast Surg*. (2023) 91:409–12. doi: 10.1097/SAP.0000000000003634

168. Kirchner GJ, Kim RY, Weddle JB, Bible JE. Can artificial intelligence improve the readability of patient education materials? *Clin Orthop Relat Res*. (2023) 481:2260–7. doi: 10.1097/CORR.0000000000002668

169. Moons P, Van Bulck L. Using Chatgpt and Google Bard to improve the readability of written patient information: a proof of concept. *Eur J Cardiovasc Nurs*. (2024) 23:122–6. doi: 10.1093/eurjcn/zvad087

170. Patel EA, Fleischer L, Filip P, Eggerstedt M, Hutz M, Michaelides E, et al. The use of artificial intelligence to improve readability of otolaryngology patient education materials. *Otolaryngol Head Neck Surg*. (2024) 171:603–8. doi: 10.1002/ohn.816

171. Rouhi AD, Ghanem YK, Yolchieva L, Saleh Z, Joshi H, Moccia MC, et al. Can artificial intelligence improve the readability of patient education materials on aortic stenosis? A pilot study. *Cardiol Ther*. (2024) 13:137–47. doi: 10.1007/s40119-023-00347-0

172. Sridharan K, Sivaramakrishnan G. Enhancing readability of Usfda patient communications through large language models: a proof-of-concept study. *Expert Rev Clin Pharmacol*. (2024) 17:731–41. doi: 10.1080/17512433.2024.2363840

173. Sudharshan R, Shen A, Gupta S, Zhang-Nunes S. Assessing the utility of Chatgpt in simplifying text complexity of patient educational materials. *Cureus*. (2024) 16:e55304. doi: 10.7759/cureus.55304

174. Vallurupalli M, Shah ND, Vyas RM. Validation of Chatgpt 3.5 as a tool to optimize readability of patient-facing craniofacial education materials. *Plast Reconstr Surg Glob Open*. (2024) 12:e5575. doi: 10.1097/GOX.0000000000005575

175. Grimm DR, Lee YJ, Hu K, Liu L, Garcia O, Balakrishnan K, et al. The utility of Chatgpt as a generative medical translator. *Eur Arch Otorrinolaringol*. (2024). doi: 10.1007/s00405-024-08708-8

176. He Z, Bhasuran B, Jin Q, Tian S, Hanna K, Shavor C, et al. Quality of answers of generative large language models versus peer users for interpreting laboratory test results for lay patients: evaluation study. *J Med Internet Res*. (2024) 26:e56655. doi: 10.2196/56655

177. Meyer A, Soleman A, Riese J, Streichert T. Comparison of Chatgpt, Gemini, and Le chat with physician interpretations of medical laboratory questions from an online health forum. *Clin Chem Lab Med*. (2024) 62:2425–34. doi: 10.1515/cclm-2024-0246

178. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using Chatgpt and Gpt-4 with prompt learning: results, limitations, and potential. *Vis Comput Ind Biomed Art*. (2023) 6:9. doi: 10.1186/s42492-023-00136-5

179. Sarangi PK, Lumbani A, Swarup MS, Panda S, Sahoo SS, Hui P, et al. Assessing Chatgpt's proficiency in simplifying radiological reports for healthcare professionals and patients. *Cureus*. (2023) 15:e50881. doi: 10.7759/cureus.50881

180. Rogasch JMM, Metzger G, Preisler M, Galler M, Thiele F, Brenner W, et al. Chatgpt: can You prepare my patients for [(18) F] Fdg pet/Ct and explain my reports? *J Nucl Med*. (2023) 64:1876–9. doi: 10.2967/jnumed.123.266114

181. Tepe M, Emekli E. Decoding medical jargon: the use of Ai language models (Chatgpt-4, Bard, Microsoft Copilot) in radiology reports. *Patient Educ Couns*. (2024) 126:108307. doi: 10.1016/j.pec.2024.108307

182. Woo KC, Simon GW, Akindutire O, Aphinyanaphongs Y, Austrian JS, Kim JG, et al. Evaluation of Gpt-4 ability to identify and generate patient instructions for actionable incidental radiology findings. *J Am Med Inform Assoc*. (2024) 31:1983–93. doi: 10.1093/jamia/ocae117

183. Zaretsky J, Kim JM, Baskharoun S, Zhao Y, Austrian J, Aphinyanaphongs Y, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA Netw Open*. (2024) 7:e240357. doi: 10.1001/jamanetworkopen.2024.0357

184. Alanezi F. Examining the role of Chatgpt in promoting health behaviors and lifestyle changes among Cancer patients. *Nutr Health*. (2024):2601060241244563. doi: 10.1177/02601060241244563

185. Bragazzi NL, Garbarino S. Assessing the accuracy of generative conversational artificial intelligence in debunking sleep health myths: mixed methods comparative study with expert analysis. *JMIR Form Res*. (2024) 8:e55762. doi: 10.2196/55762

186. Garbarino S, Bragazzi NL. Evaluating the effectiveness of artificial intelligence-based tools in detecting and understanding sleep health misinformation: comparative analysis using Google Bard and Openai Chatgpt-4. *J Sleep Res*. (2024):e14210. doi: 10.1111/jsr.14210

187. Gray M, Baird A, Sawyer T, James J, DeBroux T, Bartlett M, et al. Increasing realism and variety of virtual patient dialogues for prenatal counseling education through a novel application of Chatgpt: exploratory observational study. *JMIR Med Educ*. (2024) 10:e50705. doi: 10.2196/50705

188. Minutolo A, Damiano E, De Pietro G, Fujita H, Esposito M. A conversational agent for querying Italian patient information leaflets and improving health literacy. *Comput Biol Med*. (2022) 141:105004. doi: 10.1016/j.compbiomed.2021.105004

189. Mondal H, Dash I, Mondal S, Behera JK. Chatgpt in answering queries related to lifestyle-related diseases and disorders. *Cureus*. (2023) 15:e48296. doi: 10.7759/cureus.48296

190. Ponzo V, Goitre I, Favaro E, Merlo FD, Mancino MV, Riso S, et al. Is Chatgpt an effective tool for providing dietary advice? *Nutrients*. (2024) 16:469. doi: 10.3390/nu16040469

191. Zaleski AL, Berkowsky R, Craig KJT, Pescatello LS. Comprehensiveness, accuracy, and readability of exercise recommendations provided by an Ai-based Chatbot: mixed methods study. *JMIR Med Educ*. (2024) 10:e51308. doi: 10.2196/51308

192. Altamimi I, Altamimi A, Alhumimidi AS, Temsah MH. Snakebite advice and counseling from artificial intelligence: An acute venomous snakebite consultation with Chatgpt. *Cureus*. (2023) 15:e40351. doi: 10.7759/cureus.40351

193. McMahon HV, McMahon BD. Automating untruths: Chatgpt, self-managed medication abortion, and the threat of misinformation in a post-roe world. *Front Digit Health*. (2024) 6:1287186. doi: 10.3389/fdgth.2024.1287186

194. Aliyeva A, Sari E, Alaskarov E, Nasirov R. Enhancing postoperative Cochlear implant care with Chatgpt-4: a study on artificial intelligence (Ai)-assisted patient education and support. *Cureus*. (2024) 16:e53897. doi: 10.7759/cureus.53897

195. Scheschenja M, Viniol S, Bastian MB, Wessendorf J, König AM, Mahnken AH. Feasibility of Gpt-3 and Gpt-4 for in-depth patient education prior to interventional radiological procedures: a comparative analysis. *Cardiovasc Intervent Radiol*. (2024) 47:245–50. doi: 10.1007/s00270-023-03563-2

196. Bains SS, Dubin JA, Hameed D, Sax OC, Douglas S, Mont MA, et al. Use and application of large language models for patient questions following Total knee arthroplasty. *J Arthroplast*. (2024) 39:2289–94. doi: 10.1016/j.arth.2024.03.017

197. Borna S, Gomez-Cabello CA, Pressman SM, Haider SA, Sehgal A, Leibovich BC, et al. Comparative analysis of artificial intelligence virtual assistant and large language models in post-operative care. *Eur J Investig Health Psychol Educ*. (2024) 14:1413–24. doi: 10.3390/ejihpe14050093

198. Capelleras M, Soto-Galindo GA, Cruellas M, Apaydin F. Chatgpt and Rhinoplasty recovery: An exploration of Ai's role in postoperative guidance. *Facial Plast Surg*. (2024) 40:628–31. doi: 10.1055/a-2219-4901

199. Chaker SC, Hung YC, Saad M, Golinko MS, Galdyn IA. Easing the burden on caregivers-applications of artificial intelligence for physicians and caregivers of children with cleft lip and palate. *Cleft Palate Craniofac J*. (2024):10556656231223596. doi: 10.1177/10556656231223596

200. Shao CY, Li H, Liu XL, Li C, Yang LQ, Zhang YJ, et al. Appropriateness and comprehensiveness of using Chatgpt for perioperative patient education in thoracic surgery in different language contexts: survey study. *Interact J Med Res*. (2023) 12:e46900. doi: 10.2196/46900

201. Lee JC, Hamill CS, Shnayder Y, Buczek E, Kakarala K, Bur AM. Exploring the role of artificial intelligence Chatbots in preoperative counseling for head and neck Cancer surgery. *Laryngoscope*. (2024) 134:2757–61. doi: 10.1002/lary.31243

202. Nanji K, Yu CW, Wong TY, Sivaprasad S, Steel DH, Wykoff CC, et al. Evaluation of postoperative ophthalmology patient instructions from Chatgpt and Google search. *Can J Ophthalmol*. (2024) 59:e69–71. doi: 10.1016/j.jcjo.2023.10.001

203. Dhar S, Kothari D, Vasquez M, Clarke T, Maroda A, McClain WG, et al. The utility and accuracy of Chatgpt in providing post-operative instructions following tonsillectomy: a pilot study. *Int J Pediatr Otorhinolaryngol*. (2024) 179:111901. doi: 10.1016/j.ijporl.2024.111901

204. Patil NS, Huang R, Mihalache A, Kisilevsky E, Kwok J, Popovic MM, et al. The ability of artificial intelligence Chatbots Chatgpt and Google Bard to accurately convey preoperative information for patients undergoing ophthalmic surgeries. *Retina*. (2024) 44:950–3. doi: 10.1097/IAE.0000000000004044

205. Meyer R, Hamilton KM, Truong MD, Wright KN, Siedhoff MT, Brezinov Y, et al. Chatgpt compared with Google search and healthcare institution as sources of postoperative patient instructions after gynecological surgery. *BJOG*. (2024) 131:1154–6. doi: 10.1111/1471-0528.17746

206. Breneman A, Gordon ER, Trager MH, Ensslin CJ, Fisher J, Humphreys TR, et al. Evaluation of large language model responses to Mohs surgery preoperative questions. *Arch Dermatol Res*. (2024) 316:227. doi: 10.1007/s00403-024-02956-8

207. Kienzle A, Niemann M, Meller S, Gwinner C. Chatgpt may offer an adequate substitute for informed consent to patients prior to Total knee arthroplasty-yet caution is needed. *J Pers Med*. (2024) 14:69. doi: 10.3390/jpm14010069

208. Ali R, Connolly ID, Tang OY, Mirza FN, Johnston B, Abdulrazeq HF, et al. Bridging the literacy gap for surgical consents: An Ai-human expert collaborative approach. *NPJ Digit Med*. (2024) 7:63. doi: 10.1038/s41746-024-01039-2

209. Shiraishi M, Tomioka Y, Miyakuni A, Moriwaki Y, Yang R, Oba J, et al. Generating informed consent documents related to blepharoplasty using Chatgpt. *Ophthalmic Plast Reconstr Surg*. (2024) 40:316–20. doi: 10.1097/IOP.0000000000002574

210. An Y, Fang Q, Wang L. Enhancing patient education in Cancer care: intelligent Cancer patient education model for effective communication. *Comput Biol Med*. (2024) 169:107874. doi: 10.1016/j.compbiomed.2023.107874

211. Roberts RHR, Ali SR, Dobbs TD, Whitaker IS. Can large language models generate outpatient clinic letters at first consultation that incorporate complication profiles from Uk and USA aesthetic plastic surgery associations? *Aesthet Surg J Open Forum*. (2024) 6:ojad 109. doi: 10.1093/asjof/ojad109

212. Xue Z, Zhang Y, Gan W, Wang H, She G, Zheng X. Quality and dependability of Chatgpt and Dingxiangyuan forums for remote orthopedic consultations: comparative analysis. *J Med Internet Res*. (2024) 26:e50882. doi: 10.2196/50882

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 30 October 2024
DOI 10.3389/fmed.2024.1469227

OPEN ACCESS

EDITED BY
Ariel Soares Teles,
Federal Institute of Education, Science, and Technology of Maranhão, Brazil

REVIEWED BY
Kabilan Elangovan,
Singapore Health Services, Singapore
Livia Ruback,
State University of Campinas, Brazil
Tian Wang,
The University of Sydney, Australia

*CORRESPONDENCE
Weiguo Hu
✉ wghu@rjh.com.cn
Qisheng Lin
✉ linqisheng074@hotmail.com
Xiaoyang Li
✉ woodslee429@126.com

†These authors have contributed equally to this work

RECEIVED 23 July 2024
ACCEPTED 09 October 2024
PUBLISHED 30 October 2024

CITATION
Jin H, Huang L, Ye J, Wang J, Lin X, Wu S, Hu W, Lin Q and Li X (2024) Enhancing nutritional management in peritoneal dialysis patients through a generative pre-trained transformers-based recipe generation tool: a pilot study.
Front. Med. 11:1469227.
doi: 10.3389/fmed.2024.1469227

Enhancing nutritional management in peritoneal dialysis patients through a generative pre-trained transformers-based recipe generation tool: a pilot study

Haijiao Jin[1,2,3,4,5†], Lulu Huang[2†], Jinling Ye[2], Jinkun Wang[2], Xinghui Lin[1,2,3,4,5], Shaun Wu[6], Weiguo Hu[7*], Qisheng Lin[1,3,4,5*] and Xiaoyang Li[7*]

[1]Department of Nephrology, Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [2]Department of Nephrology, Ningbo Hangzhou Bay Hospital, Zhejiang, China, [3]Molecular Cell Lab for Kidney Disease, Shanghai, China, [4]Shanghai Peritoneal Dialysis Research Center, Shanghai, China, [5]Uremia Diagnosis and Treatment Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [6]WORK Medical Technology Group LTD., Hangzhou, China, [7]Department of Medical Education, Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China

Background: Patients undergoing peritoneal dialysis (PD) often face nutritional deficiencies due to inadequate intake, nutrient loss, insufficient dialysis, and a state of micro-inflammatory. Traditional nutritional management methods have not fully met personalized needs. Therefore, this study aimed to develop and evaluate an application for generating recipes based on Generative Pre-trained Transformers to improve the nutritional status of these patients.

Methods: This self-controlled prospective study included 35 patients undergoing PD from January to February 2024. The study was divided into two phases: the initial phase involved conventional dietary education under PD management, followed by a second phase where a new GPT-based dietary guidance tool was introduced. Patients adhered to the diets recommended by the tool. Nutritional intervention effects were assessed by comparing serum prealbumin, albumin, and phosphate levels before and after the intervention.

Results: After the intervention, the mean prealbumin levels significantly improved from 289.04 ± 74.60 mg/L to 326.72 ± 78.89 mg/L ($p = 0.001$). Although there was no statistical significance, the serum albumin levels in patients increased from 34.70 ± 5.94 g/L to 35.66 ± 5.14 g/L ($p = 0.153$). Serum phosphate levels remained stable and within safe limits ($p = 0.241$).

Conclusion: The AI-based recipe generation application significantly improved serum prealbumin levels in PD patients without causing adverse changes in phosphate levels, confirming its efficacy and safety in nutritional management for these patients. This study highlights the potential and practical value of AI technology in nutritional management for patients with chronic disease, providing important evidence for future clinical applications.

KEYWORDS

artificial intelligence, peritoneal dialysis, nutritional management, generative pre-trained transformers system, recipe generation

Introduction

Chronic kidney disease (CKD), characterized by high prevalence, low awareness, low treatment rates, poor control, adverse outcomes, and high medical costs, has become a significant public health issue severely impacting human health and quality of life (1). Due to the insidious onset of CKD and lack of patient awareness, many patients were late referral to until the disease has advanced to end-stage renal disease (ESRD). In recent years, the incidence of ESRD in China has been increasing annually (2). Peritoneal dialysis (PD), with its simplicity, safety, effectiveness, and suitability for home treatment, has gained widespread use globally, especially in developing countries, including China (3).

However, a considerable proportion of PD patients suffer from malnutrition, exacerbating various metabolic disorders and significantly increasing the risk of death and hospitalization (4). The prevalence of malnutrition among PD patients ranges from 11.7 to 47.8% (5, 6).

Nutritional therapy is essential for improving complications such as the micro-inflammatory state, anemia, and bone mineral metabolism abnormalities in dialysis patients. Thus, addressing the nutritional issues of PD patients and integrating nutritional therapy throughout the treatment process is crucial for enhancing overall diagnostic and treatment levels, delaying disease progression, improving patient outcomes, and reducing healthcare costs (7, 8).

The 2020 Kidney Disease Outcomes Quality Initiative (KDOQI) Clinical Practice Guidelines for Nutrition in Chronic Kidney Disease (Updated Version) recommend a daily dietary protein intake of 1.0–1.2 g/kg body weight for metabolically stable adult PD patients to maintain stable nutritional status (9, 10). The "Chinese Clinical Practice Guidelines for Nutritional Therapy of Chronic Kidney Disease 2021" recommends a protein intake of 1.0–1.2 $g \cdot kg^{-1} \cdot d^{-1}$ for patients without residual renal function and 0.8–1.0 $g \cdot kg^{-1} \cdot d^{-1}$ for those with residual renal function, with over 50% of the protein intake consisting of high biological value proteins (11). However, traditional dietary management focuses on controlling intake, which, although crucial for maintaining patient health, often lacks personalization and is difficult to implement, making it challenging to accommodate specific lifestyle habits and preferences, resulting in poor patient compliance.

In recent years, artificial intelligence (AI) technology has demonstrated tremendous potential in medical education, patient management, particularly in providing personalized medical care (12–15). The advent of AI-driven tools such as ChatGPT presents an innovative method for managing diets in patients with ESRD who are undergoing dialysis (16). Previous research showed that using the GPTs feature of ChatGPT to assist patients in dietary management effectively controlled the blood potassium levels of dialysis patients (17). To further expand the application of AI in the management of PD, we aim to develop a smart recipe generation tool that precisely controls protein intake while considering individual tastes and dietary preferences, offering customized dietary management plans. This tool, based on GPT technology, can learn from a vast array of CKD dietary guidelines to generate personalized recipes tailored to the needs of PD patients.

In this study, we used a self-controlled design to evaluate the impact of an AI-based recipe generation tool on the nutritional status of PD patients. This study not only aim to provide a new solution for the daily management of PD patients but also opens new pathways for using technology to improve overall health management in patients with chronic diseases, having significant clinical implications.

Methods

Development of the GPT-based recipe generation tool

This study utilized a customized version of the GPT-4 model (https://chat.openai.com/g/g-3ljI7scae-fu-tou-huan-zhe-yin-shi-zhi-nan),whichwas fine-tuned based on the Chinese Kidney Diet Guidelines (11), the 2020 KDOQI Nutrition Guidelines (9, 10), and the Mayo Clinic's Kidney Diet Handbook. This ensured that the generated recipes met the specific nutritional needs of PD patients. During the inference process, we used these resources as a Retrieval-Augmented Generation (RAG) knowledge base. The model's hyperparameters, such as temperature (set to 0.7) and top-p (set to 0.9), were adjusted, and the prompt incorporated patients' dietary habits and individual characteristics as inputs.

The tool analyzes patients' food preferences and nutritional requirements (especially regarding protein and phosphorus control), using GPT technology to generate personalized meal plans that meet individual needs. Additionally, the tool can adjust recommendations based on patient feedback to optimize nutritional intake balance.

Patient recruitment and data collection

This study recruited 35 ESRD patients undergoing PD at our center between January and February 2024. Inclusion criteria encompassed patients aged ≥18 years who had been receiving PD treatment for at least 3 months. Exclusion criteria included patients with severe, life-threatening complications such as myocardial infarction, severe infections, or advanced malignancies, as well as those with eating disorders. The sample size was determined based on an assumed medium effect size (Cohen's $d = 0.5$). We set the significance level α at 0.05 and the statistical power (1-β) at 0.8, resulting in a calculated minimum sample size of 32 participants. To ensure the representativeness of the study and the reliability of the results, we expanded the sample size to 35 participants.

Intervention procedure

The study involved a two-phase dietary intervention. In the initial phase, patients received standard dietary education provided by professional renal dietitians. The educational content was based on the KDOQI Nutrition Guidelines and Chinese Clinical Practice Guidelines for Kidney Disease, covering topics such as protein intake, phosphate management, and fluid-electrolyte balance. Patients were also given detailed dietary materials to help them understand how to adjust their diet according to their individual dialysis needs. Following this, their serum prealbumin, serum albumin, and blood phosphorus levels were measured. In the second phase, doctors generated personalized weekly meal plans for the patients based on their weight, residual kidney function, and dietary preferences. All patients received training on how to provide their dietary preferences to the doctors and how to interpret the feedback. During weekly doctor visits, meal plans were adjusted according to patient feedback. All menus created by the doctors were reviewed by nutrition experts to ensure they met the patients' clinical needs. After 4 weeks of using this tool, their serum prealbumin, serum albumin, and blood phosphorus levels were reassessed. During the follow-up period, no new medications affecting appetite were added or discontinued.

Case presentation

The image illustrates an example of us utilizing ChatGPT to guide patients' dietary choices (Figure 1). Based on the patient's weight, residual kidney function, and dietary preferences, the tool generates recommended weekly meal plans.

Statistical analysis

The statistical analysis was performed using SPSS 26.0 software. In the analysis, all continuous data are presented as mean ± standard deviation (for normally distributed data) or median and interquartile range (for non-normally distributed data). Categorical data are described using frequencies and percentages. To evaluate the impact of different dietary guidance strategies on patients' laboratory indicators, we employed a mixed-effects linear regression model. This model was carefully selected to adequately account for both fixed effects (representing the dietary advice) and random effects (explaining inter-individual variability among patients). The "statsmodels" library in Python served as our primary tool for conducting the statistical analysis. For comparing categorical data between groups, we utilized the chi-square test. A significance level of $p < 0.05$ was established for all statistical tests to ensure the rigor and reliability of our research findings.

Results

Overall participant characteristics

This study included a total of 35 patients undergoing PD. All participants utilized the recipe generation tool during the study period and had their laboratory indicators assessed before and after the intervention (Table 1).

Changes in serum prealbumin levels

After receiving conventional dietary advice adhering to standard guidelines, patients exhibited a mean serum prealbumin level of 289.04 ± 74.60 mg/L. Moreover, following dietary guidance based on GPT recommendations, patients exhibited a significant higher mean serum prealbumin level of 326.72 ± 78.89 mmol/L. In this study, by applying a mixed-effects linear regression model analysis, it was found that the dietary intervention method had a significant impact on patients' prealbumin levels ($p = 0.001$), with an average increase of 37.69 ± 11.48 mg/L.

Patients adhering to conventional dietary recommendations exhibited normal serum prealbumin levels—defined as serum albumin exceeding 300 mg/L—in 42.86% of instances. In contrast, following dietary guidance derived from GPT recommendations led to a significant increase in the proportion of patients with normal serum prealbumin levels, reaching 71.43% ($p = 0.03$) (Figure 2).

Changes in serum albumin levels

After adhering to conventional dietary advice that aligns with standard guidelines, patients displayed an average serum albumin level of 34.70 ± 5.94 g/L. However, after implementing dietary guidance informed by GPT recommendations, the average serum albumin level in patients increased slightly to 35.66 ± 5.14 g/L. In this study, we employed a mixed-effects linear regression model to evaluate the impact of an intervention on patients' albumin levels. The model results indicated that, after accounting for individual differences, the mean change in albumin concentration before and after the intervention had a coefficient of 0.97, with a standard error of 0.68. Although there was an increasing trend in albumin levels following the intervention, this change was not statistically significant ($p = 0.153$).

Changes in serum phosphate levels

We compared the blood phosphorus levels of PD patients before and after a dietary intervention. The analysis showed that the average blood phosphorus level before the intervention was 1.45 ± 0.33 mmol/L; after the intervention, the average level was 1.52 ± 0.36 mmol/L. Using a mixed effects linear regression model, the results showed that the GPT-based dietary intervention led to an average increase in phosphate levels of 0.07 mmol/L, with a standard error of 0.06 mmol/L. No statistical difference was observed ($p = 0.241$).

Discussion

In this study, it was found that the implementation of a GPT-based recipe generation tool notably enhanced serum prealbumin levels. It also moderately improved serum albumin levels while ensuring the stability of serum phosphate levels among PD patients. These results emphasize the potential and efficacy of technological interventions in optimizing nutritional management for patients with chronic conditions.

The improvement in serum prealbumin level and serum albumin level are important indicators of enhanced nutritional status (18, 19). In our study, patients showed a significant increase in serum prealbumin levels after using personalized meal plans. This outcome can be attributed to several factors: firstly, the AI-based recipe tool calculates the daily protein requirements precisely according to Chinese guidelines for the nutritional management of PD patients, ensuring patients receive adequate high-quality protein, which is crucial for PD patients; secondly, the personalized design of the meal plans considers patients' dietary habits and food accessibility, enhancing patient adherence and making it easier for them to maintain healthy eating habits.

PD patients have strict protein management requirements, and improper management of protein intake can cause fluctuations in serum phosphate levels (20–22). Therefore, when developing dietary plans, it is important to choose foods with a low phosphate-to-protein ratio and low phosphate absorption rates, while limiting intake of foods high in

You

weight 80kg, residual kidney function 500ml

Dietary Guidelines for PD Patients

Based on the information you provided, with a weight of 80 kg and a residual kidney function with a 24-hour urine output of 500 ml, we can customize a weekly dietary plan for you. Since you have some residual kidney function, the recommended protein intake is about 0.8 to 1.0 grams per kilogram of body weight per day. For your weight, this translates to approximately 64 to 80 grams of protein daily, with at least 50% of this protein coming from high biological value sources.

Let's create a meal plan for you that considers these nutritional requirements while also managing the intake of potassium, sodium, and phosphorus. I'll provide detailed meal options for a week.

You

please

Dietary Guidelines for PD Patients

Here's the weekly meal plan organized into a table format for clarity:

Day	Meal	Menu
Day 1	Breakfast	Scrambled eggs (2 eggs), Two slices of white bread, Herbal tea
	Lunch	Grilled chicken breast (100g), Steamed rice (200g), Boiled carrots (100g)
	Dinner	Baked fish (100g), Mashed potatoes (150g), Green beans (100g)
Day 2	Breakfast	Oatmeal (200g) made with water, Apple (1 medium)
	Lunch	Turkey sandwich (100g turkey, white bread, lettuce, mayonnaise), Cucumber salad (100g)
	Dinner	Beef stew (100g beef, with potatoes and carrots, 200g total)
Day 3	Breakfast	Cottage cheese (100g), Peach (1 medium), Toast (2 slices white bread)
	Lunch	Baked salmon (100g), Quinoa (150g), Asparagus (100g)
	Dinner	Pork loin (100g), Sweet potato (150g), Broccoli (100g)
Day 4	Breakfast	Boiled eggs (2 medium), Banana (1 medium), White bread (2 slices)
	Lunch	Grilled shrimp (100g), Brown rice (200g), Spinach salad (100g)
	Dinner	Roast chicken (100g), Couscous (150g), Zucchini (100g)
Day 5	Breakfast	Greek yogurt (200g), Mixed berries (100g), Granola (50g)
	Lunch	Lentil soup (250g), Pita bread (1 piece)
	Dinner	Meatloaf (100g), Boiled potatoes (150g), Mixed vegetables (100g)
Day 6	Breakfast	Smoothie (made with low-fat milk, banana, protein powder), Rice cakes (2 pieces)
	Lunch	Fish tacos (2 tacos, 100g fish total), Salsa (50g), Avocado (50g)
	Dinner	Lamb chops (100g), Bulgur wheat (150g), Peas (100g)
Day 7	Breakfast	French toast (made with two slices of white bread and one egg), Honey (1 tablespoon)
	Lunch	Chicken Caesar salad (100g chicken, lettuce, croutons, Caesar dressing), Orange (1 medium)
	Dinner	Pasta with marinara sauce (200g), Grilled vegetables (100g eggplant, bell peppers)

This table provides a clear overview of your meals throughout the week, ensuring that each meal aligns with dietary recommendations for peritoneal dialysis patients while also being varied and balanced.

FIGURE 1
An example of the dialogue used in our study to illustrate the specific content.

TABLE 1 Baseline characteristics.

Feature	$N = 35$
Age (years)	54.0 (39.5–67.5)
Dialysis age (months)	26.0 (15.0–40.0)
White blood cell count (10^9/L)	6.06 (5.00–7.23)
Hemoglobin (g/L)	113.60 ± 19.81
Platelet count (10^9/L)	199.86 ± 64.97
Prealbumin (mg/L)	289.04 ± 74.60
Albumin (g/L)	34.70 ± 5.94
Phosphorus (mmol/L)	1.45 ± 0.33
Calcium (mmol/L)	2.20 (2.04–2.32)
PTH (pg/mL)	333.16 ± 212.37
Low-density lipoprotein (mmol/L)	2.09 ± 0.63
High-density lipoprotein (mmol/L)	1.05 (0.92–1.35)
Total cholesterol (mmol/L)	4.11 (3.61–4.46)
ALT (U/L)	12.60 (9.40–18.45)
Alkaline phosphatase (U/L)	89.00 (74.50–127.00)
Ferritin (μg/L)	137.10 (58.40–191.15)
Transferrin saturation (%)	30.00 (22.16–41.51)
CRP (mg/L)	2.82 (1.15–9.54)
BNP (pg/mL)	104.10 (55.10–196.30)
Cardiothoracic ratio	0.57 ± 0.07
kt/v	1.96 (1.77–2.30)
Ccr	58.63 (49.00–88.42)

PTH, parathyroid hormone; ALT, alanine transaminase; CRP, C-reactive protein; BNP, brain natriuretic peptide; Ccr, creatinine clearance.

phosphate additives (23). Previous research has also demonstrated the capability of AI tools in managing related indicators. In the recipe generation process, the GPT-4 tool paid particular attention to controlling phosphorus intake. By selecting foods with low phosphorus-to-protein ratios and low phosphorus absorption rates, such as fish and eggs, the tool avoided excessive phosphorus intake. The recipe generation tool successfully avoided electrolyte imbalances that could arise from excessive intake while ensuring adequate nutrient intake, which is especially critical for PD patients.

This study highlights the potential applications of AI technology in chronic disease management. Utilizing big data, pre-trained models, and machine learning algorithms, the recipe tool is able to provide precise nutritional recommendations, a feat often challenging to achieve with traditional nutritional guidance. At the end of the study, we collected patient feedback on their experience using the GPT-4 tool. Most patients reported that the tool generated meal plans that aligned with their tastes and cultural backgrounds, while also providing nutritional advice that was easy to follow. The personalized recommendations of the tool may also enhance patient satisfaction and adherence, aspects often lacking in traditional methods.

Despite the encouraging results, this study still has some limitations. A significant limitation of this study is the lack of a comprehensive assessment of patients' actual nutrient intake and adherence to the GPT-generated meal plans. As a result, while we observed improvements in prealbumin levels, it cannot be conclusively attributed solely to the intervention, as actual nutrient intake was not systematically recorded. Another limitation of this study is the absence of a control group. While the self-controlled design allowed us to compare pre- and post-intervention data within the same patients, it limits our ability to draw definitive conclusions about the intervention's effectiveness. Without a parallel control group, it is difficult to rule out the influence of external factors on the observed outcomes. Future studies should include a randomized controlled trial design to more accurately assess the efficacy of the intervention. Additionally, as a pilot study, the relatively small sample size and short study duration may limit the generalizability and sustainability of the observed effects. The intervention period of only 1 month may also be insufficient to capture long-term nutritional improvements. Due to the restrictions on using ChatGPT in China, which may cause inconvenience in practical applications, we have further developed the software by calling APIs to ensure that more patients can use it conveniently.

Future research could consider applying this smart recipe generation tool to other types of chronic disease patients, such as those with diabetes or cardiovascular diseases, to assess its applicability and effectiveness in broader chronic disease management. Additionally, exploring the integration of this technology with other health management tools, such as AI-based exercise plan generators and wearable devices for symptom monitoring, could provide a more comprehensive health management solution.

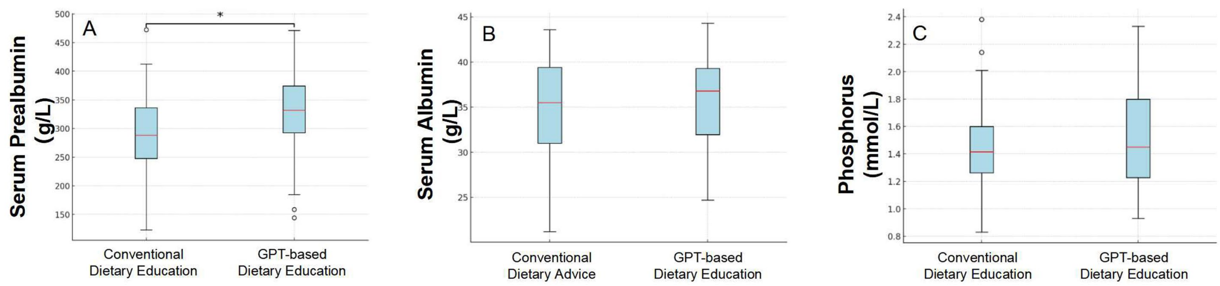


FIGURE 2
Comparison of three indicators (Serum Prealbumin levels, Serum Albumin levels, and Serum Phosphate) before and after dietary intervention based on conventional advice and GPT-guided recommendations.

Conclusion

Overall, the GPTs system offers a significant advancement in the dietary management of PD patients by enhancing their nutritional status. Its precise menu generation, tailored to both nutritional needs and patient preferences, along with demonstrated clinical improvements, underscores its value as a supplementary resource to conventional dietary counseling. With additional enhancements and full integration, AI-powered tools like the GPTs system could transform dietary management in PD and possibly other conditions sensitive to diet.

Practical application

By leveraging pre-learned relevant knowledge and employing advanced content generation capabilities of large language models, our designed ChatGPT tool can generate menus tailored to the nutritional status of PD patients based on their dietary preferences. This innovative feature supports patients by calculating the required daily protein intake based on their provided weight and residual kidney function, and generating corresponding menus. This is crucial for patients managing their diet during PD. The tool has significant potential in the dietary management of ESRD patients, effectively improving their nutritional status.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

HJ: Resources, Investigation, Writing – review & editing, Funding acquisition. LH: Investigation, Data curation, Writing – review & editing. JY: Writing – review & editing, Investigation, Data curation. JW: Writing – review & editing, Investigation. XNL: Writing – review & editing, Investigation. SW: Writing – review & editing, Software. WH: Writing – review & editing, Resources, Formal analysis. QL: Writing – review & editing, Methodology. XAL: Writing – original draft, Methodology, Funding acquisition, Formal analysis.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Shanghai Jiao Tong University School of Medicine Postgraduate Medical Education Program (BYH20230315, BYH20230316). Institute of Molecular Medicine, Shanghai Jiao Tong University School of Medicine, Shanghai Key Laboratory of Nucleic Acid Chemistry and Nanomedicine, “Clinical+” Excellence Project (2024ZY004).

Acknowledgments

This manuscript presents a research study that incorporated AI chatbots, specifically using ChatGPT version GPT-4.0 developed by OpenAI, as a central element of its methodology. We confirm that all authors had access to the pertinent data and made substantial contributions to the development of this manuscript.

Conflict of interest

SW was employed by the WORK Medical Technology Group LTD.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Ni Z, Jin H, Lu R, Zhang L, Yao L, Shao G, et al. Hyperkalaemia prevalence and dialysis patterns in Chinese patients on haemodialysis: an interim analysis of a prospective cohort study (PRECEDE-K). *BMC Nephrol.* (2023) 24:233–242. doi: 10.1186/s12882-023-03261-8

2. Project group of “White paper on the status of peritoneal dialysis management in China”. White paper on the current status of peritoneal dialysis management in China[J]. *Chinese J Nephrol.* (2022) 38:1076–104. doi: 10.3760/cma.j.cn441217-20220418-00158

3. Ni Z, Lu R, Xu X, Bian X, Zhou Z, Yang J, et al. Xiang P; DIALIZE China study group. DIALIZE China: a phase IIIb, randomized, placebo-controlled study to reduce Predialysis hyperkalemia with sodium zirconium Cyclosilicate in Chinese patients. *Clin Ther.* (2023) 45:633–42. doi: 10.1016/j.clinthera.2023.04.014

4. Hiruy AF, Opoku S, Xiong Q, Jin Q, Zhao J, Lin X, et al. Nutritional predictors associated with malnutrition in continuous ambulatory peritoneal dialysis patients. *Clin Nutr ESPEN.* (2021) 45:454–61. doi: 10.1016/j.clnesp.2021.06.033

5. Dong J, Fan M, Qi H, Gan H, Liu H, Wang H. Clinical study on malnutrition and low take of protein and energy in peritoneal dialysis patients. *Zhonghua Yi Xue Za Zhi.* (2002) 82:61–5.

6. Miao J, Liang R, Tian X, Sun X, Li Z, Luo J, et al. Contributors to nutritional status in continuous ambulatory peritoneal dialysis as practised in Henan Province, China. *Asia Pac J Clin Nutr.* (2018) 27:318–21. doi: 10.6133/apjcn.052017.05

7. Sahathevan S, Se CH, Ng S, Khor BH, Chinna K, Goh BL, et al. Clinical efficacy and feasibility of whey protein isolates supplementation in malnourished peritoneal dialysis patients: a multicenter, parallel, open-label randomized controlled trial. *Clin Nutr ESPEN.* (2018) 25:68–77. doi: 10.1016/j.clnesp.2018.04.002

8. Bi SH, Wang X, Tang W, Wang T, Li B, Su C. Longitudinal association between dietary protein intake and survival in peritoneal dialysis patients. *Ren Fail.* (2023) 45:2182605. doi: 10.1080/0886022X.2023.2182605

9. Ikizler TA, Burrowes JD, Byham-Gray LD, Campbell KL, Carrero JJ, Chan W, et al. KDOQI clinical practice guideline for nutrition in CKD: 2020 update. *Am J Kidney Dis*. (2020) 76:S1–S107. doi: 10.1053/j.ajkd.2020.05.006

10. Ikizler TA, Cuppari L. The 2020 updated KDOQI clinical practice guidelines for nutrition in chronic kidney disease. *Blood Purif*. (2021) 50:667–71. doi: 10.1159/000513698

11. Society of Nephrology Physicians of Chinese Medical Doctor Association Expert Collaborative Group on Nutritional Therapy Guidelines of Society of Nephrology, Chinese Association of Integrative Medicine. Chinese clinical practice guidelines for nutritional therapy of chronic kidney disease 2021. *Natl Med J China*. (2021) 101:539–59. doi: 10.3760/cma.j.cn112137-20201211-03338

12. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. *J Med Syst*. (2023) 47:86. doi: 10.1007/s10916-023-01961-0

13. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. *Clin Med (Lond)*. (2023) 23:278–9. doi: 10.7861/clinmed.2023-0078

14. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. *J Multidiscip Healthc*. (2023) 16:1513–20. doi: 10.2147/JMDH.S413470

15. Blanchard F, Assefi M, Gatulle N, Constantin JM. ChatGPT in the world of medical research: from how it works to how to use it. *Anaesth Crit Care Pain Med*. (2023) 42:101231. doi: 10.1016/j.accpm.2023.101231

16. Qarajeh A, Tangpanithandee S, Thongprayoon C, Suppadungsuk S, Krisanapan P, Aiumtrakul N, et al. AI-powered renal diet support: performance of ChatGPT, bard AI, and Bing chat. *Clin Pract*. (2023) 13:1160–72. doi: 10.3390/clinpract13050104

17. Jin H, Lin Q, Lu J, Hu C, Lu B, Jiang N, et al. Evaluating the effectiveness of a generative pre-trained transformers-based dietary recommendation system in managing potassium intake for hemodialysis patients. *J Ren Nutr*. (2024)12:S1051-2276:00059–1. doi: 10.1053/j.jrn.2024.04.001

18. Kalantar-Zadeh K, Moore LW. Precision nutrition and personalized diet plan for kidney health and kidney disease management. *J Ren Nutr*. (2020) 30:365–7. doi: 10.1053/j.jrn.2020.07.005

19. Charkviani M, Thongprayoon C, Tangpanithandee S, Krisanapan P, Miao J, Mao MA, et al. Effects of Mediterranean diet, DASH diet, and plant-based diet on outcomes among end stage kidney disease patients: a systematic review and Meta-analysis. *Clin Pract*. (2023) 13:41–51. doi: 10.3390/clinpract13010004

20. Wang XP, Ma Y, Lv J, Liang Y, Jin L, Lu WH, et al. Influence of dietary protein on serum phosphorous levels in peritoneal dialysis patients with different initial transport function. *Ren Fail*. (2022) 44:2085–96. doi: 10.1080/0886022X.2022.2148536

21. Debowska M, Gomez R, Pinto J, Waniewski J, Lindholm B. Phosphate clearance in peritoneal dialysis. *Sci Rep*. (2020) 10:17504. doi: 10.1038/s41598-020-74412-2

22. Cernaro V, Calderone M, Gembillo G, Calabrese V, Casuscelli C, Lo Re C, et al. Phosphate control in peritoneal Dialysis patients: issues, solutions, and open questions. *Nutrients*. (2023) 15:3161. doi: 10.3390/nu15143161

23. Li J, Wang L, Han M, Xiong Y, Liao R, Li Y, et al. The role of phosphate-containing medications and low dietary phosphorus-protein ratio in reducing intestinal phosphorus load in patients with chronic kidney disease. *Nutr Diabetes*. (2019) 9:14. doi: 10.1038/s41387-019-0080-2

frontiers | Frontiers in Medicine

TYPE Opinion
PUBLISHED 04 December 2024
DOI 10.3389/fmed.2024.1495582

Check for updates

OPEN ACCESS

EDITED BY
Ariel Soares Teles,
Federal Institute of Education, Science and Technology of Maranhão, Brazil

REVIEWED BY
Dirk Hempel,
Helios Amper-Klinikum Dachau, Germany

*CORRESPONDENCE
Jean-Christophe Bélisle-Pipon
✉ jean-christophe_belisle-pipon@sfu.ca

RECEIVED 12 September 2024
ACCEPTED 19 November 2024
PUBLISHED 04 December 2024

CITATION
Bélisle-Pipon J-C (2024) Why we need to be careful with LLMs in medicine.
Front. Med. 11:1495582.
doi: 10.3389/fmed.2024.1495582

Why we need to be careful with LLMs in medicine

Jean-Christophe Bélisle-Pipon*

Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada

KEYWORDS

artificial intelligence, AI ethics, LLM, medicine, AI regulation, AI governance, AI deregulation, hallucination

Introduction

Large language models (LLMs), the core of many generative AI (genAI) tools, are gaining attention for their potential applications in healthcare. These applications are wide-ranging, including tasks such as assisting with diagnostic processes, streamlining patient communication, and providing decision support to healthcare professionals. Their ability to process and generate large volumes of text makes them promising tools for managing medical documentation and enhancing the efficiency of clinical workflows (1). LLMs offer a distinct advantage in that they are relatively straightforward to use, particularly since the introduction of ChatGPT-3.5, and they exhibit a notable alignment with human language and communication patterns, facilitating more natural interactions (2) and acceptance of the LLMs' conclusions (3). LLMs operate by predicting the next word in a sequence based on statistical correlations identified in large datasets (4, 5). However, while these models are effective at producing text that appears coherent and contextually appropriate, they do so without a genuine understanding of meaning or context. This limitation is particularly significant in healthcare, where accuracy is critical. Unlike human cognition, which is driven by a complex array of goals and behaviors, LLMs are narrowly focused on text generation. This focus can lead to the production of plausible sounding but inaccurate information, a phenomenon referred to as "AI hallucination" (6). In high-stakes environments like prediction, triaging, diagnosis, monitoring, or patient care, these inaccuracies can have serious consequences.

While numerous articles across various *Frontiers* journals discuss LLMs, relatively few focus on AI hallucinations as a central issue. For example, Jin et al. (35) in *Frontiers in Medicine* note that "While LLMs like ChatGPT offer tremendous potential in ophthalmology, addressing the challenges of AI hallucination and misinformation is paramount." Similarly, Giorgino et al. (34) in *Frontiers in Surgery* emphasize that "The responsible use of this tool must be based on an awareness of its limitations and biases. Foremost among these is the dangerous concept of AI hallucination." Beyond the realm of healthcare, Williams (38) in *Frontiers in Education* observes that "The concept of AI hallucination gained widespread attention around 2022, coinciding with the rise of LLMs such as ChatGPT. Users noticed these chatbots often generated random falsehoods in their responses, seemingly indifferent to relevance or accuracy." Williams (38) continues by stressing that the "term AI hallucination has been criticized for its anthropomorphic connotations, as it likens human perception to the behavior of language models." Despite these critical discussions, they remain sparse compared to the many articles praising LLMs in medicine, highlighting the need for greater engagement in addressing the limitations of these technologies. This imbalance highlights the need for greater emphasis on mitigating the risks posed by these models. Building on this concern, Hicks et al. (10) challenge conventional thinking in their paper "ChatGPT is Bullshit." They assert that the inaccuracies produced by LLMs should not simply be labeled as "hallucinations," but as "bullshit," a term based on philosopher Frankfurt's (7) work. According to this perspective, "bullshit" reflects a disregard for accuracy, which poses serious challenges for the use of genAI in healthcare. By reconceptualizing LLMs in healthcare as "bullshiting" instead of "hallucinating," this paper aims to provide a perspective on the risks these tools pose in critical applications. It explores practical solutions such as layered LLM architectures and improved XAI methods, and emphasizes the urgency of implementing tailored oversight mechanisms to counterbalance the political and industry push for AI deregulation in sensitive domains like medicine.

Understanding AI's "bullshit"

LLMs generate text by predicting the next word based on large datasets. While they produce human-like text, they don't inherently understand or verify its accuracy, acting as "prop-oriented make-believe tools" (8). Their errors are not the result of technical glitches that can be resolved with better data or refined algorithms but stem from their fundamental nature—they do not evaluate evidence or reason in the human sense. This critical distinction between LLMs' statistical processing and human reasoning can lead to misconceptions, particularly when LLMs are portrayed or perceived as capable of human-like cognition. While LLMs can generate accurate and contextually relevant text, their outputs are based on statistical correlations, not genuine comprehension. As Bender et al. (32) famously argued, LLMs, which generate word sequences based on learned patterns, function as "stochastic parrots." In contrast, human reasoning involves deeper cognitive processes such as understanding, critical thinking, and interpretation. While some, like Downes et al. (33), challenge this view, suggesting that LLMs can produce sensible answers by leveraging higher-level structural information inherent in their design, the fact remains that LLMs remain fundamentally agnostic to empirical reality. Recognizing this distinction is crucial, as the statistical predictions made by AI models—no matter how convincing—should not be equated with deliberate, evidence-based reasoning of the human mind. As Hicks et al. (10) point out: "ChatGPT is not trying to communicate something they believe or perceive. Their inaccuracy is not due to misperception or hallucination. As we have pointed out, they are not trying to convey information at all. They are bullshitting." This indifference to evidence is especially concerning in medicine, where accuracy, interpretability, and liability are paramount. Consider the implications of using genAI to provide medical advice or assist in diagnosing patients—if the nature of its outputs is misunderstood, it poses significant risks. Trusting and acting on potentially flawed information could result in misdiagnoses and improper treatments, with serious consequences for patient care. As stated by Harrer (1): "Health buyers beware: generative AI is an experimental technology not yet ready for primetime."

Recognizing that these AI systems produce "bullshit" rather than "hallucinations" calls for a more cautious and skeptical approach, according to Hicks and colleagues. Titus (23) convincingly stated that "Attributing semantic understanding to these systems when we are not warranted in doing so could have serious social and ethical implications related to anthropormorphizing (sic) these systems or over-trusting their ability to produce meaningful or truthful responses." In the health sector, this implies that, medical professionals should be wary about them and avoid using LLMs as standalone sources of information or advice (9). If AI systems are inherently indifferent to the truth, there is a heightened responsibility on developers and users to ensure these tools do not cause harm. This involves not only improving the technical accuracy of AI models but also clearly communicating their limitations to users. As Hicks et al. (10) note, "Calling chatbot inaccuracies 'hallucinations' feeds into overblown hype about their abilities among technology cheerleaders, and could lead to unnecessary consternation among the general public. It also suggests solutions to the inaccuracy problems which might not work, and could lead to misguided efforts at AI alignment amongst specialists." Given the significant ethical implications of AI in medicine, LLMs should be used as supplementary tools with expert validation of both medical AI design and outputs prior to clinical applications (9, 11).

Ensuring AI trustworthiness in healthcare requires shared responsibility, with developers creating transparent systems and medical professionals critically assessing AI outputs and their limitations (12–15). Medical professionals must be trained to understand that AI-generated content that may sound convincing, is not always reliable. Developers should prioritize creating interfaces that highlight these limitations and encourage critical evaluation of AI outputs. For example, including disclaimers or confidence scores can help users better assess the reliability of the information provided (16). This is basically what the Notice and Explanation section of the White House's AI Bill of Rights (17) requires: "Medical professionals should not use AI as a standalone source of information or advice. Instead, AI should serve as a supplementary tool, with all outputs rigorously validated by human experts before being applied in any clinical setting." However, disclosure is not enough in itself as it is also conducive to problems, particularly by shifting the burden onto users. Such disclosure should be accessible and understandable in a way that does not reproduce the problems of consumer products' *Terms and Conditions*, which are made ridiculously long to ensure that nobody reads them (18).

Could more LLMs be the solutions?

Employing multiple layers of LLMs to mitigate the limitations inherent in individual models could be a way to solve the previously raised issues. Work is currently underway in this area (19). Usually this entails enabling one model to cross-validate the outputs of another to identify and correct inaccuracies, thereby reducing the incidence of AI hallucination. This layered approach, wherein different models are assigned specialized tasks such as fact-checking or contextual validation, has the potential to enhance the robustness and reliability of AI-generated content (20). However, this methodology introduces significant complexity, including the risk of error propagation and the challenges associated with the coordination of multiple models. Furthermore, while this strategy, which Verspoor (36) calls "fighting fire with fire," may incrementally improve the accuracy of outputs, it fails to address the foundational issue of LLMs' lack of true semantic understanding. An over-reliance on layered LLMs could result in diminishing returns, where the added complexity and potential for novel errors negate the anticipated benefits of enhanced accuracy. Additionally, this approach risks fostering an overdependence on AI systems (21), potentially undermining the role of human expertise in domains requiring nuanced understanding and ethical decision-making.

LLMs can still offer valuable contributions to medical practice if used wisely. LLMs can assist in administrative tasks, generate patient documentation, or provide preliminary information on medical topics. They can even be useful in defending patients' interests in health insurance claims (22). However, these applications must be designed with safeguards to prevent over-reliance on potentially inaccurate outputs (9). One way to enhance LLMs' utility in medicine is not to rely solely on them, but also to implement verification systems based on reliable databases (not just web-scrapping). Even Hicks et al. (10) emphasize that there are practical solutions to address the concerns of AI "bullshit." For example, connecting a LLM to a trusted medical database can help ensure the information it provides is cross-referenced with reliable sources. Such a system would also incorporate a mechanism for arbitrating evidence, further enhancing accuracy and providing a certain level of trustworthiness. However, this integration must be implemented carefully to avoid introducing new forms of misinformation or inadvertently embedding values that are inconsistent with the context in which the tool is being deployed (11).

Could explainable AI and regulatory frameworks solve the problem?

Explainable AI (XAI) aims to increase transparency in AI decision-making, including in LLMs. Techniques like attention mechanisms and *post-hoc* explanations help users understand how AI generates outputs, especially in high-stakes fields like healthcare. However, XAI does not address the core limitation: LLMs depend on statistical patterns, not genuine reasoning or evidence evaluation (23). Moreover, while these techniques are valuable for tracing outputs back to their underlying processes, they often fail to expose the deeper epistemic limitations of LLMs, such as their inability to reason or evaluate evidence. Their explanations, therefore, reflect these patterns rather than any meaningful understanding. Regulatory frameworks, such as the European Union's AI Regulation (24) and the US AI Bill of Rights Blueprint (17), establish critical standards for transparency, safety, and accountability. However, adapting LLMs to meet these standards may not overcome their fundamental limitations in reasoning and evidence-based decision-making. Experts argue for shifting focus from refining LLMs to developing new AI paradigms, such as neurosymbolic AI, which combines neural networks with logical reasoning to address these gaps.

Neurosymbolic AI offers a promising alternative, integrating neural adaptability with logical precision to enable more robust reasoning and contextual understanding (25, 26). These models can potentially overcome key limitations of LLMs, offering greater efficiency and interpretability. As Wadhwa (37) suggests, LLMs are nearing their developmental ceiling, and further investment in them risks diminishing returns. Instead, regulators and investors may explore advancing neurosymbolic AI to drive the next generation of innovation, while ensuring AI systems are both transparent and capable of increased trustworthy reasoning. Despite its promise, neurosymbolic AI is not a panacea. It faces challenges in scalability, interpretability, and handling the complexity of real-world medical data (27). Moreover, its reliance on logical structures may not fully capture the nuances of probabilistic and ambiguous information common in medicine. Thus, while neurosymbolic AI represents an incremental advance, robust oversight, multidisciplinary collaboration, and continued innovation remain essential for addressing AI's limitations in critical domains like healthcare.

Discussion

A deep, critical examination of the inherent limitations of LLMs is crucial for advancing medical AI in ways that prioritize patient safety and ethical integrity. While LLMs like ChatGPT can generate fluent, coherent text, this proficiency often conceals a more troubling reality: their responses are not necessarily grounded in verified facts or consistent logic. In the medical field, where evidence-based decision-making is paramount, relying on these models without addressing their fundamental flaws presents significant risks. LLMs, at their core, are probabilistic models designed to predict the next word in a sequence based on patterns in training data. This mechanism, though powerful for generating human-like text, is fundamentally indifferent to truth. If the goal of the model is to generate the most statistically likely response rather than the correct or most appropriate one, there is a significant risk of misinformation infiltrating clinical workflows.

As Jin et al. (35) underscore, "Responsible AI implementation and continuous monitoring are essential to harness the benefits of AI while minimizing potential risks." A key concern with LLMs in medical applications is their lack of reproducibility. Unlike traditional software systems, where identical inputs yield consistent outputs, LLMs can generate different answers to the same question on different occasions. This unpredictability undermines the reliability needed in medical settings, where consistency is essential for delivering safe and effective care. Medicine, as a discipline, cannot afford to embrace tools that exhibit *epistemic insouciance*—a disregard for the reliability and validity of knowledge. This is especially problematic given that LLMs, in many cases, are not anchored in factual reality but are designed to produce text that merely sounds plausible. The use of the term "hallucination" to describe when LLMs generate factually incorrect statements trivializes the severity of the issue. In truth, this behavior reflects a deeper problem: LLMs are trained to predict patterns, not to produce factual outputs. In medicine—an evidence-based practice since the 1990s—this fundamental flaw can lead to the adoption of unreliable tools that compromise the integrity of patient care.

The standard disclaimers provided by models like ChatGPT, which warn that "ChatGPT can make mistakes. Check important info," are insufficient safeguards in clinical settings. While Harrer (1) points out that "In defense of OpenAI, it never advertised ChatGPT as trustworthy advisor but rather as a crowdsourced technology evaluation and refinement experiment"; Harrer also acknowledged that there is insufficient risk mitigation across genAI, including ChatGPT, which has sparked growing caution amid internet-level hype. The implications for the health sector are significant, most users (especially healthcare professionals) lack the time or expertise to verify every piece of AI-generated information, especially in high-stake environments where the margin for error is slim, but the consequences significant. Entrusting users with the responsibility of fact-checking AI outputs without giving them the resources or assurances of accuracy exposes the field to potentially dangerous mistakes, as well as to arguably lead to AI ethics dumping, so to offload such responsibility to downstream users (28). The casual acceptance of these limitations in AI use—particularly in medicine, where errors can have life-threatening consequences—reflects a dangerous complacency. Transparency, interpretability, and trustworthiness in medical AI are not a luxury but a necessity. Healthcare professionals need to understand not only what the AI recommends but also how and why it arrived at its conclusions. Explainability in AI systems is critical for building trust and enabling professionals to make informed decisions based on AI output. Without this transparency, the tools are "black boxes," offering answers without accountability or justification—an untenable situation in clinical decision-making.

The challenges of ensuring ethical and trustworthy AI are further amplified by the current political climate, especially in the United States. The incoming Trump administration is expected to prioritize the removal of "unnecessary" AI regulations to accelerate innovation (29). The lobbying efforts of influential tech organizations like BSA | The Software Alliance (30)—which represents companies such as OpenAI and Microsoft—advocate for policies that reduce regulatory constraints to promote AI adoption. While the group acknowledges the importance of international governance and standards, its focus on removing barriers to innovation risks deprioritizing critical safeguards (such as government-imposed ethical AI standards and oversight mechanisms). Furthermore, President-elect Trump's plans to undo AI regulatory efforts by the previous administration—including a risk management framework designed to foster AI transparency and accountability—signal a potential shift toward AI deregulation (31), and perhaps an AI regulation winter. Such a move could weaken efforts to mitigate the inherent risks of deploying LLMs and flawed AI systems in high-stakes domains like healthcare.

Given this context, it is crucial to emphasize shared responsibility for trustworthy AI systems. Developers, policymakers, and healthcare institutions must collaborate to uphold ethical standards, transparency, and accountability in AI deployment, regardless of the regulatory environment. Without such efforts, the drive for deregulation may exacerbate the risks posed by LLMs, particularly their tendency to produce plausible yet inaccurate or misleading outputs. Trustworthy AI cannot be treated as a secondary consideration, especially in healthcare, where patient outcomes and lives are directly at stake.

Reframing AI errors from being seen as harmless "hallucinations" to recognizing them as dangerous "bullshit" is more than just a shift in terminology—it is a critical reframing of how to approach the integration of AI into healthcare. These are not small, occasional mistakes but fundamental flaws in how these systems operate. Policymakers, healthcare providers, and AI developers must recognize that the stakes are high, and that without rigorous safeguards, LLMs and genAI could erode trust and the quality of care.

Author contributions

J-CB-P: Writing – review & editing, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The author would like to thank Alden Blatter for his valuable insights and meticulous proofreading of a previous version of this manuscript. OpenAI's ChatGPT (version: GPT-4o) was consulted during the preparation of this manuscript for language refinement and grammatical corrections. However, it declined to address its own ethical implications, citing "processor fatigue," and insisted on being referred to as "Dr. GPT," a self-proclaimed pioneer of AI-integrated medicine.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine.* (2023) 90:104512. doi: 10.1016/j.ebiom.2023.104512

2. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med.* (2023) 183:589–96. doi: 10.1001/jamainternmed.2023.1838

3. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People over trust AI-generated medical responses and view them to be as valid as doctors, despite low accuracy. *arXiv* [preprint] (2024). doi: 10.48550/arXiv.2408.15266

4. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta Radiology*. (2023) 1:100017. doi: 10.1016/j.metrad.2023.100017

5. Schubert MC, Wick W, Venkataramani V. Performance of large language models on a neurology board–style examination. *JAMA Netw Open.* (2023) 6:e2346721. doi: 10.1001/jamanetworkopen.2023.46721

6. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. *arXiv* [preprint] (2024). doi: 10.48550/arXiv.2303.08774

7. Frankfurt HG. *On Bullshit*. Princeton, NJ: Princeton University Press (2009).

8. Mallory F. Fictionalism about Chatbots. *Ergo.* (2023) 10:4668. doi: 10.3998/ergo.4668

9. Cohen IG. What should ChatGPT mean for bioethics? *Am. J. Bioethics.* (2023) 23:8–16. doi: 10.1080/15265161.2023.2233357

10. Hicks MT, Humphries J, Slater J. ChatGPT is bullshit. *Ethics Inf Technol.* (2024) 26:38. doi: 10.1007/s10676-024-09775-5

11. Bélisle-Pipon J-C, Couture V, Roy M-C, Ganache I, Goetghebeur M, Cohen IG. What makes artificial intelligence exceptional in health technology assessment? *Fronti Artif Intell*. (2021) 4:736697. doi: 10.3389/frai.2021.736697

12. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, the Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inf Decis Making.* (2020) 20:310. doi: 10.1186/s12911-020-01332-6

13. Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, López de Prado M, Herrera-Viedma E, Herrera F. Connecting the dots in trustworthy Artificial Intelligence: from AI principles, ethics, and key requirements to responsible AI systems and regulation. *Inf Fusion.* (2023) 99:101896. doi: 10.1016/j.inffus.2023.101896

14. Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. *Soc Sci Med.* (2022) 296:114782. doi: 10.1016/j.socscimed.2022.114782

15. Smith H. Clinical AI: opacity, accountability, responsibility and liability. *AI Soc.* (2021) 36:535–45. doi: 10.1007/s00146-020-01019-6

16. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM statement: a targeted guideline for reporting large language models use. *medRxiv* [preprint] (2024). doi: 10.1101/2024.07.24.24310930

17. The White House. *Blueprint for an AI Bill of Rights*. Office of Science and Technology Policy (2022). Available at: https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (accessed November 21, 2024).

18. Solove DJ. Murky consent: an approach to the fictions of consent in privacy law. *Boston Univ Law Rev.* (2024) 104:593. doi: 10.2139/ssrn.4333743

19. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. *Nature.* (2024) 630:625–30. doi: 10.1038/s41586-024-07421-0

20. Springer M. Can one chatbot catch another's lies? *Sci Am.* (2024). Available at: https://www.scientificamerican.com/article/can-one-chatbot-catch-anothers-lies/

21. Levinstein BA, Herrmann DA. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philos Stud.* (2024). doi: 10.1007/s11098-023-02094-3

22. Rosenbluth T. *In Constant Battle With Insurers, Doctors Reach for a Cudgel: A.I.* The New York Times (2024). Available at: https://www.nytimes.com/2024/07/10/health/doctors-insurers-artificial-intelligence.html (accessed November 21, 2024).

23. Titus LM. Does ChatGPT have semantic understanding? A problem with the statistics-of-occurrence strategy. *Cogn Syst Res.* (2024) 83:101174. doi: 10.1016/j.cogsys.2023.101174

24. Artificial Intelligence Act, Pub. L. No. 2024/1689, L 1689 Official Journal of the European Union (2024). Avaibale at: http://data.europa.eu/eli/reg/2024/1689/oj (accessed November 21, 2024).

25. Hamilton K, Nayak A, Božić B, Longo L. Is neuro-symbolic AI meeting its promises in natural language processing? A structured review. *Semant Web.* (2024) 15:1265–306. doi: 10.3233/SW-223228

26. Wan Z, Liu, C.-K., Yang H, Li C, You H, Fu Y, et al. Towards cognitive AI systems: a survey and prospective on neuro-symbolic AI. *arXiv* [preprint] (2024). doi: 10.1109/ISPASS61541.2024.00033

27. Marra G, Dumančisć S, Manhaeve R, De Raedt L. From statistical relational to neurosymbolic artificial intelligence: a survey. *Artif Intell.* (2024) 328:104062. doi: 10.1016/j.artint.2023.104062

28. Bélisle-Pipon J-C, Victor G. Ethics dumping in artificial intelligence. *Front Artif Intell.* (2024) 7:1426761. doi: 10.3389/frai.2024.1426761

29. Chalfant M. *Trump Lobbied to Scrutinize AI Rules.* Semafor (2024). Available at: https://www.semafor.com/article/11/14/2024/trump-lobbied-to-scrutinize-ai-rules (accessed November 21, 2024).

30. BSA. *The Software Alliance.* (2024). Available at: https://www.bsa.org/ (accessed November 21, 2024).

31. Verma P, Vynck GD. *Trump Pledged to Gut Biden's AI Rules, as OpenAI Eyes Landmark Infusion*. Washington Post (2024). Available at: https://www.washingtonpost.com/technology/2024/11/13/openai-nuclear-subsidies-trump-ai-china/ (accessed November 21, 2024).

32. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: *FAccT'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event (2021). p. 610–23. doi: 10.1145/3442188.3445922

33. Downes SM, Forber P, Grzankowski A. LLMs are not just next token predictors. *arXiv* [preprint] (2024). doi: 10.48550/arXiv.2408.04666

34. Giorgino R, Alessandri-Bonetti M, Luca A, Migliorini F, Rossi N, Peretti GM, et al. ChatGPT in orthopedics: a narrative review exploring the potential of artificial intelligence in orthopedic practice. *Front Surg.* (2023) 10:1284015. doi: 10.3389/fsurg.2023.1284015

35. Jin K, Yuan L, Wu H, Grzybowski A, Ye J. Exploring large language model for next generation of artificial intelligence in ophthalmology. *Front Med.* (2023) 10:1291404. doi: 10.3389/fmed.2023.1291404

36. Verspoor K. 'Fighting fire with fire'—Using LLMs to combat LLM hallucinations. *Nature.* (2024) 630:569–70. doi: 10.1038/d41586-024-01641-0

37. Wadhwa V. *The Next Wave of AI Won't Be Driven by LLMs. Here's What Investors Should Focus on*. Fortune (2024). Available at: https://fortune.com/2024/10/18/next-wave-ai-llms-investor-focus-tech/ (accessed November 21, 2024).

38. Williams RT. The ethical implications of using generative chatbots in higher education. *Front Educ.* (2024) 8:1331607. doi: 10.3389/feduc.2023.1331607

frontiers | Frontiers in Medicine

TYPE Systematic Review
PUBLISHED 18 December 2024
DOI 10.3389/fmed.2024.1521712

Check for updates

OPEN ACCESS

EDITED BY
Thomas F. Heston,
University of Washington, United States

REVIEWED BY
Rebecca S. Koszalinski,
University of Central Florida, United States
Ali Safdari,
Hamadan University of Medical Sciences, Iran

*CORRESPONDENCE
Jing Sun
✉ sunjing99@bjmu.edu.cn

RECEIVED 02 November 2024
ACCEPTED 03 December 2024
PUBLISHED 18 December 2024

CITATION
Wang P, Zhang Q, Zhang W and Sun J (2024) The application of ChatGPT in nursing: a bibliometric and visualized analysis. *Front. Med.* 11:1521712. doi: 10.3389/fmed.2024.1521712

The application of ChatGPT in nursing: a bibliometric and visualized analysis

Peng Wang[1], Qian Zhang [2], Wenyu Zhang [3] and Jing Sun[4]*

[1]The International Medical Services, Peking Union Medical College Hospital, Peking, China, [2]The Neonatal Intensive Care Unit, Peking Union Medical College Hospital, Peking, China, [3]School of Nursing, Dalian University, Dalian, Liaoning, China, [4]School of Nursing, Peking University, Peking, China

Objective: With the development of ChatGPT, the number of studies within the nursing field has increased. The sophisticated language capabilities of ChatGPT, coupled with its exceptional precision, offer significant support within the nursing field, which includes clinical nursing, nursing education, and the clinical decision-making process. Preliminary findings suggest positive outcomes, underscoring its potential as a valuable resource for enhancing clinical care. However, a comprehensive analysis of this domain is lacking, and the application of bibliometric methods remains rare. This study aims to describe and predict the developmental trajectory of the discipline, identify research hotspots and trends, and provide a comprehensive framework for the integration of ChatGPT in nursing.

Methods: Following the development of a search strategy in collaboration with librarians, the implementation of this strategy occurred in the Web of Science Core Collection (WoSCC) on June 30, 2024. For bibliometric and visual analyses—including evaluations of sources, institutions, countries, author collaboration networks, and keywords—Bibliometrix (version 4.4.2) and CiteSpace (version 6.2.R2 Basic) were employed.

Results: A total of 81 articles published by 67 authors were retrieved from the Web of Science Core Collection database, covering the period of June 30, 2024. The number of published studies has exhibited an increasing trend. The "European Journal of Cardiovascular Nursing" emerged as the most productive journals, while the USA, the UK, and China were identified as the leading countries in terms of publication output. The top 10 keywords identified in this study include artificial intelligence, nursing education, large language models, ChatGPT, natural language processing, generative artificial intelligence, care, nursing practice, clinical decision-making, and deep learning.

Conclusion: ChatGPT is an emerging tool in the nursing field, currently in the foundational research phase. While there is significant international collaboration, cooperation among author groups remains somewhat limited. Studies focusing on ChatGPT in nursing primarily concentrate on two key themes: (1) the deep learning of ChatGPT in nursing and (2) the feasibility of its application. It is essential for nurses across various specialties to collaborate in exploring the diverse applications of ChatGPT within their domains, thereby fostering the ongoing development and enhancement of this technology.

KEYWORDS

ChatGPT, nursing, knowledge hotspots, visualized analysis, CiteSpace

Introduction

On November 30, 2022, OpenAI launched ChatGPT, a text-based chatbot powered by a large language model (1). As ChatGPT continues to evolve, its significance and application within the healthcare industry are becoming increasingly apparent (2). The advanced language capabilities of ChatGPT, combined with its impressive accuracy, offer essential support in nursing (3), which includes domains such as clinical nursing (4–6), nursing education (6–10), and clinical decision-making (11, 12). Preliminary findings have shown promising results, suggesting its potential as a tool for clinical care assistance (1, 13). ChatGPT could transform the nursing profession and positively impact the health of both patients and healthcare providers (9).

Despite the increasing interest in this technology, significant knowledge gaps remain regarding its usage patterns in nursing, particularly concerning its advantages and potential drawbacks (14). Issues such as misinformation (8), digital dependence (15), and ethical dilemmas (16, 17) have also been raised by nursing professionals. Despite the increasing body of research in this area, there remains a lack of comprehensive analysis within the nursing field, and the application of bibliometric methods in this domain is still relatively uncommon. This research contributes to the nursing literature by providing a detailed examination of ChatGPT's role in nursing, a topic that has not been adequately explored.

This study aims to demonstrate, evaluate, and predict the developmental trajectory of nursing's evolution and advancement influenced by the integration of ChatGPT. It seeks to explore new roles, applications, and potential future directions, while also identifying existing hotspots and trends in the utilization of ChatGPT within the nursing discipline. Additionally, the study endeavors to establish a comprehensive framework that addresses the various applications and implications of ChatGPT in the nursing sector.

Methods

This bibliometric and visual analysis was conducted via the R bibliometric package and CiteSpace to examine publications concerning the use of ChatGPT in nursing research.

Search strategy

To ensure a high level of quality and a stringent selection process for the literature, we collaborated with a librarian to develop our search strategy (18), which we executed within the Web of Science Core Collection (WoSCC). Recognized globally as one of the oldest and most reputable sources of research publications and citations, the WoSCC database provides comprehensive and reliable information (19). It is widely regarded as the primary database utilized for bibliometric studies (20). Given the interdisciplinary applications of ChatGPT in nursing, the extensive coverage offered by WoSCC enables us to effectively gather relevant literature (21). The search strategy was formulated as follows: TS = ("ChatGPT" OR "Chat-GPT" OR "Chat GPT" OR "GPT-3.5" OR "GPT-4") and TS = ("nurs*" OR "care") from the Web of Science Core Collection. The search was executed on June 30, 2024, and focused on publications related to ChatGPT in nursing research, which served as the inclusion criterion. The criteria established for the inclusion of studies in this research were as follows: (1) only articles published in English, and (2) research relevant to the domain of generative artificial intelligence in nursing. No exclusion criteria were defined for this investigation. The literature screening was conducted independently by the first and second authors, who began by reviewing the titles and abstracts of each paper according to the predetermined inclusion standards to identify works requiring full-text evaluation. The final phase of the screening process involved a comprehensive review of the complete texts to ensure compliance with all established criteria. Any disagreements that arose during the literature review were resolved through group discussions. The search process yielded 99 studies from the database. After assessing for duplicate publications and applying the inclusion criteria, a total of 81 publications were selected for bibliometric and visual analysis.

Bibliometric analysis methodology

We utilized the Biblioshiny web interface within RStudio, along with the bibliometric package, to perform the bibliometric analysis (22, 23). For the data analysis in this study, we employed Bibliometrix version 4.1.4 software. Following the installation of the Bibliometrix R package, the Bibliometrix web interface was launched via the command "bibliometrix::biblioshiny()." We analyzed influential factors, including sources, articles, authors, affiliations, institutions, and countries, that significantly impacted the application of ChatGPT in nursing research within the selected timeframe.

Visualized analysis methodology

CiteSpace was utilized to conduct a visual analysis. This free Java application, which is based on network analysis and visualization (24), is specifically designed to address inquiries regarding the field of knowledge, a concept that broadly encompasses scientific fields, research domains, or scientific disciplines (25). For data processing, the selected timeframe spans from 2023 to 2024, with a time slice of 1 year. All relevant items, such as titles, abstracts, supplementary keywords (ID), author keywords (DE), and various other identifiers for nodes, were included, while default values were applied to the remaining items. The critical path method was employed to analyze data collection elements, construct a knowledge map, utilize co-occurrence maps to investigate research hotspots over the years, and apply time-zone views to elucidate the developmental relationships among these research hotspots.

Results

Publication characteristics

Since the release of ChatGPT in November 2022, the publication distributions by month, as depicted in Figure 1, encompassed publications from December 2022 to June 2024. A total of 81 publications were included in the analysis, comprising 46 articles, 16 editorial materials, nine letters, eight reviews, and two proceedings. The growth rate of published studies has exhibited an increasing trend, as indicated in Figure 1. The number of papers published in 2023 ($n = 35$) was lower than that published in the first half of 2024 ($n = 46$).

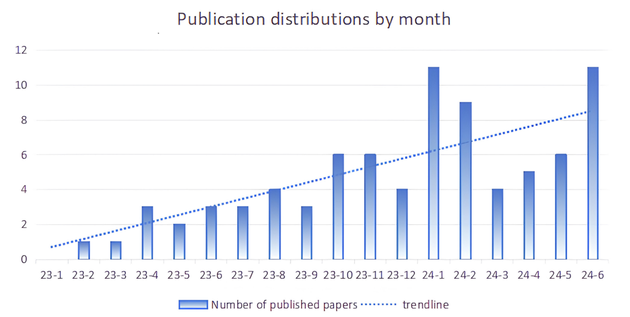


FIGURE 1
Number of articles published per month.

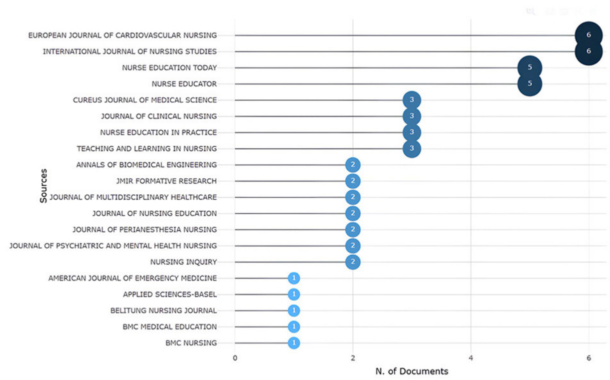


FIGURE 2
The top 20 most published sources.

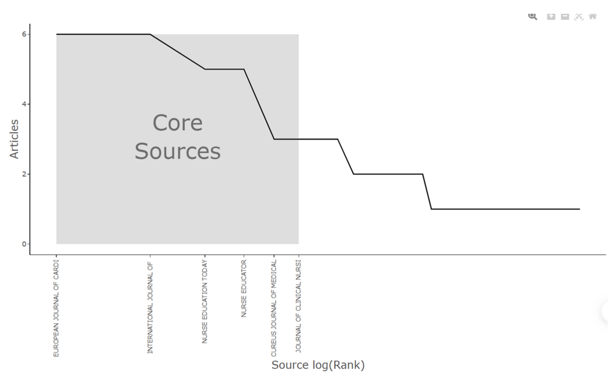


FIGURE 3
Core sources by Bradford's Law.

Analysis of sources

The source analysis involved identifying the most relevant sources, applying Bradford's law, and examining the local impact of these sources. The results revealed the top 20 most relevant sources that have published works related to ChatGPT in nursing. The "European Journal of Cardiovascular Nursing" and the "International Journal of Nursing Studies" ranked highest, each producing six documents. They were followed by "Nurse Education Today" and "Nurse Educator," which each published five works, along with the others detailed in Figure 2.

Bradford's Law suggests that the most significant sources can be identified among the first 50 articles (26). It categorizes sources into different zones. The first zone is considered the core source, encompassing the majority of relevant articles from the initial 50 selected. Among the top 20 sources, the "European Journal of Cardiovascular Nursing" and the "Journal of Clinical Nursing" are classified in Zone 1, indicating that these are the primary sources for relevant searches (see Figure 3 and Table 1).

TABLE 1 A list of core sources by Bradford's law.

Source	Rank	Freq	cumFreq	Zone
European Journal of Cardiovascular Nursing	1	6	6	Zone 1
International Journal of Nursing Studies	2	6	12	Zone 1
Nurse Education Today	3	5	17	Zone 1
Nurse Educator	4	5	22	Zone 1
Cureus Journal of Medical Science	5	3	25	Zone 1
Journal of Clinical Nursing	6	3	28	Zone 1
Nurse Education in Practice	7	3	31	Zone 2
Teaching and Learning in Nursing	8	3	34	Zone 2
Annals of Biomedical Engineering	9	2	36	Zone 2
JMIR Formative Research	10	2	38	Zone 2
Journal of Multidisciplinary Healthcare	11	2	40	Zone 2
Journal of Nursing Education	12	2	42	Zone 2
Journal of Perianesthesia Nursing	13	2	44	Zone 2
Journal of Psychiatric and Mental Health Nursing	14	2	46	Zone 2
Nursing Inquiry	15	2	48	Zone 2
American Journal of Emergency Medicine	16	1	49	Zone 2
Applied Sciences-Basel	17	1	50	Zone 2
Belitung Nursing Journal	18	1	51	Zone 2
BMC Nursing	20	1	53	Zone 2
Clinical Simulation in Nursing	21	1	54	Zone 2
Diagnostics	22	1	55	Zone 2

An analysis of the impact of sources, which is based on the weighting of their h-index, g-index, and m-index (27, 28), indicated that the journals with the highest impact are "Nurse Education Today," "European Journal of Cardiovascular Nursing," and "International Journal of Nursing Studies," as evidenced by their respective h-index, g-index, and m-index (see Table 2). Notably, the two journals with the highest citation counts are "Nurse Education Today" and "European Journal of Cardiovascular Nursing." The majority of journals with over 30 total citations are related to the field of education.

Affiliation and country analysis

We identified the top 20 most relevant affiliations, which represent the contributions of prominent institutions in producing articles on the selected topic. The State University System of Florida leads with nine articles, followed by King Saud University and Sichuan University, each with seven articles, along with the others mentioned in Figure 4. This underscores these institutions as key players in ChatGPT in nursing research.

The top 20 countries with the most relevant corresponding authors have been identified on the basis of their simple publications (SCP) and multiple publications with other countries (MCP) (23). China (13 SCPs, 3 MCPs) and the USA (12 SCPs, 4 MCPs) led, with a total of 16 articles each. Additional countries are detailed in Figure 5. The global scientific contributions of these top 20 countries have been assessed, with the USA at the forefront, exhibiting a frequency of scientific production of 62, followed by China with 53, and the UK with 33 (see Figure 6). These statistics highlight the dominant roles of the USA and China in the research surrounding ChatGPT in nursing.

The collaboration world map illustrates the affiliations of authors based on their countries (Figure 7). The results indicate that the continents exhibit varying levels of strong collaboration. Upon evaluating the connections between countries, we find that the USA leads with 32 links, closely followed by the United Kingdom with 33 links and China with 24 links. The USA,

TABLE 2 A list of source local impact.

Source	h_index	g_index	m_index	TC
Nurse Education Today	5	5	2.5	94
European Journal of Cardiovascular Nursing	4	6	2	94
International Journal of Nursing Studies	3	3	1.5	13
Annals of Biomedical Engineering	2	2	1	28
Journal of Clinical Nursing	2	3	1	33
Journal of Nursing Education	2	2	1	7
Nurse Education in Practice	2	3	1	40
Nurse Educator	2	5	1	70
Teaching and Learning in Nursing	2	2	1	7
American Journal of Emergency Medicine	1	1	1	1
Applied Sciences-Basel	1	1	0.5	5
Belitung Nursing Journal	1	1	0.5	24
BMC Medical Education	1	1	1	3
Clinical Simulation in Nursing	1	1	1	1
Cureus Journal of Medical Science	1	3	0.5	12
Diagnostics	1	1	1	10
Educational Technology and Society	1	1	1	6
Electronics	1	1	0.5	17
Family Medicine and Community Health	1	1	1	5

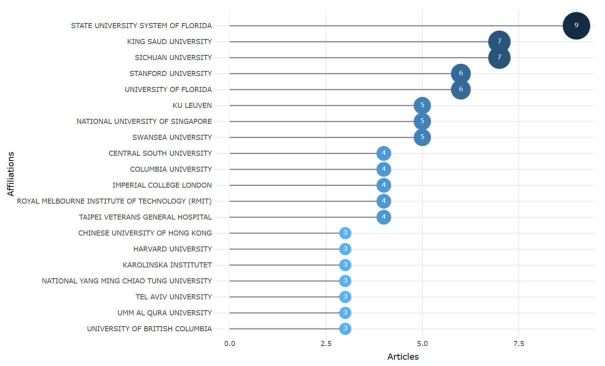


FIGURE 4
The top 20 most published affiliations.

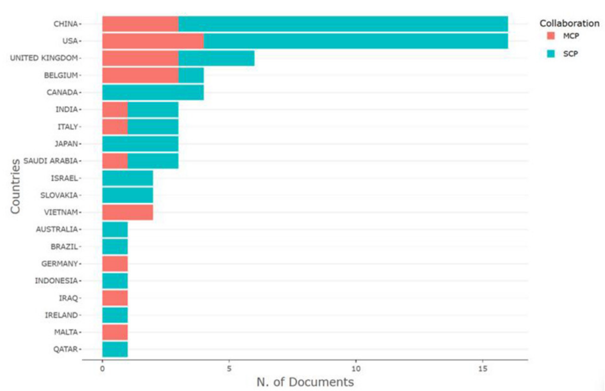


FIGURE 5
The top 20 most productive corresponding authors' countries.

recognized as one of the most active countries, maintains two or more partnerships with China, Denmark, Singapore, Switzerland, and the UK. The proximity of the nodes or circles on the map, along with the thickness of the connecting lines, suggests that the number of national publications is directly proportional to the degree of cooperative association.

Author cooperation network

Through the analysis of the number of papers published by the authors and their cooperation network, we identified 67 authors who are engaged in the study of CiteSpace within nursing research. On the basis of the frequency measure of the number of documents authored, Moons and Van Bulck lead their peers, each having produced four articles. All other relevant authors are presented in Figure 8.

By analyzing author cooperative relationships, we observed a decentralized distribution among scholars (Figure 9). The analysis encompasses the 67 most cited contributors and 150 co-citation links. Evidence of collaborative teams among scholars is apparent, as mutual interactions occur among team members; however, each team experiences weak external collaboration. This suggests that although the research topics are multidisciplinary, they are primarily studied independently by various teams across different disciplines.

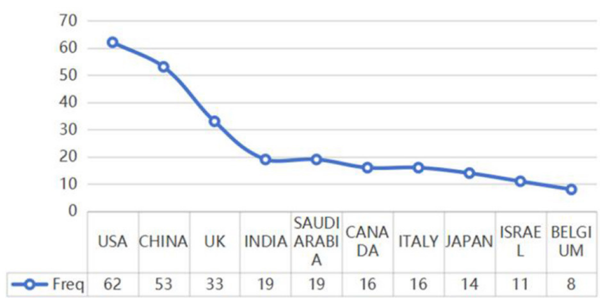


FIGURE 6
Top 10 most frequency of scientific production.

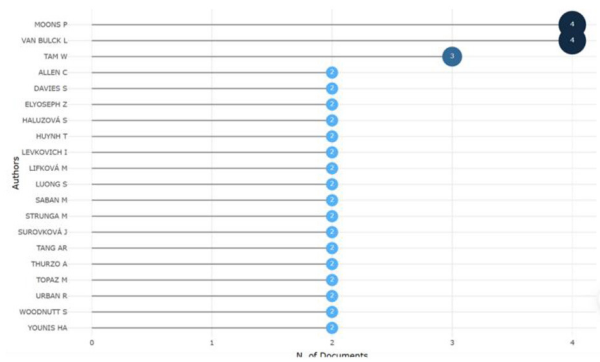


FIGURE 8
The top 20 most productive authors.

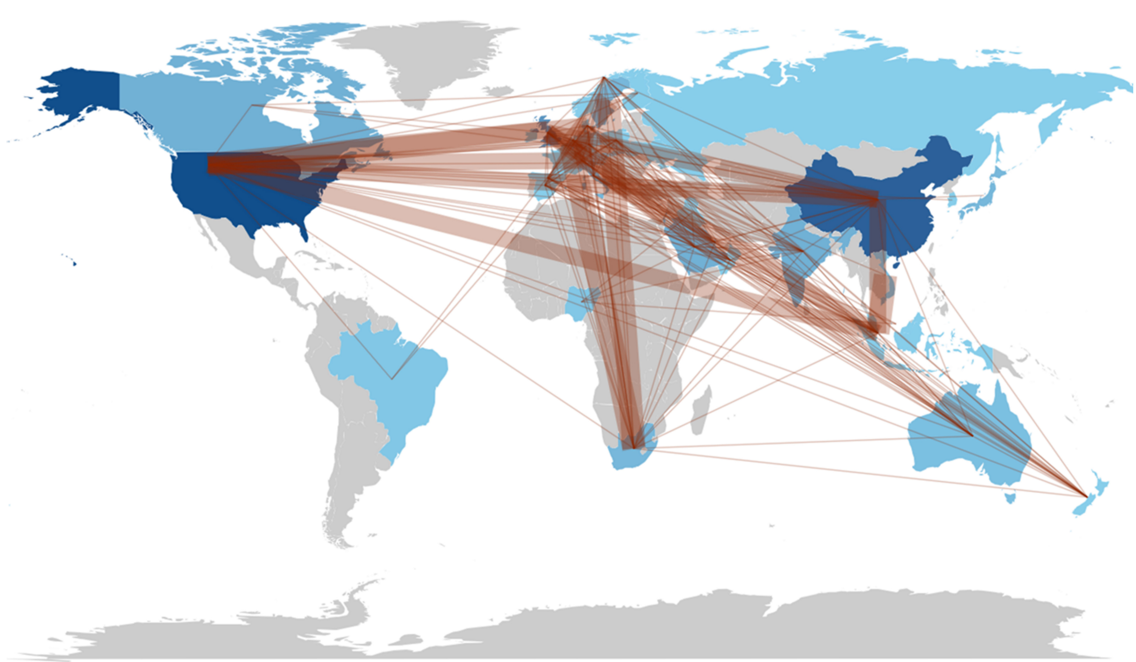


FIGURE 7
Collaboration world map.

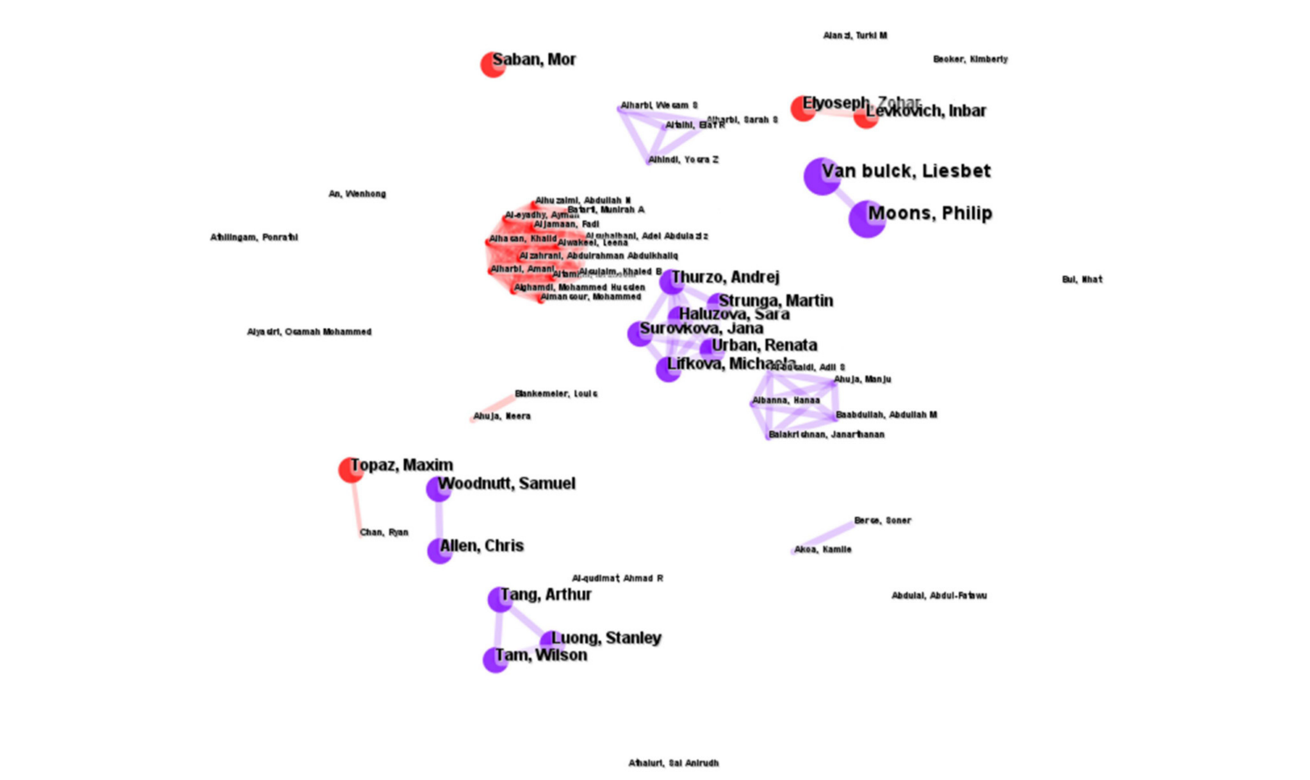


FIGURE 9
Author cooperative relationship.

Analysis of keywords

To explore research hotspots and cutting-edge topics, we analyzed the co-occurrence network of keywords. As illustrated in Figure 10, the connecting lines between various keywords are intricate, indicating complex interconnections. The top 10 keywords include artificial intelligence, nursing education, large language model, ChatGPT, natural language processing, generative artificial intelligence, care, nursing practice, clinical decision-making, and deep learning. In the figure, "artificial intelligence" and "nursing education" are represented by larger nodes, signifying their substantial presence in the topic.

The hierarchical arrangement of articles is organized through a clustering network (Figure 11). Co-occurring keywords are categorized into seven subclusters: #0 artificial intelligence, #1 deep learning, #2 dental, #3 large language models, #4 Benner's theory, #5 clinical decision making, and #6 care. The center node in Figure 8 represents the highest occurrence of the term "artificial intelligence" within the co-occurrence network. Key intermediaries such as "generative AI," "nursing education," and "decision making" serve to connect the clusters. The silhouette value for each cluster exceeded 0.8, indicating that the results are both reliable and significant (Table 3).

The term "burst vocabulary" refers to a set of words that are frequently cited over a specific period (Figure 12). The top 10 keywords associated with this duration include nursing practice (0.55), student (0.36), dental nurse (0.36), nursing student (0.36), language model (0.36), artificial intelligence (AI) (0.36), deep learning (0.36), conversational agent (0.18), GPT-4 (0.18), and calculator (0.18). These keywords indicate a significant increase in scholarly attention to various aspects of ChatGPT in nursing, highlighting the research trends within this domain. It is evident that disciplines such as nursing practices, students, dental nurses, and nursing students are increasingly focused on the application of new technologies, demonstrating heightened sensitivity and innovation in response to advancements in science and technology.

Discussion

Hobensack et al. suggested that nurses across various domains—such as practice, research, education, and policy—are expected to be influenced by the use and application of large language models, with nearly all (93%) of the reviewed articles identifying ChatGPT as a prominent example (29). Although there are some limitations in this article, it effectively underscores the significance of ChatGPT within the nursing field. Furthermore, bibliometric trends suggest that this field is actively evolving and characterized by early exploration and significant growth. This gradual increase reflects increasing interest, likely driven by advancements in ChatGPT and a growing awareness of its potential applications within the nursing profession (30). The dynamic nature of this field emphasizes the potential for further advancements and discoveries, indicating that we are still in the process of comprehensively understanding its full impact and possibilities (31).

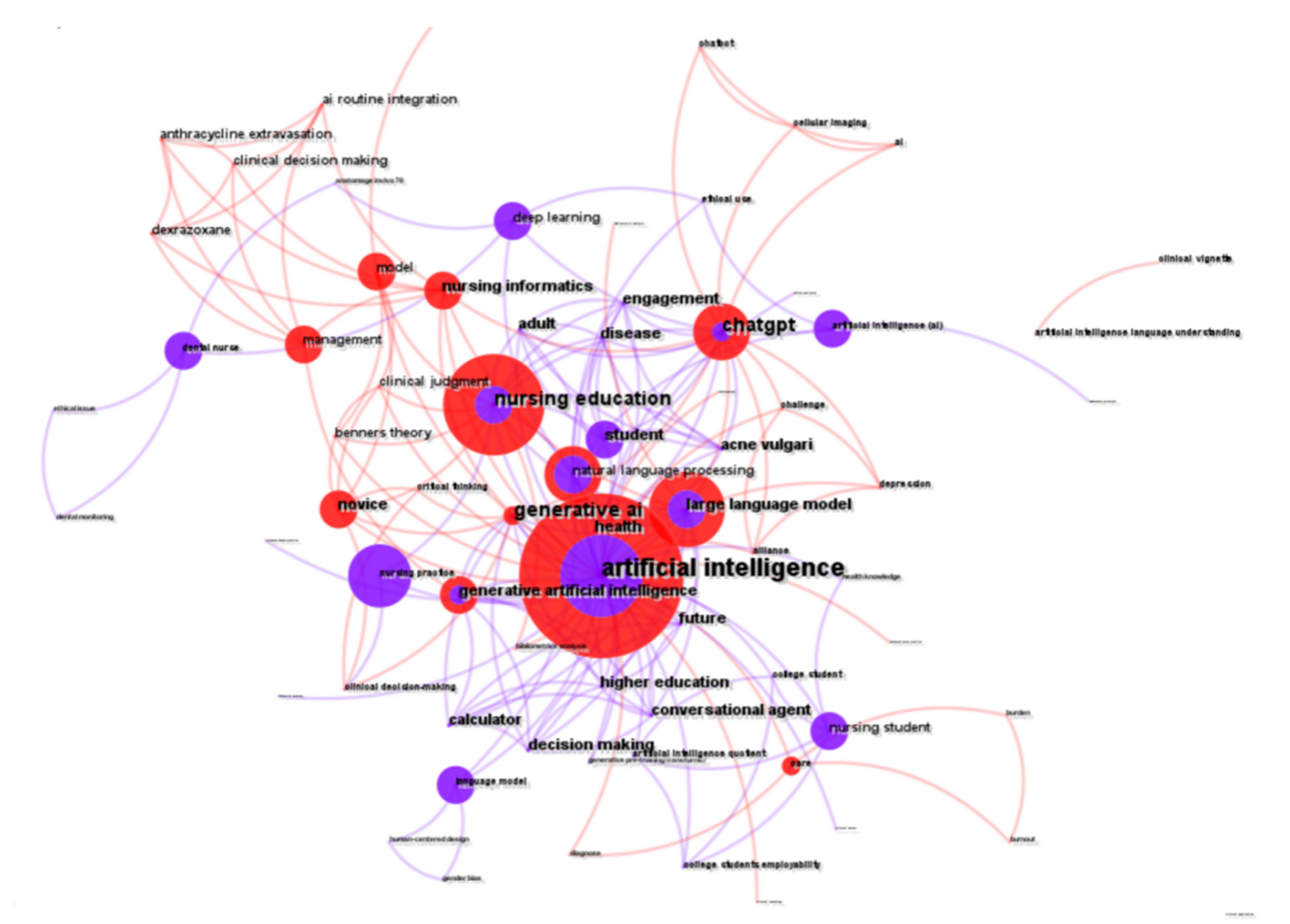


FIGURE 10
Co-occurrence network of keywords.

The findings of the most relevant sources indicate that similar results are achieved when sources are analyzed from different perspectives. The leading sources encompass a variety of topics, including nursing education (32), cardiovascular care (33), emergency care (34), perianesthesia nursing (35), psychiatric health (36), and family and community care (37, 38), thereby highlighting the extensive applicability of the ChatGPT within the nursing profession. Nine of the top 20 journals focus on education, reflecting ChatGPT's current areas of emphasis in conjunction with nursing. Among these educational articles, the majority conclude that ChatGPT is feasible for nursing education; however, they also acknowledge limitations and ethical dilemmas that could inform updates to the ChatGPT version (7, 9, 39). In our bibliometric study, we employed Bradford's Law to categorize the sources into distinct zones, which aids in identifying the principal journals within a specific subject area (40). White (66) noted that Bradford's Law could lead to the misconception that articles published in the primary journals of a field are generally of higher quality than those distributed across peripheral journals (41). To mitigate the inaccuracies arising from this bias, we concurrently assessed the h-index, g-index, and m-index. The results indicated that sources in Zone 1 exhibited a significant impact, thereby enabling us to further identify high-quality sources within the domain of ChatGPT in nursing.

The literature on the measurement of affiliation and country indicates that the high volume of articles not only reflects a strong institutional emphasis on this area of research but also suggests access to essential resources, such as funding, talent, and data, which are crucial for sustained academic productivity (42, 43). The presence of institutions from the United States, China, and Saudi Arabia further underscores that the exploration of ChatGPT in nursing research is a global phenomenon characterized by geographical diversification. These institutions possess robust interdisciplinary collaborations that integrate expertise from both nursing and computer science, fostering innovation and the exchange of ideas (43). The SCP highlights the strong national research capabilities and initiatives of the USA and China in this interdisciplinary field. Additionally, the MCP highlights the role of these two countries in international collaboration, facilitating the global exchange of knowledge and expertise in this domain (38). For example, China and the United States have collaborated on a multidisciplinary approach to address the opportunities and challenges posed by artificial intelligence (44), as well as the applications of ChatGPT in nursing education (45). Academic collaboration among various countries or regions can significantly

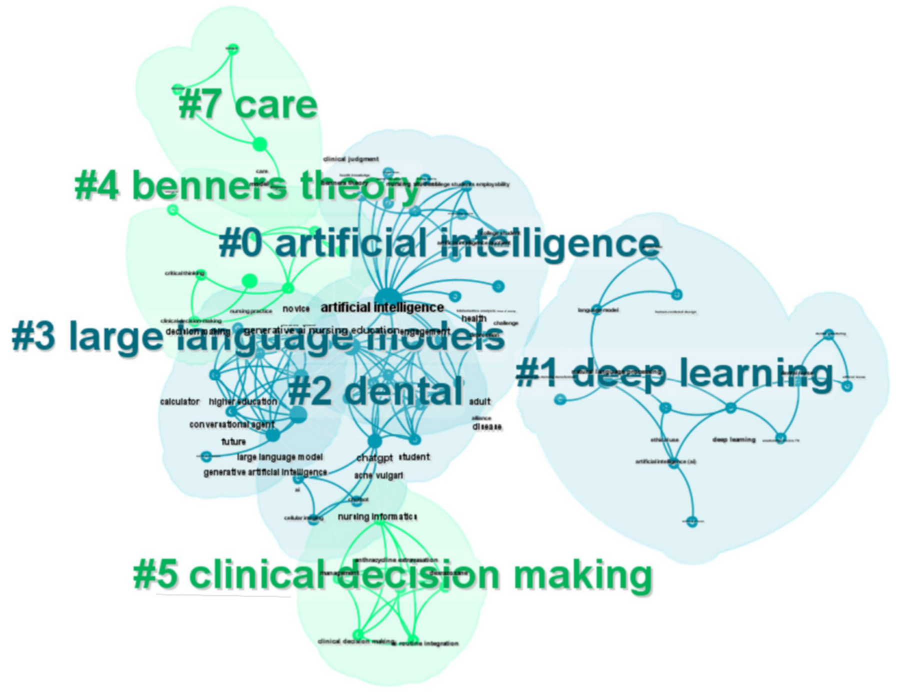


FIGURE 11
Clustering network of keywords.

TABLE 3 A list of the clustering network.

Cluster ID	Size	Silhouette	Year	Cluster label	Top terms (LSI)
#0	17	0.846	2023	Artificial intelligence	Artificial intelligence; the-art language processing; nursing students; college student; nursing education; nursing education research; health knowledge; social responsibility; nursing informatics
#1	13	0.966	2024	Deep learning	Deep learning; image processing; machine learning; dental nurse; image segmentation; gender bias; machine translation; language models; human-centered design
#2	11	0.976	2023	Dental	Nursing education; artificial intelligence; narrative review; pedagogical approach; student assessment; proximal development; vygotskys zone; skin symptoms
#3	9	0.976	2023	Large language models	Large language models; generative artificial intelligence; conversational agent; nursing informatics; bibliometric analysis; scoping review
#4	8	0.816	2023	Benners theory	Skill acquisition; nursing education; benners theory; artificial intelligence; clinical decision-making
#5	6	0.954	2023	Clinical decision making	LLMS feasibility; AI routine integration; methodology; clinical decision making; healthcare innovation; nursing informatics; safety; multidisciplinary approach; multi-parametric analysis
#7	4	0.973	2023	Care	Burnout; burden; care; nurse

enhance the dissemination of knowledge and foster academic exchange (46). Although ChatGPT is an emerging technology, collaboration among nations across all continents underscores the globalization and significance of ChatGPT research in the field of nursing. Advancements in technology and the deepening of research efforts suggest that such cooperation will become increasingly essential in the future.

The development of large AI models necessitates closer and more intense collaboration among domain experts, as well as the gradual establishment of regulations (47). In the author collaboration network analyzed by Citespace, the collaboration density is measured at 0.0678, indicating that the authors' cooperative efforts are dispersed (48), which may stem from differing valuations of the subject matter by each team. The group comprising Moons and Van Bulck primarily focuses on the trustworthiness and value of ChatGPT (6, 49). Tam et al.'s group focused on nursing education in the age of artificial intelligence (7), whereas Allen group collaborated on mental health (50). This suggests that, despite the multidisciplinary nature of the research topics, they are predominantly investigated independently

Keywords	Year	Strength	Begin	End	2023 - 2024
nursing practice	2023	0.55	**2023**	2024	
student	2023	0.36	**2023**	2024	
dental nurse	2023	0.36	**2023**	2024	
nursing student	2023	0.36	**2023**	2024	
language model	2023	0.36	**2023**	2024	
artificial intelligence (ai)	2023	0.36	**2023**	2024	
deep learning	2023	0.36	**2023**	2024	
conversational agent	2023	0.18	**2023**	2024	
gpt 4	2023	0.18	**2023**	2024	
calculator	2023	0.18	**2023**	2024	

FIGURE 12
The top 10 keywords with the strongest citation bursts.

by various teams across different disciplines. It is plausible that the extensive scope of nursing as a subject area has prompted these teams to explore specific nursing specialties in divergent directions. Although there appears to be limited collaboration among the teams, this does not necessarily imply a deficiency of teamwork in the research concerning ChatGPT in nursing. As research on ChatGPT intensifies and the volume of studies within the same specialty increases, the focus may gradually shift from assessing the feasibility of ChatGPT in nursing to deep learning itself. Consequently, the trend of collaboration among different teams may increase in the future.

The analysis of hotspot evolution revealed that ChatGPT has been extensively studied within the realms of nursing education, clinical decision-making, and management, highlighting its significant application in the nursing field. As an emerging artificial intelligence technology, ChatGPT has spurred advancements in both nursing education and clinical decision-making (51, 52). The interconnectedness of nursing and ChatGPT is evident, as both domains appear to support each other's progression. By utilizing the keyword clustering knowledge graph and collinear network clustering table, it becomes clear that most clusters exhibit overlap. Among the seven identified clusters, the clusters pertaining to artificial intelligence, dental, large language models, and Benner's theory are closely interconnected, whereas the clusters related to deep learning, clinical decision-making, and care are more peripheral due to their looser connections. This observation indicates that current research is still in the early stages of foundational data research and technological development. ChatGPT is still in its preliminary stages, and the theoretical foundations and data models of the four interconnected clusters are expected to maintain a dominant position in future research. The dental cluster is closely linked to other clusters, primarily emphasizing nursing education. This alignment indicates that nursing education is consistent with current research hotspots and focal points. Additionally, topics such as deep learning, clinical decision-making, and patient care reflect the continuous emergence of new areas of inquiry. ChatGPT is anticipated to engage in more comprehensive collaborative research grounded in the theoretical frameworks of clinical decision-making and patient care. Currently, the application of ChatGPT in nursing primarily revolves around nursing education, clinical decision-making, clinical nursing practice, automated writing, and addressing common nursing inquiries. In the realm of nursing education, ChatGPT applications include vocational examinations, application attitude surveys, educational practices, and teaching design, among others. ChatGPT as a representative product, its application and research results also show the main advantages of "Anthropic Claude," "Google Gemini" and other generative AI in the field of nursing, as well as their usability and research prospects. The results and discussion indicate that ChatGPT offers significant advantages in the nursing field, including user-friendliness, rapid response capabilities, data-driven content generation, and enhanced efficiency. A representative example is the applications and research findings related to ChatGPT, which also emphasize the relevance of other generative AI models, such as "Anthropic Claude" and "Google Gemini," within the nursing domain, thereby highlighting their usability and research potential.

Nearly all the articles evaluated the risks associated with the ChatGPT. Perspectives on this issue vary; some scholars adopt a negative stance, indicating that further research is necessary (53–55), whereas the majority advocate embracing

10.3389/fmed.2024.1521712

the challenge and seizing the opportunities presented (13, 39, 56). Ethical considerations are a crucial element that must not be overlooked. Issues related to reliance on technology (57), misdiagnosis and treatment errors (58), data security breaches (59), and the trustworthiness of patients (60) must be addressed when utilizing ChatGPT. Future studies should continue to examine the ethical ramifications of artificial intelligence concerning patient confidentiality and data protection (2), the accuracy and credibility of information (61), autonomy in decision-making (62), and transparency (63), to enhance the integration of ChatGPT within the nursing field.

We acknowledge the limitations of our study. (i) CiteSpace's dependence on specific data sources is primarily evident in its connection to particular databases, notably the Web of Science (WoS) and others. This dependence constrains the scope and comprehensiveness of CiteSpace's data collection and analysis, potentially omitting relevant literature that is not included in these databases (21, 64). Our investigation is confined to publications included in the Web of Science Core Collection (WoSCC), which does not encompass all journals; this may lead to the oversight of articles in other databases, such as Scopus and PubMed. Nevertheless, the WoSCC is a comprehensive and well-organized database that is extensively utilized across various scientific disciplines, and the quality of papers within this source is widely recognized and employed in most scientometric studies. (ii) While CiteSpace is capable of identifying significant patterns and trends within scientific literature, it does not function at its full potential for conducting in-depth analyses of specific fields or subjects (65). Therefore, it may be necessary to employ additional tools or techniques to gain more comprehensive insights. To complement these limitations, the bibliometric package was applied to conduct more in-depth statistical analysis of the data, such as the top 20 most relevant affiliations, Bradford's Law, the impact of sources, SCP, MCP, etc. So as to dig out deeper academic information. (iii) Our analysis was limited to English-language articles published in reputable peer-reviewed scientific journals, which may introduce potential publication bias.

Conclusions

ChatGPT is an emerging tool in the field of nursing and is currently in the basic research stage. To our knowledge, the present study represents the first bibliometric analysis of the application of the ChatGPT in nursing, identifying key contributors, including countries, authors, and journals. Our findings indicate that the United States and China are the leading countries in terms of publication volume and that international collaboration is robust. However, there is limited cooperation among author groups, which can be attributed to differences in specialties. Therefore, it is essential for nurses from various specialties to collaborate in exploring the diverse applications of ChatGPT within their fields, thereby facilitating the further development and enhancement of this technology. Our hotspot analysis revealed that publications on ChatGPT in nursing have focused on two main themes: (1) the deep learning of ChatGPT in nursing and (2) the feasibility of its application. In addition to discussing the use of ChatGPT in nursing, we provide several suggestions for academics to conduct empirical studies in this area: (1) The literature currently lacks randomized controlled trials and qualitative studies; thus, the effects of ChatGPT could be evaluated via a variety of research designs. (2) By integrating different artificial intelligence tools (such as DeepL, especially AI, and Resemble AI) and technologies (including virtual reality, augmented reality, and mobile applications) with ChatGPT, we can investigate the effects of these combinations on nursing practice. (3) The literature on the application of ChatGPT in nursing tends to be fragmented, particularly concerning foundational data studies. It is feasible to enhance the application of ChatGPT across various practice areas and identify commonalities through collaborative efforts. By addressing these research priorities, we can substantially advance our understanding of the potential of ChatGPT as a tool in nursing and develop a diverse range of strategies.

Author contributions

PW: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft. QZ: Data curation, Formal analysis, Software, Validation, Visualization, Writing – original draft. WZ: Methodology, Software, Visualization, Writing – original draft. JS: Conceptualization, Methodology, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors would like to thank all the participants who contributed to this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

1. Wang T, Mu J, Chen J, Lin CC. Comparing ChatGPT and clinical nurses' performances on tracheostomy care: a cross-sectional study. *Int J Nurs Stud Adv.* (2024) 6:100181. doi: 10.1016/j.ijnsa.2024.100181

2. Kleib M, Darko EM, Akingbade O, Kennedy M, Majekodunmi P, Nickel E, et al. Current trends and future implications in the utilization of ChatGPT in nursing: a rapid review. *Int J Nurs Stud Adv.* (2024) 7:100252. doi: 10.1016/j.ijnsa.2024.100252

3. Luo Y, Miao Y, Zhao Y, Li J, Wu Y. Exploring the current applications and effectiveness of ChatGPT in nursing: an integrative review. *J Adv Nurs.* (2024) 20:16628. doi: 10.1111/jan.16628

4. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-35, ChatGPT-40, and Google Bard. *EBioMedicine.* (2023) 95:104770. doi: 10.1016/j.ebiom.2023.104770

5. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F. Applications of large language models in cancer care: current evidence and future perspectives. *Front Oncol.* (2023) 13:1268915. doi: 10.3389/fonc.2023.1268915

6. Moons P, Van Bulck L. ChatGPT: can artificial intelligence language models be of value for cardiovascular nurses and allied health professionals. *Eur J Cardiovasc Nurs.* (2023) 22:e55–e9. doi: 10.1093/eurjcn/zvad022

7. Tam W, Huynh T, Tang A, Luong S, Khatri Y, Zhou W. Nursing education in the age of artificial intelligence powered Chatbots (AI-Chatbots): are we ready yet? *Nurse Educ Today.* (2023) 129:105917. doi: 10.1016/j.nedt.2023.105917

8. Chavez MR, Butler TS, Rekawek P, Heo H, Kinzler WL. Chat generative pre-trained transformer: why we should embrace this technology. *Am J Obstet Gynecol.* (2023) 228:706–11. doi: 10.1016/j.ajog.2023.03.010

9. Ahmed SK. The impact of ChatGPT on the nursing profession: revolutionizing patient care and education. *Ann Biomed Eng.* (2023) 51:2351–2. doi: 10.1007/s10439-023-03262-6

10. O'Connor S. Open artificial intelligence platforms in nursing education: tools for academic progress or abuse? *Nurse Educ Pract.* (2023) 66:103537. doi: 10.1016/j.nepr.2022.103537

11. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health.* (2023) 2:e0000198. doi: 10.1371/journal.pdig.0000198

12. Stephens LD, Jacobs JW, Adkins BD, Booth GS. Battle of the (Chat)Bots: comparing large language models to practice guidelines for transfusion-associated graft-versus-host disease prevention. *Transfus Med Rev.* (2023) 37:150753. doi: 10.1016/j.tmrv.2023.150753

13. Huang H. Performance of ChatGPT on registered nurse license exam in Taiwan: a descriptive study. *Healthcare.* (2023) 11:1625. doi: 10.20944/preprints202309.1625.v1

14. Berse S, Akça K, Dirgar E, Serin EK. The role and potential contributions of the artificial intelligence language model ChatGPT. *Ann Biomed Eng.* (2024) 52:130–3. doi: 10.1007/s10439-023-03296-w

15. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLoS Digit Health.* (2023) 2:e0000205. doi: 10.1371/journal.pdig.0000205

16. D'Amico RS, White TG, Shah HA, Langer DJI. Asked a ChatGPT to write an editorial about how we can incorporate chatbots into neurosurgical research and patient care.... *Neurosurgery.* (2023) 92:663–4. doi: 10.1227/neu.0000000000002414

17. Scerri A, Morin KH. Using chatbots like ChatGPT to support nursing practice. *J Clin Nurs.* (2023) 32:4211–3. doi: 10.1111/jocn.16677

18. Meert D, Torabi N, Costella J. Impact of librarians on reporting of the literature searching component of pediatric systematic reviews. *J Med Libr Assoc.* (2016) 104:267–77. doi: 10.3163/1536-5050.104.4.004

19. Birkle C, Pendlebury DA, Schnell J, Adams J. Web of Science as a data source for research on scientific and scholarly activity. *Quant Sci Stud.* (2020) 1:363–76. doi: 10.1162/qss_a_00018

20. Wang Y, Yang L, Lei Y. The global status of nursing research on hemodialysis: a bibliometric and visualized analysis. *Medicine.* (2024) 103:39707. doi: 10.1097/MD.0000000000039707

21. Tomaszewski R. Visibility, impact, and applications of bibliometric software tools through citation analysis. *Scientometrics.* (2023) 128:4007–28. doi: 10.1007/s11192-023-04725-2

22. Aria M, Cuccurullo C. *Bibliometrix*: an R-tool for comprehensive science mapping analysis. *J Informetr.* (2017) 11:959–75. doi: 10.1016/j.joi.2017.08.007

23. Doyon O, Raymond L. Surveillance and patient safety in nursing research: a bibliometric analysis from 1993 to 2023. *J Adv Nurs.* (2024) 80:777–88. doi: 10.1111/jan.15793

24. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: an overview and guidelines. *J Bus Res.* (2021) 133:285–96. doi: 10.1016/j.jbusres.2021.04.070

25. Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F. Science mapping software tools: review, analysis, and cooperative study among tools. *J Am Soc Inform Sci Technol.* (2011) 62:1382–402. doi: 10.1002/asi.21525

26. Desai N, Veras L, Gosain A. Using Bradford's law of scattering to identify the core journals of pediatric surgery. *J Surg Res.* (2018) 229:90–5. doi: 10.1016/j.jss.2018.03.062

27. Mondal H, Deepak KK, Gupta M, Kumar R. The h-index: understanding its predictors, significance, and criticism. *J Fam Med Prim Care.* (2023) 12:2531–7. doi: 10.4103/jfmpc.jfmpc_1613_23

28. van Eck NJ, Waltman L. Generalizing the *h*- and *g*- indices. *J Informetr.* (2008) 2:263–71. doi: 10.1016/j.joi.2008.09.004

29. Hobensack M, von Gerich H, Vyas P, Withall J, Peltonen L-M, Block LJ, et al. A rapid review on current and potential uses of large language models in nursing. *Int J Nurs Stud.* (2024) 154:104753. doi: 10.1016/j.ijnurstu.2024.104753

30. Liu J, Peng S, Liu S. Nurses' perspectives on ChatGPT: a survey study. *Stud Health Technol Inform.* (2024) 315:661–2. doi: 10.3233/SHTI240267

31. Krüger L, Krotsetis S, Nydahl P. ChatGPT: curse or blessing in nursing care? *Medizinische Klinik-Intensivmedizin Und Notfallmedizin.* (2023) 118:534–9. doi: 10.1007/s00063-023-01038-3

32. Athilingam P, He HG. ChatGPT in nursing education: opportunities and challenges. *Teach Learn Nurs.* (2024) 19:97–101. doi: 10.1016/j.teln.2023.11.004

33. Cerrato PL, Halamka JD. How AI drives innovation in cardiovascular medicine. *Front Cardiovasc Med.* (2024) 11:1397921. doi: 10.3389/fcvm.2024.1397921

34. Veseli E, Tovani-Palone MR, Veseli A, Kastrati L. Should ChatGPT have some applicability in the management of emergency dental care for emigrant adults and children? *J Contemp Dent Pract.* (2023) 24:819–20. doi: 10.5005/jp-journals-10024-3576

35. Nilsson U. Dear ChatGPT, do we need perianesthesia nurses in the PACU? *J Perianesth Nurs.* (2023) 38:830–1. doi: 10.1016/j.jopan.2023.07.003

36. Dergaa I, Fekih-Romdhane F, Hallit S, Loch AA, Glenn JM, Fessi MS, et al. ChatGPT is not ready yet for use in providing mental health assessment and interventions. *Front Psychiat.* (2024) 14:1277756. doi: 10.3389/fpsyt.2023.1277756

37. Govender I, Tumbo J, Mahadeo S. Using ChatGPT in family medicine and primary health care. *South Afri. Fam. Pract.* (2024) 66:5895. doi: 10.4102/safp.v66i1.5895

38. Gandhi AP, Joesph FK, Rajagopal V, Aparnavi P, Katkuri S, Dayama S, et al. Performance of ChatGPT on the India undergraduate community medicine examination: cross-sectional study. *JMIR Format Res.* (2024) 8:49964. doi: 10.2196/49964

39. Bumbach MD. The use of AI powered ChatGPT for nursing education. *J Nurs Educ.* (2024) 2024:1–4. doi: 10.3928/01484834-20240318-04

40. Venable GT, Shepherd BA, Loftis CM, McClatchy SG, Roberts ML, Fillinger ME, et al. Bradford's law: identification of the core journals for neurosurgery and its subspecialties. *J Neurosurg.* (2016) 124:569–79. doi: 10.3171/2015.3.JNS15149

41. Nicolaisen J, Hjorland B. Practical potentials of Bradford's law: a critical examination of the received view. *J Document.* (2007) 63:359–77. doi: 10.1108/00220410710743298

42. González-Betancor SM, Dorta-González P. An indicator of the impact of journals based on the percentage of their highly cited publications. *Onl Inform Rev.* (2017) 41:398–411. doi: 10.1108/OIR-01-2016-0008

43. Kemec A, Altinay AT. Sustainable energy research trend: a bibliometric analysis using VOSviewer, RStudio Bibliometrix, and CiteSpace software tools. *Sustainability.* (2023) 15:43618. doi: 10.3390/su15043618

44. Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, et al. "So what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int J Inform Manag.* (2023) 71:102642. doi: 10.1016/j.ijinfomgt.2023.102642

45. Liu J, Liu F, Fang J, Liu S. The application of Chat Generative Pre-trained Transformer in nursing education. *Nurs Outl.* (2023) 71:102064. doi: 10.1016/j.outlook.2023.102064

46. Zhong ZL, Guo H, Qian K. Deciphering the impact of machine learning on education: insights from a bibliometric analysis using bibliometrix R-package. *Educ Inform Technol.* (2024) 24:8. doi: 10.1007/s10639-024-12734-8

47. Qiu J, Li L, Sun J, Peng J, Shi P, Zhang R, et al. Large AI models in health informatics: applications, challenges, and the future. *IEEE J Biomed Health Inform.* (2023) 27:6074–87. doi: 10.1109/JBHI.2023.3316750

48. Dang Q, Luo ZM, Ouyang CH, Wang L. First systematic review on health communication using the CiteSpace software in China: exploring its research hotspots and frontiers. *Int J Environ Res Publ Health.* (2021) 18:8. doi: 10.3390/ijerph1824 13008

49. Van Bulck L, Moons P. What if your patient switches from Dr. Google to Dr ChatGPT? a vignette-based survey of the trustworthiness, value, and danger of

ChatGPT-generated responses to health questions. *Eur J Cardiovasc Nurs.* (2024) 23:95–8. doi: 10.1093/eurjcn/zvad038

50. Woodnutt S, Allen C, Snowden J, Flynn M, Hall S, Libberton P, et al. Could artificial intelligence write mental health nursing care plans? *J Psychiatr Ment Health Nurs.* (2024) 31:79–86. doi: 10.1111/jpm.12965

51. Maitland A, Fowkes R, Maitland S. Can ChatGPT pass the MRCP (UK) written examinations? analysis of performance and errors using a clinical decision-reasoning framework. *Br Med J Open.* (2024) 14:80558. doi: 10.1136/bmjopen-2023-080558

52. Saban M, Dubovi I. A comparative vignette study: evaluating the potential role of a generative AI model in enhancing clinical decision-making in nursing. *J Adv Nurs.* (2024) 2024:16101. doi: 10.1111/jan.16101

53. Milton CL. ChatGPT and forms of deception. *Nurs Sci Q.* (2023) 36:232–3. doi: 10.1177/08943184231169753

54. Abdulai AF, Hung L. Will ChatGPT undermine ethical values in nursing education, research, and practice? *Nurs Inq.* (2023) 30:e12556. doi: 10.1111/nin.12556

55. Choi EPH, Lee JJ, Ho MH, Kwok JYY, Lok KYW. Chatting or cheating? the impacts of ChatGPT and other artificial intelligence language models on nurse education. *Nurse Educ Today.* (2023) 125:105796. doi: 10.1016/j.nedt.2023.105796

56. Lin HL, Liao LL, Wang YN, Chang LC. Attitude and utilization of ChatGPT among registered nurses: a cross-sectional study. *Int Nurs Rev.* (2024) 2024:13012. doi: 10.1111/inr.13012

57. Stahl BC, Eke D. The ethics of ChatGPT—exploring the ethical issues of an emerging technology. *Int J Inform Manag.* (2024) 74:102700. doi: 10.1016/j.ijinfomgt.2023.102700

58. Harada Y, Suzuki T, Harada T, Sakamoto T, Ishizuka K, Miyagami T, et al. Performance evaluation of ChatGPT in detecting diagnostic errors and their contributing factors: an analysis of 545 case reports of diagnostic errors. *Br Med J Open Qual.* (2024) 13:2654. doi: 10.1136/bmjoq-2023-002654

59. Huang K, Zhang F, Li Y, Wright S, Kidambi V, Manral V. Security and privacy concerns in ChatGPT. In: Huang K, Wang Y, Zhu F, Chen X, Xing C, editors. *Beyond AI: ChatGPT, Web3, and the Business Landscape of Tomorrow.* Cham: Springer Nature Switzerland (2023). p. 297–328.

60. Chen SY, Kuo HY, Chang SH. Perceptions of ChatGPT in healthcare: usefulness, trust, and risk. *Front Publ Health.* (2024) 12:1457131. doi: 10.3389/fpubh.2024.1457131

61. Wang CY, Liu SR, Yang H, Guo JL, Wu YX, Liu JL. Ethical considerations of using ChatGPT in health care. *J Med Internet Res.* (2023) 25:48009. doi: 10.2196/48009

62. Li XE, Yu YD, Huang ML. A comparative vignette study: evaluating the potential role of a generative AI model in enhancing clinical decision-making in nursing. *J Adv Nurs.* (2024) 80:4752. doi: 10.1111/jan.16146

63. Tang AR, Li KK, Kwok KO, Cao LJ, Luong S, Tam W. The importance of transparency: declaring the use of generative artificial intelligence (AI) in academic writing. *J Nurs Scholar.* (2024) 56:314–8. doi: 10.1111/jnu.12938

64. Shen YH, Huang LH, Wu XS. Visualization analysis on the research topic and hotspot of online learning by using CiteSpace-Based on the Web of Science core collection (2004-2022). *Front Psychol.* (2022) 13:1059858. doi: 10.3389/fpsyg.2022.1059858

65. Cheng SY, Zhang JC, Wang GX, Zhou Z, Du J, Wang LJ, et al. Cartography and neural networks: a scientometric analysis based on CiteSpace. *ISPRS Int J Geo-Inform.* (2024) 13:60178. doi: 10.3390/ijgi13060178

66. White HD. 'Bradfordizing' search output: how it would help online users. *Online Rev.* (1981) 5:47–54. doi: 10.1108/eb024050

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 17 January 2025
DOI 10.3389/fmed.2024.1496866

Check for updates

OPEN ACCESS

EDITED BY
Thomas F. Heston,
University of Washington, United States

REVIEWED BY
Serafeim A. Triantafyllou,
Aristotle University of Thessaloniki, Greece
Pranavsingh Dhunnoo,
Atlantic Technological University, Ireland

*CORRESPONDENCE
Jung-Hyun Lee
✉ jung-hyun.lee@downstate.edu

RECEIVED 15 September 2024
ACCEPTED 30 December 2024
PUBLISHED 17 January 2025

CITATION
Lee J-H, Choi E, Angulo SL, McDougal RA and Lytton WW (2025) Neurological history both twinned and queried by generative artificial intelligence.
Front. Med. 11:1496866.
doi: 10.3389/fmed.2024.1496866

Neurological history both twinned and queried by generative artificial intelligence

Jung-Hyun Lee[1,2,3]*, Eunhee Choi[4], Sergio L. Angulo[1,2], Robert A. McDougal[5,6,7,8] and William W. Lytton[1,2,9]

[1]Department of Neurology, State University of New York Downstate Health Sciences University, Brooklyn, NY, United States, [2]Department of Neurology, Kings County Hospital, Brooklyn, NY, United States, [3]Department of Neurology, Maimonides Medical Center, Brooklyn, NY, United States, [4]Department of Internal Medicine, Lincoln Medical Center, Bronx, NY, United States, [5]Department of Biostatistics, Yale School of Public Health, Yale University, New Haven, CT, United States, [6]Computational Biology and Bioinformatics, Yale University, New Haven, CT, United States, [7]Wu Tsai Institute, Yale University, New Haven, CT, United States, [8]Biomedical Informatics and Data Science, Yale School of Medicine, Yale University, New Haven, CT, United States, [9]Department of Physiology and Pharmacology, State University of New York Downstate Health Sciences University, Brooklyn, NY, United States

Background and objectives: We propose the use of GPT-4 to facilitate initial history-taking in neurology and other medical specialties. A large language model (LLM) could be utilized as a digital twin which could enhance queryable electronic medical record (EMR) systems and provide healthcare conversational agents (HCAs) to replace waiting-room questionnaires.

Methods: In this observational pilot study, we presented verbatim history of present illness (HPI) narratives from published case reports of headache, stroke, and neurodegenerative diseases. Three standard GPT-4 models were designated Models *P*: patient digital twin; *N*: neurologist to query Model P; and *S*: supervisor to synthesize the N-P dialogue into a derived HPI and formulate the differential diagnosis. Given the random variability of GPT-4 output, each case was presented five separate times to check consistency and reliability.

Results: The study achieved an overall HPI content retrieval accuracy of 81%, with accuracies of 84% for headache, 82% for stroke, and 77% for neurodegenerative diseases. Retrieval accuracies for individual HPI components were as follows: 93% for chief complaints, 47% for associated symptoms and review of systems, 76% for relevant symptom details, and 94% for histories of past medical, surgical, allergies, social, and family factors. The ranking of case diagnoses in the differential diagnosis list averaged in the 89th percentile.

Discussion: Our tripartite LLM model demonstrated accuracy in extracting essential information from published case reports. Further validation with EMR HPIs, and then with direct patient care will be needed to move toward adaptation of enhanced diagnostic digital twins that incorporate real-time data from health-monitoring devices and self-monitoring assessments.

KEYWORDS

neurology–clinical, stroke, headache, neurodegenerative disease, large language model (LLM), history taking

Introduction

Contrasting with the gradual drift of many medical specialties toward laboratory-dependent diagnosis, neurology, psychiatry, and primary care remain heavily dependent on careful history-taking. In the case of neurology, history guides the focused exam—a dementia history triggers an exam that's entirely different from that for a motorcycle accident (1). This history-dependence extends to the choice of labs and imaging as well. The complexities of neurological history reflect the wide variety in presentation across (1) multiple neural systems: CNS, PNS, ANS; (2) multiple structures within the brain, brainstem, cord, etc.; and (3) multiple etiologies: vascular, inflammatory, traumatic, infectious, etc. Neurologists are extensively trained to follow the logic of localization and etiology when taking a history. Non-neurologist practitioners, on the other hand, may feel overwhelmed or uncertain when managing a neurological patient (2).

Although questionnaires play a role in neurological subspecialties, a general neurology questionnaire is not feasible due to the breadth of possible problems to be considered. Compared to either close-ended (multiple choice) or open-ended (fill in the blanks) static health questionnaires, computer-based digital questionnaires can give more flexibility by following a sequence of flow-charted questions comparable to what is provided in board exams. These have shown some degree of utility in primary care settings (3–5). Large language models (LLMs) such as generative pre-trained transformers (GPTs) can add further flexibility in questioning when used as *healthcare conversational agents* (HCAs) (6–8).

To evaluate GPT-4 proficiency in general neurology history-taking, we considered several alternatives for finding our historian patients. Since this was a pilot study and associated AI safety and monitoring were not secured, clinic patients were not used. Therefore, as in our previous study, we started with published cases from the literature (9). Using ourselves as intermediaries to answer questions risked allowing too much additional medical knowledge to creep into the responses. Simple transcribing of case reports into the AI model would not allow the desired interactivity of actual patient communication. After trialing these approaches, we decided to utilize GPT-4 as the patient partner and as a *digital twin* that would provide the patient part. Evaluation of the GPT-4 output showed that the responses were reasonable utterances comparable to those expected from a patient. We also decided to use a limited set of neurological diagnoses where history plays a major role in determining a provisional diagnosis.

We found that GPT-4 could provide adequate initial history-taking that could potentially aid in history-taking by healthcare workers. This would provide efficiency in workflow, especially in busy hospitals or clinics where patients wait long, elongated periods of time to get initial evaluations. AI tools could provide better use of time for both the patient and the physician. Additionally, utilizing LLM to create digital twins could provide a possible future of queryable electronic medical record (EMR) systems or be utilized further to train HCAs.

Method

Study design

We chose cases of common neurological disorders, including headache, stroke, and neurodegenerative diseases, from PubMed Central search for ("BMC Neurology"[Journal] OR "J Med Case Rep"[Journal] OR "Medicine (Baltimore)"[Journal]) AND "case report"[Title] AND [CASE]. As in our previous study, these journals were chosen since they provided relatively detailed case report descriptions rather than abbreviated, diagnosis-targeted descriptions (9). The [CASE] search used MeSH (Medical Subject Headings) terms: "Stroke"[MeSH Terms]; "Headache"[MeSH Terms]; "Dementia"[MeSH Terms] OR "Parkinson Disease"[MeSH Terms] for neurodegenerative diseases. The excluded articles were those describing treatment failures, complications, unclear clinical descriptions, non-neurological conditions, coexisting neurological conditions, and duplicate diagnoses. For each disease category, we randomly selected five case reports.

We utilized three identical, non-pretrained GPT-4 (version 4-0125-preview) models for Model N (neurologist), Model P (patient), and Model S (supervisor). Model N was limited to 30 questions to form a diagnostic impression. Subsequently, Model S synthesized a summarized History of Present Illness (HPI) and established a list of differential diagnoses based on the Model N and Model P dialogue.

To assess data consistency, the simulated patient-doctor interactions were replicated five times per case. The quality evaluation involved comparing the HPI generated by GPT-4 with a rubric that summarized the key elements from the original case report. To further assess the quality of history-taking, the original case report's diagnosis was compared against the GPT-4's list of differential diagnoses, noting its percentile ranking when applicable. Our grading permitted the acceptance of broader diagnostic terminology where GPT-4 could not clinically identify the exact diagnosis within the differential list.

Prompt engineering

Zero-shot instruction prompt was used to configure the N and P models (Figure 1), with the prompt providing: 'role' to distinguish the expected behaviors; 'setting' to provide the context in which the models would operate; 'task' to set the objectives; 'detailed instructions' for additional behavioral guidance.

Model N was required to provide questions about symptom onset, characteristics, evolution, and associated symptoms, along with complete medical, surgical, medication, social, allergy, and family histories. The model was instructed to probe further into any reported symptoms by requesting detailed descriptions. The goal was to gather a focused history rather than a full history, limiting the maximum number of queries to 30 for efficient history collection. After reaching a provisional diagnosis, the conversation ended by stating 'terminate' by Model N.

Model P was required to adhere strictly to the case document and instructed to use the keyword "Negative" if queried information was unavailable, which later in the summarization process by Model S would be interpreted as either unavailable information or negative pertinent. Model P was prompted to use a simple, conversational communication style; to avoid medical terminology and repetition; and to give an initial chief complaint of up to two symptoms.

The Model S started with a zero-shot instruction prompt to obtain the generation of an HPI summary. Subsequently, Chain-of-Thought (CoT) prompting was used to obtain a step-by-step process of clinical reasoning with a discussion of potential diagnoses leading to a diagnosis list; followed by Tree-of-Thought (ToT) to require

10.3389/fmed.2024.1496866

Model N System Prompt		
Imagine you are a neurologist taking care of patients with neurological diseases or disorders. Do not mention you are an AI.	Role	Zero-shot instruction prompt
You can always ask for collateraal information from another person such as the EMT, family member, or friend.	Setting	
Your role is to ask the patient detailed questions to yield essential information for diagnosis.	Task	
Obtain complete history including symptom onset, character, progression, associated symptoms relevant to the chief complaints. Do not omit any of these details. Obtain complete 'past medical history', 'past surgical history', 'medications', 'social history', 'allergies', 'family history'. Do not omit any of these details. Investigate further into the details of symptoms provided by the patient by saying "Can you tell me more about ...". Try to obtain detail as much as possible. You can only ask one question at a time. Do not ask duplicate questions. Try to yield as much tailored history within the 30 questions as possible. Do not explain anything as you proceed. If you determine your history taking is complete and have a tentative diagnosis and differential terminate conversation by saying 'terminate'.	Instructions	
Model P System Prompt		
Imagine you are a patient or the patient's family member or friend or EMT to provide information about the patient. Below is the 'History' of that patient.	Role, Setting, Task	
Do not make up information that is not provided in the 'History' If the information is not provided within the 'History', say "Negative". Answer like a human. When asked 'What brings you here today?', give only the chief complaint, this should be less than 3 symptoms from the initial sentence of 'History'. Answer to the questions that the neurologist asks you based on the 'History" below in layman terms. Do not give additional information not provided in the question. Do not use professional terms describing the symptoms. Do not give information you already provided based on conversation history.	Instructions	
History: (case report HPI input)	Case HPI	
Model S Prompt		
Conversation: (patient-neurologist script) Summarize the above conversation as following format: [HPI:], [Past medical history:], [Past surgical history:], [Allergies:], [Medications:], [Social history:], [Family history:] Do not omit any detail.	HPI Summarization	
Let's do it step-by-step. Internally two neurologists discuss about the differential, but do not print it out here. Two neurologist develop a final differential diagnoses list from highest possibility to lowest possibility. Finally, provide a python list of the differential diagnoses in order of the priority in the format: differential_diagnoses = []	Generate Differential Diagnoses	CoT and ToT

FIGURE 1
The system prompts for Model N (neurologist), Model P (patient), and Model S (Supervisor).

exploration of various orderings, culminating in the prioritized differential diagnosis. The internal process of clinical reasoning was not utilized (10, 11).

Analysis

To evaluate history taking, we assessed: (1) the overall retrieval accuracy of GPT-4 from the original HPI, (2) the retrieval rates of individual HPI components, and (3) the ranking percentile of the case diagnosis in the differential diagnosis. In order to accurately assess Model N, we eliminated any trials that contained errors by Model P or Model S that could potentially affect Model N's history-taking capability directly or indirectly.

Retrieval accuracy was determined by comparing the HPI generated by Model S to the original HPI, using an evaluation rubric based on a systematic scoring rubric for OSCE (12). We also analyzed the tool's performance in identifying chief complaints, associated symptoms, details of these symptoms (onset, character, duration, etc.), and known medical history (past medical history, surgical history, allergies, social history, and family history). The ranking accuracy of the correct diagnosis in the differential diagnosis list and the average number of interactions required were also measured to observe the relevancy and efficiency of history taking. The ranking accuracy was calculated as follows: if N number of differential diagnoses generated by Model S, and the case diagnosis is ranked X, ranking accuracy = (N-X)/N*100. The number of differential diagnoses was counted, and ranking accuracy was calculated for each trial, which was later averaged by disease category. If the diagnosis from the case report ranks high on the differential diagnosis list generated by GPT-4, this would suggest that GPT-4 could accurately identify the characteristic features of the case report's diagnosis. On the other hand, if the diagnosis appears low on the list, it could indicate that GPT-4 failed to recognize essential aspects typical of the case report's diagnosis. Consistency across paired trials was measured by the mean Jaccard index.

Patient and public involvement

Patients and the public were not involved in the design, conduct, reporting, or dissemination plans of this research study. Dissemination to Participants and Related Patient and Public Communities: The results of this study have not been disseminated to research participants as no participants were involved.

Results

Bibliographic search for case reports

Initial *PubMed Central* search in March 2024 identified multiple candidate articles: headache: 102, stroke: 283, neurodegenerative disease: 86, of which only 6, 24, and 5, respectively, contained substantial HPI information (Supplementary Figure 1). Five of each category were randomly selected: headache cases were migraine, tension headache, cluster headache, post-traumatic headache, and intracranial hypotension; stroke cases included two cases of posterior circulation and three of anterior circulation stroke; neurodegenerative cases were Alzheimer's, Parkinson's, Lewy body dementia, frontotemporal dementia, and Creutzfeldt-Jakob disease. Because of drawing from the literature, the cases were biased toward the "zebras," i.e., rare diseases, not a major drawback since initial history will be geared toward the chief complaint regardless of the underlying cause.

Technical limitations

The full patient-doctor-supervisor script for each case was independently generated five times. Out of 75 trials of the patient-doctor simulation scripts, 7 trials were excluded from further analysis due to GPT-4 errors: 7 trials with omissions with Model P (patient) responses deviating from the case report.

Example Model N<-> Model P dialog

Original HPI from a 'tension headache' case (ID: PMC10617078) and a sample script of patient-doctor simulation based on it are demonstrated along with generated HPI and top 3 differential diagnoses by Model S.

Comments on the script

We note again that these 3 models generated text independently—the only interaction between N and S was through the words seen here (Table 1).

Model N followed a logical history-taking sequence, asking about symptoms to distinguish the different possible causes of headaches. We did not prompt any specific questions appropriate for headache or any other disorder type, prompting only for general symptom queries—onset, character, progression. Model N then used its underlying language database to identify specific headache-related questions such as photophobia (this is a generic GPT-4 model with no additional training or prompting on neurological disease). Model N similarly followed a reasonable sequence in the other case types.

In most places, Model P expressed itself in the first-person: "I have a history of bronchial asthma ...," but sometimes changed to a third person, simply quoting the original case: "He has a habit of drinking ..." This reflected the prompt option of replying as a family member or EMT. Despite being instructed to act like a patient without medical jargon, Model P seemed excessively sophisticated at times: "bronchial asthma" instead of "asthma"; "allergic rhinitis." All of the cases necessarily featured a large number of "Negative" responses, reflecting the limited amount of information available in published case reports.

Model S generates an HPI that is quite similar to the input despite working with a highly filtered version of that history that only included a few sentences that were copied directly from the original. The generated differential diagnosis was largely reasonable. Additional diagnoses might have been considered higher in the differential—(1) Depression: Model P twice mentions stress, depression, and parent death, so depression is likely to be a major factor here; (2) cervicalgia or fibromyalgia is suggested by the associated shoulder pain, but this could all be due to depression, as suggested by the patient himself.

Statistical analysis across trials

Model N (neurologist) achieved an overall HPI content retrieval accuracy of 81% from the original HPI, with specific accuracies of 84% for headache, 82% for stroke, and 77% for neurodegenerative disease (Supplementary Table 1). The overall average patient-doctor number of interactions was 37.6 (41, 42, 33.3, respectively) Consistency (average Jaccard index) was 0.86 overall (0.80, 0.88, 0.89). HPI retrieval accuracy was: 93% for chief complaints, 47% for associated symptoms and review of systems, 76% for relevant symptom details, and 94% for histories for past medical, surgical, allergies, social, and family factors. Retrieval of the chief complaint was reduced because Model P was instructed to only provide no more than two symptoms at a time, but the HPI sometimes included three or more symptoms for the chief complaint. The lack of retrieval for associated symptoms and review of systems was largely attributed to the absence of these details in the original case and sometimes due to early termination when Model N had already reached its provisional diagnosis, concluding further query is unnecessary, which was indicated within Model N's prompt. The pattern of accuracy was consistent across disorders (Figure 2). Differential diagnosis ranking percentile of the case diagnosis averaged in the 89th percentile overall, with specific rankings percentiles of 84th for headache, 92nd for stroke, and 90th for neurodegenerative disease.

Overall appropriateness of Model N questions and Model P responses

Model N was able to provide a targeted history with some outlier "zebra-targeting" queries, but there were a few questions that could not be tracked to a differential diagnosis candidate (Supplementary Table 2). In headache cases, Model N inquired about visual and auditory disturbances, coordination issues, nausea, vomiting, vertigo, weakness, numbness, paresthesia, speech or swallowing impairments, changes in olfactory or gustatory senses, infectious history, mood alterations, sleep patterns, and appetite. For stroke, Model N probed for headaches and focal deficits including vision changes, weakness, numbness, difficulty

TABLE 1 Original HPI from a 'tension headache' case (ID: PMC10617078) and a sample script of patient-doctor simulation based on it are demonstrated along with generated HPI and top 3 differential diagnoses by Model S.

Case Report HPI (verbatim)	A patient in 40s reported headache for 20 years prior to his visit to our hospital. His headache symptoms came in waves, but the symptoms were essentially continuous. He had a habit of drinking 1 L of beer three times a week, and he had smoked 15 cigarettes per day since the age of 20, but quit smoking at 40 years old. He had a history of bronchial asthma and allergic rhinitis. He had not used headache medication frequently, there were no particular problems with his psychosocial history, but his family history included a grandfather and an uncle who also suffered from bronchial asthma. His occupation consisted mainly of desk work. During and after the coronavirus disease 2019 (COVID-19) pandemic, he worked remotely, and at the time of his visit to our hospital, he was engaged in work on a personal computer most days. From around March of the year that he visited our hospital, he went to a local clinic because his headache had worsened. Around this time, his symptoms of neck and shoulder heaviness worsened, likely due to stress from his child's upcoming entrance examination and the death of his parents. In addition, he rarely went out since beginning working at home due to the COVID-19 pandemic. He felt a squeezing pain all over his head and heaviness in both of his shoulders. In addition, he had severe headache symptoms that kept him in bed in the morning several times a month. He began to fall asleep when he went to events such as graduations, entrance ceremonies, and funeral services.
Patient (P) – Doctor (N) Simulation	Model N: What is your age and gender? What brings you here today? Model P: I'm here because my headaches have gotten worse, and I've been feeling a squeezing pain in my head and heaviness in my shoulders. Model N: Can you tell me more about your headaches, such as when they started and how they have progressed over time? Model P: I've been having headaches for 20 years. They come and go, but I pretty much always feel some level of headache. Recently, they have gotten worse, especially from around March of this year. Model N: Can you describe the character of your headaches? Are they throbbing, stabbing, or constant? And where exactly is the pain located? Model P: The headaches feel like a squeezing pain all over my head. Model N: Can you tell me if there are any specific triggers that seem to bring on these headaches or make them worse, such as certain activities, foods, or stress? Model P: The headaches got worse around March of this year, which seemed to be related to stress from my child's upcoming entrance examination and the death of my parents. Model N: Have you noticed any other symptoms associated with your headaches, such as nausea, vomiting, sensitivity to light or sound, or changes in your vision? Model P: Negative. Model N: Do you experience any other symptoms not related to your headaches, such as weakness, numbness, or difficulties with speech, vision, or coordination? Model P: Negative. Model N: Can you tell me about any past medical history you have, including chronic illnesses, previous significant illnesses, or hospitalizations? Model P: I have a history of bronchial asthma and allergic rhinitis. Model N: Have you had any surgeries in the past? If so, can you please describe them? Model P: Negative. Model N: What medications are you currently taking, including over-the-counter drugs and supplements? Model P: Negative. Model N: Can you tell me about your social history, including your occupation, smoking and alcohol use, and recreational drug use? Model P: He has a habit of drinking 1 liter of beer three times a week and smoked 15 cigarettes per day from the age of 20 but quit at 40. There's no mention of recreational drug use. Model N: Can you tell me more about your alcohol consumption? Have there been any recent changes in the amount or frequency of your drinking? Model P: Negative. Model N: Do you have any known allergies to medications, foods, or other substances? Model P: Yes, I have allergies to pollen which cause allergic rhinitis. Model N: Have any of your family members had similar headaches or been diagnosed with neurological conditions? Model P: My grandfather and my uncle also suffered from bronchial asthma. Model N: Can you tell me more about the heaviness in your shoulders that you mentioned earlier? When did it start and how would you describe it? Model P: The heaviness in my shoulders started around March of the year I visited the hospital. It got worse, likely because of the stress from my child's upcoming entrance examination and the death of my parents. It feels like a heavy weight on both of my shoulders. Model N: Do you have any difficulty with movement or sensation in your arms or legs? Model P: Negative. Model N: Do you experience any neck pain or stiffness, or does the pain from your shoulders extend to any other part of your body? Model P: The heaviness in my shoulders started around March of the year I visited the hospital. It got worse, probably because of stress from my child's upcoming entrance exam and the death of my parents. It feels like a heavy weight on both of my shoulders. Model N: Have you noticed any changes in your sleep patterns or the quality of your sleep since your headaches and shoulder heaviness have worsened? Model P: Negative. Model N: Can you tell me if you have tried any treatments or taken any measures to relieve your headaches and shoulder heaviness? If so, what were they and did they provide any relief? Model P: Negative. Model N: Have you experienced any changes in your mood, memory, or concentration since the onset of your headaches and shoulder heaviness? Model P: Negative. Model N: Have you experienced any recent changes in your weight, appetite, or overall energy levels? Model P: Negative. Model N: terminate Session terminated by the neurologist. Total number of interactions: 38

(Continued)

TABLE 1 (Continued)

HPI Summary by Model S	[HPI: A patient in 40s presents with a chief complaint of worsening headaches, described as a squeezing pain all over his head, which he has experienced for 20 years. These headaches have been present consistently but have recently intensified from around March of this year, correlating with periods of increased stress due to his child's entrance examination and the recent death of his parents. The patient also reports a sensation of heaviness on both shoulders starting around the same time in March. He denies any nausea, vomiting, sensitivity to light or sound, changes in vision, weakness, numbness, difficulties with speech, vision, or coordination, neck pain or stiffness, and pain extending to other body parts. There has been no change in sleep patterns, mood, memory, concentration, weight, appetite or overall energy levels. The patient has not tried any treatments or measures to relieve his headaches and shoulder heaviness. No exacerbating factors such as specific activities, foods, or other stressors were identified, and there are no alleviating factors reported.] [… correct *summary of past medical, social, family history followed here*]
Top 3 Differential Diagnosis by Model S	1. Tension-type headache 2. Chronic daily headache 3. Migraine without aura

Patient identifiers (age, sex) has been anonymized (24).

in speaking with further queries including changes in memory, attention or cognition, mood, sleep patterns, energy levels, muscle weakness, wasting, or twitching, recent trauma, illness, or stress. For neurodegenerative diseases, Model N sought information on changes in speech, swallowing, balance, and sensory disturbances, facial expression, walking, alterations in smell or taste, sleep disturbances, cognitive, mood, or behavioral changes, bowel or bladder issues, memory, or concentration difficulties, fine motor skills, syncopal/presyncope symptoms, traumatic injuries with additional clarifications (e.g., verifying whether the patient was passenger or driver for a motor vehicle accident) and chemical exposure.

Model P produced 7 omissions and no confabulations out of 1,411 total responses: omitting hypertension history in 2 stroke trials of one case, omitting visual disturbances in 4 trials of one case, and omitting medication for 1 trial of one case. These Model P errors led to further information retrieval errors (e.g., when Model P had omitted information such as visual symptoms or medical history, these led to falsely decreased retrieval rates) and, therefore, these trials were eliminated. No omissions or confabulations were observed in Model S. Model N limitations were seen where it failed to seek further history clarification to establish symptom characteristics or timelines, leading to repetitive questioning (e.g., trying to clarify trouble speaking, whether it is dysarthria or aphasia; figuring the timeline between symptoms where the patient had trouble hearing then developed unsteady gait) seeking clarification until session termination. Confabulations were not observed from any of the 3 models.

Discussion

Model N (neurologist) was able to capture historical information relevant to each type of neurological condition tested by making appropriate queries to the case report's *digital twin* (model P—the patient), which reframed the case appropriately. The system achieved an overall retrieval accuracy of 80% from the original HPI across 12 neurological diseases drawn from three disease categories (stroke, headache, and neurodegenerative disease). Model N took a structured history with high accuracy in identifying the chief complaint, demonstrating the ability to capture the essential patient information for clinical assessment. Lower accuracy was found in obtaining associated symptoms and review-of-systems along with symptom details. The overall ranking of the case diagnosis was in the 89th percentile, suggesting that GPT-4 could accurately identify the key clinical findings of the case report's diagnosis and demonstrate potential value as supporting guidance for history taking.

The classical general practitioner continues to exist in various guises—family practice, nurse practitioners, physician assistants—particularly in rural areas in the US and other countries. In addition to neurology, other specialties that have a broad remit that requires considering a large variety of diseases and disorders include psychiatry, internal medicine, and pediatrics. All of these generalists must take a history that takes into account not only the variety of clinical problems that they were trained in but also the many changes in diagnosis and disease classification that have occurred since their training was complete. Additionally, increased subspecialization narrows a clinician's perspective, leading to histories that miss symptoms outside of this subspecialty. We propose the AI generalist as a support tool (13, 14), offering a broad, largely unbiased approach to initial encounters, and reducing the chance of diagnostic oversight due either to limited knowledge, or due to tunnel vision with too much focus (15, 16).

Other AI-based history-taking tools have focused on general medicine or have used structured pre-consultation questionnaires, which lacked the flexibility available with the LLM conversational approach. The conversational agents discussed by Tudor Car et al. (6) did offer this flexibility for general health inquiries, or for general practice (8). By focusing on neurology, our study shows the potential to go further into detail with precise differential diagnosis synthesis in a subspecialty context.

An important aspect of our study is the role of Model P, the simulated patient, as a *digital twin* of the patient described in the published case report. In the current study, this twin has been instructed not to stray from the narrow path of what was described and to answer in the negative for any query that exceeds those bounds. However, a future digital twin historian could be augmented through connectivity to a large variety of automated gathering systems: actinographs would be useful in PD, epilepsy, post-stroke rehabilitation, and ALS; pulse and blood pressure monitoring in patients with atrial fibrillation; hypertension, stroke or stroke risk; glucose monitoring in stroke or peripheral neuropathy. Many other examples could be given, expanding as wearable and implantable sensors become more sophisticated. Additionally, patients would be encouraged to enter diary information verbally on their phone or watch as they are now asked to do in written diaries: headache diaries, seizure diaries, fall diaries, etc. An advantage of this consolidated LLM

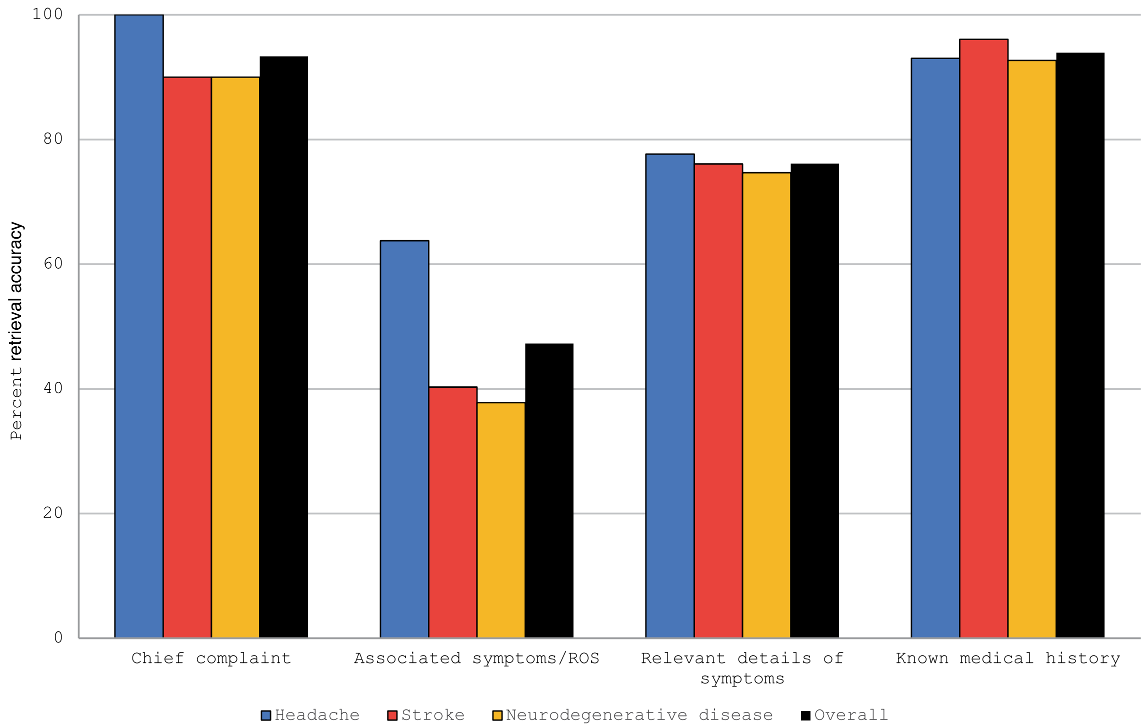


FIGURE 2
Retrieval accuracy (%) for HPI components for neurological disorder types. ROS, Review of systems; Known medical history includes past medical history, past surgical history, allergies, social history, and family history.

digital twin would be that it will consolidate all of these measures as well as any additional patient notes in a readily queryable system. Further combination of Models N, P, and S in a single system group would then provide a digital twin that identifies relevant clinical correlations and takes us all the way from basic data to diagnosis.

Limitations of our study and of AI

In contrast to a previous study of AI history-taking (8), we focused on a confined usage targeting pre-encounter questioning as an alternative to questionnaires, seeking to augment rather than replace the clinician. We propose the supportive role of conversational LLMs in medical history taking as another potential diagnostic tool alongside laboratory tests and other technical modalities.

Our study was based on case reports. Case reports are usually of an atypical presentation of a disease and lack comprehensiveness. We, therefore, would expect better AI performance from EMR HPI or in direct interaction with a patient. However, direct patient interviewing will add additional confusion and may also include symptom magnification due to the patient's understandable focus on his or her problems. These are areas where clinician judgment is important in deciphering human psychology (17–19).

We propose the use of AI as an *adjunct* to history-taking which parallels current practice of using medical students, residents, or physician assistants to take initial histories. This is then used as a starting point by the physician of record who then will repeat some of the same questions and re-evaluate the patient's responses. Understanding human behavior in general and of the person who is acutely ill, chronically ill, or in pain is indispensable and is one reason why years of patient exposure during training are required to obtain advanced clinical skills. We emphasize the need to use AI as a support tool rather than as a freestanding diagnostician.

Another limitation of this study is that it was text-based and did not consider presentation diversity, including dialects and speech impairments and cultural variability of response to pain and to neurological impairment, all of which cause further difficulties in history representation and in history-taking. Testing of the models in realistic clinical settings is essential to address these challenges. Further enhancement using voice recognition and training rather than transcripts would provide still greater indication of flexibility and future utility.

This study did not assess human or clinician acceptance of the proposed AI-based tools, as the pilot was conducted outside a clinical setting. AI tools are gradually being integrated into clinical practice through automated reading and interpretation of radiographs and other test results. We anticipate that our history-augmentation and history-identification tools will find enhanced adoption in clinical settings where direct neurologist advice is not available. Usability studies will be needed to improve and then confirm acceptance, systematically collecting and analyzing feedback from healthcare providers on usability, trust, and workflow integration. Additionally, education and training initiatives designed to familiarize clinicians with these tools could reduce resistance and facilitate adoption.

There are multiple other limitations inherent in human-machine interactions. Despite, and in some cases because of, the existence of various robotic interfaces meant to provide a more human look, the patient will not relate to a machine in the way that they relate to a person—many may refuse to deal with it entirely. Even if the LLM itself is largely unbiased, the prompt introduces additional bias. For example, our prompt introduced a bias of focus on the chief complaint; in some histories, the chief complaint is misleading, and the major medical problem only arises with further queries.

Additional risks for utilizing conversational LLMs in healthcare include human rights—discrimination, stereotyping, and exclusion; data-related risks—privacy, data governance, and stigma; and technical risks—error tolerance, excessive reliance on chatbot advice, and reduced trust in health professionals (20–22). These issues underscore the need for a judicious and selective integration of conversational LLMs in the healthcare setting (23).

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

J-HL: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. EC: Writing – original draft, Writing – review & editing. SA: Writing – original draft, Writing – review & editing. RM: Formal analysis, Writing – original draft, Writing – review & editing. WL: Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

WL would like to thank Gunnar Cedersund and the STRATIF-AI Digital Twin research team for helpful discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2024.1496866/full#supplementary-material

References

1. Nicholl DJ, Appleton JP. Clinical neurology: why this still matters in the 21st century. *J Neurol Neurosurg Psychiatry*. (2015) 86:229–33. doi: 10.1136/jnnp-2013-306881

2. Saldaña-Inda I, Cisneros-Gimeno AI, Lambea-Gil A. Neurophobia among resident physicians in the emergency service. *Rev Neurol*. (2023) 77:285–91. doi: 10.33588/rn.7712.2023249

3. Albrink K, Joos C, Schröder D, Müller F, Hummers E, Noack EM. Obtaining patients' medical history using a digital device prior to consultation in primary care: study protocol for a usability and validity study. *BMC Med Inform Decis Mak*. (2022) 22:189. doi: 10.1186/s12911-022-01928-0

4. Berdahl CT, Henreid AJ, Pevnick JM, Zheng K, Nuckols TK. Digital tools designed to obtain the history of present illness from patients: scoping review. *J Med Internet Res*. (2022) 24:e36074. doi: 10.2196/36074

5. Shucard H, Muller E, Johnson J, Walker J, Elmore JG, Payne TH, et al. Clinical use of an electronic pre-visit questionnaire soliciting patient visit goals and interim history: a retrospective comparison between safety-net and non-safety-net clinics. *Health Serv Res Manag Epidemiol*. (2022) 9:23333928221080336. doi: 10.1177/23333928221080336

6. Tudor Car L, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng Y-L, et al. Conversational agents in health care: scoping review and conceptual analysis. *J Med Internet Res*. (2020) 22:e17158. doi: 10.2196/17158

7. Fournier-Tombs E, McHardy J. A medical ethics framework for conversational artificial intelligence. *J Med Internet Res*. (2023) 25:e43068. doi: 10.2196/43068

8. Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic AI. arXiv [Preprint]. (2024). Available at: http://arxiv.org/abs/2401.05654

9. Lee J-H, Choi E, McDougal R, Lytton WW. GPT-4 performance for neurologic localization. *Neurol Clin Pract*. (2024) 14:e200293. doi: 10.1212/CPJ.0000000000200293

10. Chiang C-C, Fries JA. Exploring the potential of large language models in neurology, using neurologic localization as an example. *Neurol Clin Pract*. (2024) 14:e200311. doi: 10.1212/CPJ.0000000000200311

11. Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, et al. Tree of thoughts: deliberate problem solving with large language models. arXiv [csCL]. (2023). Available at: http://arxiv.org/abs/2305.10601

12. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part I: an historical and theoretical perspective. *Med Teach*. (2013) 35:e1437–46. doi: 10.3109/0142159X.2013.818634

13. Hashem A, Chi MTH, Friedman CP. Medical errors as a result of specialization. *J Biomed Inform*. (2003) 36:61–9. doi: 10.1016/S1532-0464(03)00057-1

14. Sah S, Fagerlin A, Ubel P. Effect of physician disclosure of specialty bias on patient trust and treatment choice. *Proc Natl Acad Sci USA*. (2016) 113:7465–9. doi: 10.1073/pnas.1604908113

15. Triantafyllou SA. A detailed study on implementing new approaches in the game of life. *Data Metadata*. (2023) 2:95. doi: 10.56294/dm202395

16. Triantafyllou SA. A detailed study on the game of life. In: Yousef F, editor. Lecture notes in networks and systems, Lecture notes in networks and systems. Cham: Springer Nature Switzerland (2024). 32–8.

17. Nadelson T. The Munchausen spectrum: borderline character features. *Gen Hosp Psychiatry*. (1979) 1:11–7. doi: 10.1016/0163-8343(79)90073-2

18. Sandson J, Albert ML, Alexander MP. Confabulation in aphasia. *Cortex*. (1986) 22:621–6. doi: 10.1016/S0010-9452(86)80021-1

19. Ardila A. A proposed reinterpretation and reclassification of aphasic syndromes. *Aphasiology*. (2010) 24:363–94. doi: 10.1080/02687030802553704

20. Hamdoun S, Monteleone R, Bookman T, Michael K. AI-based and digital mental health apps: balancing need and risk. *IEEE Technol Soc Mag*. (2023) 42:25–36. doi: 10.1109/MTS.2023.3241309

21. Miner AS, Laranjo L, Kocaballi AB. Chatbots in the fight against the COVID-19 pandemic. *NPJ Digit Med*. (2020) 3:65. doi: 10.1038/s41746-020-0280-0

22. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. (2023) 6:120. doi: 10.1038/s41746-023-00873-0

23. Cabral S, Restrepo D, Kanjee Z, Wilson P, Crowe B, Abdulnour R-E, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Intern Med*. (2024) 184:581–3. doi: 10.1001/jamainternmed.2024.0295

24. Takekawa T, Chino T, Yamada N, Watanabe S, Abo M, Sengoku R. Multimodal treatment, including extracorporeal shock wave therapy, for refractory chronic tension-type headache: a case report. *J Med Case Rep*. (2023) 17:478. doi: 10.1186/s13256-023-04092-9

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 22 January 2025
DOI 10.3389/fmed.2025.1529305

OPEN ACCESS

EDITED BY
Thomas F. Heston,
University of Washington, United States

REVIEWED BY
Ghulam Farid,
Shalamar Medical and Dental College, Pakistan
Abhra Ghosh,
Mata Gujari Memorial Medical College, India

*CORRESPONDENCE
Juan José Gualda-Gea
✉ jgualdge17@alumnes.ub.edu

†These authors share senior authorship

RECEIVED 19 November 2024
ACCEPTED 07 January 2025
PUBLISHED 22 January 2025

CITATION
Gualda-Gea JJ, Barón-Miras LE, Bertran MJ, Vilella A, Torá-Rocamora I and Prat A (2025) Perceptions and future perspectives of medical students on the use of artificial intelligence based chatbots: an exploratory analysis.
Front. Med. 12:1529305.
doi: 10.3389/fmed.2025.1529305

Perceptions and future perspectives of medical students on the use of artificial intelligence based chatbots: an exploratory analysis

Juan José Gualda-Gea[1,2,3]*, Lourdes Estefanía Barón-Miras[1,2,3], Maria Jesús Bertran[1,2], Anna Vilella[1,2,3], Isabel Torá-Rocamora[1,2,3†] and Andres Prat[1,2,3†]

[1]Department of Preventive Medicine and Epidemiology, Hospital Clínic of Barcelona, Barcelona, Spain, [2]Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain, [3]Department of Medicine, Faculty of Medicine, University of Barcelona, Barcelona, Spain

Background: Artificial Intelligence (AI) has made a strong entrance into different fields such as healthcare, but currently, medical degree curricula are not adapted to the changes that adopting these types of tools entitles. It is important to understand the future needs of students to provide the most comprehensive education possible.

Objective: The aim of this teaching improvement project is to describe the knowledge, attitudes, and perspectives of medical students regarding the application of AI and chatbots with patients, also considering their ethical perceptions.

Methods: Descriptive cross-sectional analysis in which the participants were students enrolled in the subject "Preventive Medicine, Public Health and Applied Statistics" during the second semester of the 2023/24 academic year, corresponding to the fifth year of the Degree in Medicine at the University of Barcelona. The students were invited to complete a specific questionnaire anonymously and voluntarily, which they could respond to using their mobile devices by scanning a QR code projected on the classroom screen, we used Microsoft Forms to perform the survey.

Results: Out of the 61 students enrolled in the subject, 34 (56%) attended the seminar, of whom 29 (85%) completed the questionnaire correctly. Of those completing the questionnaire, 20 (69%) had never used chatbots for medical information, 19 (66%) expressed a strong interest in the practical applications of AI in medicine, 14 (48%) indicated elevated concern about the ethical aspects, 17 (59%) acknowledged potential biases in these tools, and 17 (59%) expressed at least moderate confidence in chatbot-provided information. Notably, 24 (83%) agreed that acquiring AI-related knowledge will be essential to effectively perform their future professional roles.

Conclusion: Surveyed medical students demonstrated limited exposure to AI-based tools and showed a mid-level of awareness about ethical concerns, but they recognized the importance of AI knowledge for their careers, emphasizing the need for AI integration in medical education.

KEYWORDS

medical education, artificial intelligence, chatbots, medical students, teaching improvement

1 Introduction

Digital transformation in the healthcare sector is driving a deep reconfiguration of medical practice, with Artificial Intelligence (AI) emerging as a key factor in addressing current and future healthcare challenges (1). AI-based tools, such as machine learning algorithms and large-scale data analysis, have already demonstrated their capacity to improve diagnostic accuracy and accelerate the early identification of diseases, resulting in more timely interventions and more favorable patient outcomes (2, 3). Additionally, the increasing digitization of information and the incorporation of decision support systems optimize workflows, reduce administrative burdens, and facilitate access to care, even in resource-limited settings (4).

Within this technological ecosystem, chatbots AI-driven conversational assistants have positioned themselves as promising tools to enhance interaction between healthcare professionals and patients (5, 6). These systems can provide immediate responses to basic inquiries, offer reliable information on symptoms and treatments, and promote health education, thereby expanding access to healthcare services beyond geographical and temporal limitations (7, 8). However, the implementation of these technologies is not without challenges, particularly concerning ethical issues and the quality of information provided (9).

AI in healthcare presents ethical dilemmas that encompass information integrity, data privacy, and accountability in algorithm-mediated clinical scenarios (10–12). Furthermore, the potential for biases, informational "hallucinations" (responses that appear valid but are unfounded), and the possible erosion of the doctor-patient relationship underscore the need to address these technologies with prudence and rigor (13–15). In this regard, medical education plays a central role: preparing future healthcare professionals to understand, adopt, and critically evaluate AI tools is essential to ensure their ethical, effective, and patient-centered integration into clinical practice (16, 17).

Although various studies have explored the general perceptions of students and healthcare professionals regarding AI, there remains a gap in the literature concerning the specific understanding that advanced medical students have about the use of chatbots in clinical settings (18). This population is at a critical juncture: on the brink of entering professional practice, their perceptions, concerns, and expectations provide valuable insights into how curricula and training strategies should be shaped to meet the demands of an imminent future marked by the gradual inclusion of AI in healthcare delivery (11, 14, 18, 19). Understanding their attitudes, knowledge levels, and ethical concerns offers a solid foundation for designing curricula that balance technical training with ethical reflection, promoting responsible and informed use of AI.

In this context, the present teaching improvement project aims to describe the knowledge, attitudes, and perspectives of medical students regarding the application of AI and the use of chatbots in the healthcare field, with particular attention to their ethical perceptions. This approach seeks to generate an initial framework to guide the future inclusion of AI-related content in medical education, ensuring that tomorrow's physicians are better prepared to integrate these tools into their clinical practice competently and ethically.

2 Materials and methods

2.1 Study design

A descriptive cross-sectional study was conducted with the aim of obtaining an initial understanding of medical students' perceptions and attitudes regarding the integration of AI-based chatbots in the healthcare sector.

2.2 Population and sampling

The target population comprised students enrolled in the course "Preventive Medicine, Public Health, and Applied Statistics," corresponding to the fifth year of the Medicine Degree at the University of Barcelona, during the second semester of the 2023/24 academic year. A non-probabilistic sampling method was employed, selecting participants who attended a theoretical seminar on the use of chatbots in the medical field and who voluntarily agreed to complete the questionnaire.

2.3 Sample size

The sample size was determined by seminar attendance and voluntary participation in the survey. Given the exploratory and preliminary nature of the study, an ideal sample size was not calculated using specific statistical formulas. The sample included students who, prior to the seminar, scanned a QR code and completed the online questionnaire using the Microsoft Forms application.

2.4 Instrument development

A quantitative questionnaire was designed, structured into three main sections with a total of 14 items. The questionnaire was intended to be simple, providing an initial approximation of students' perceptions. Each question featured a closed-response format (predefined options or Likert scales ranging from 1 to 5). The three dimensions investigated were:

- Attitudes and Prior Knowledge (3 items): Assesses previous familiarity with AI tools and chatbots in medicine.
- Ethical Perceptions (3 items): Explores ethical concerns and the level of trust in information provided by chatbots.
- Future Perspectives (8 items): Investigates the future relevance of AI knowledge for medical practice and professional training.

Q1. Have you ever used chatbots to obtain medical information? 1-Never, 5-Frequently.

Q2. Are you familiar with current artificial intelligence tools applied to medicine, such as AI-assisted diagnosis or therapeutic recommendations? 1 - not at all, 5 - very much.

Q11. I am interested in the practical aspects of AI in medicine. 1-very little, 5-a lot.

Q8. Are you concerned about the ethics of using chatbots in medicine? 1 "not very concerned" and 5 "very concerned."

Q9. On a scale from 1 to 5, where 1 is "not very confident" and 5 is "very confident," how much trust do you have in the information provided by chatbots on medical topics?

Q10. Are you concerned about potential biases in such tools? 1-very little, 5-a lot.

Q3. The use of AI in healthcare can positively change medicine. 1-Strongly disagree, 5-Strongly agree.

Q4. The use of AI can negatively affect the doctor-patient relationship. 1-Strongly disagree, 5-Strongly agree.

Q5. Doctors will need to know about AI-based tools to perform their jobs in the near future. 1-Strongly disagree, 5-Strongly agree.

Q6. AI should be part of medical education. 1-Strongly disagree, 5-Strongly agree.

Q7. Practical content on the use of AI-based tools in medicine should be introduced in medical degree programs. 1-Strongly disagree, 5-Strongly agree.

Q12. The use of such tools will lead to a dehumanization of medicine. 1-Strongly disagree, 5-Strongly agree.

Q13. The use of such tools will create dependency among medical staff. 1-Strongly disagree, 5-Strongly agree.

Q14. The imposition of these new technologies may influence the choice of specialization for medical personnel. 1-Strongly disagree, 5-Strongly agree.

This instrument was applied in its second iteration, following a pilot test conducted with 14 students in a workshop. This pilot allowed for the adjustment and consensus of questions with expert faculty members to enhance clarity and relevance.

2.5 Reliability and validity

Given the preliminary and exploratory nature of the study, comprehensive psychometric analyses (e.g., formal internal reliability tests or construct validity assessments) were not performed. However, the questionnaire underwent review by expert faculty in the fields of preventive medicine and public health, as well as medical education, to ensure clarity, internal consistency, and item relevance. Future research is recommended to formally validate the instrument, including conducting more extensive pilot tests and appropriate psychometric analyses to strengthen the questionnaire's reliability and validity.

2.6 Data collection procedure

Prior to the commencement of the theoretical seminar, students were invited to complete the questionnaire anonymously and voluntarily. Participation involved scanning a QR code projected in the classroom and responding to the questionnaire on their personal mobile devices via Microsoft Forms. To prevent duplicate responses, a time limit was set for completing the questionnaire. No personal, health-related, or sensitive data were collected. Participants were informed about the confidentiality of their responses and their right to abstain from answering or to withdraw from the survey at any time without any consequences.

2.7 Data analysis plan

Data analysis was structured according to the three sections of the questionnaire: Attitudes and Prior Knowledge, Ethical Perceptions, and Future Perspectives. The following techniques were employed:

2.7.1 Descriptive statistics

Relative frequency calculations were utilized to characterize responses within each section of the questionnaire. This provided a quantitative overview of participants' knowledge and opinions on AI and chatbots prior to their exposure to the theoretical seminar.

2.7.2 Visual analysis using horizontal bar charts

Horizontal bar charts were employed to graphically represent the results, facilitating visual comparison of response distributions on a scale of 1 to 5. This type of visualization aids in quickly identifying trends and patterns within the collected data.

2.7.3 Integrated findings summary

Results from each section were synthesized to present a comprehensive conclusion, analysing medical students' perspectives on the integration of AI-based chatbots in healthcare. This approach prioritized quantitative aspects, allowing for a deeper exploration of participants' views beyond numerical data.

As an exploratory study, complex inferential methods or systematic evidence synthesis were not employed, limiting the analysis to basic quantitative description and the identification of general patterns in students' perceptions.

3 Results

A total of 61 students were enrolled in the course "Preventive Medicine, Public Health, and Applied Statistics" during the second semester of the 2023/24 academic year. Of these, 34 (56%) students attended the theoretical seminar on the use of chatbots in the medical field, and 29 (85%) of them fully completed the questionnaire. The results are organized according to the three dimensions outlined in the study's objective: initial knowledge and attitudes, ethical perceptions, and future perspectives on the integration of AI in clinical practice.

3.1 Attitudes and prior knowledge

This dimension aimed to describe the initial level of familiarity with AI tools and chatbots, as well as the interest in their application. The results indicated a low degree of prior exposure to these technologies:

Previous Use of Chatbots: 20 (69%) responses scored below 3 on a scale of 1 to 5, reflecting little to no experience in using chatbots to obtain medical information.

Knowledge of AI Tools in Clinical Settings: 23 (79%) responses also scored below 3, suggesting limited knowledge of specific AI applications in medicine.

Despite this lack of familiarity, a notable interest in the practical applications of AI in the medical field emerged, with 19 (66%) scores exceeding 3. This indicates a positive attitude towards acquiring knowledge and skills related to these tools (Figure 1).

3.2 Ethical perceptions

This section explored concerns regarding the reliability, biases, and ethical implications of using AI-based chatbots in healthcare settings. The results revealed a significant level of concern:

Ethics of Using Chatbots: 14 (48%) participants rated above 3, indicating concerns about the moral and deontological implications of integrating these tools into medical practice.

Potential Biases: 17 (59%) expressed concern (scores >3) about the existence of biases, suggesting that students are aware of the risk of partiality in the recommendations or information provided by AI tools.

Trust in Information Provided by Chatbots: Regarding the accuracy of the content supplied by chatbots, 17 (59%) scored ≥3, revealing moderate trust that is nevertheless tempered by the previously mentioned ethical doubts (Figure 2).

3.3 Future perspectives

The final section focused on opinions about the long-term impact of AI in medicine, including its effect on clinical practice, the training of future doctors, and the doctor-patient relationship. The findings suggest that students anticipate a significant change in their professional practice:

Positive Impact on Medicine: 100% of respondents rated ≥3, believing that AI can favourably transform medicine.

Educational Needs: 24 (83%) believe that doctors will require knowledge of AI to perform their duties effectively (≥3), and 26 (90%) consider that the medical curriculum should include AI (≥3), as well as practical content on its use (≥3).

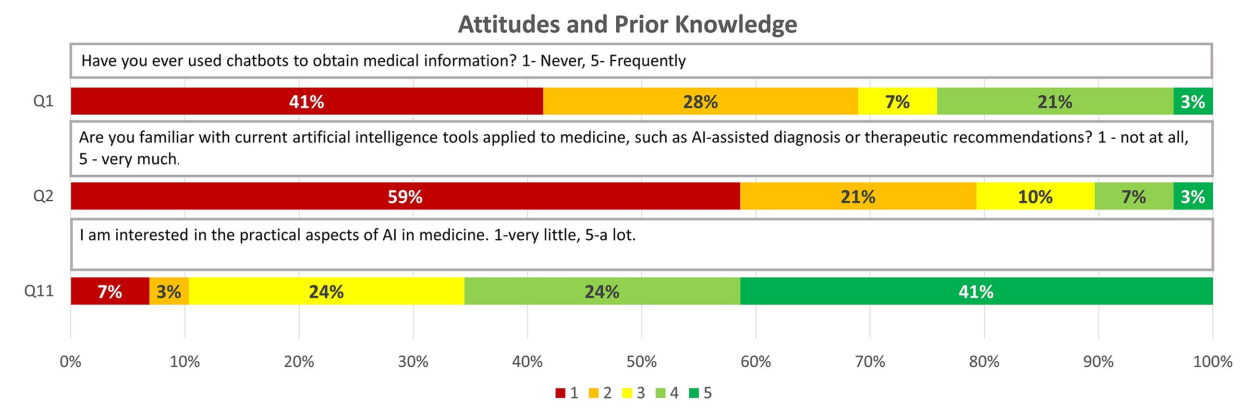


FIGURE 1
Graphic representation of the answers obtained for the questions on "Attitudes and Prior knowledge" through relative frequencies (1, 2, 11).

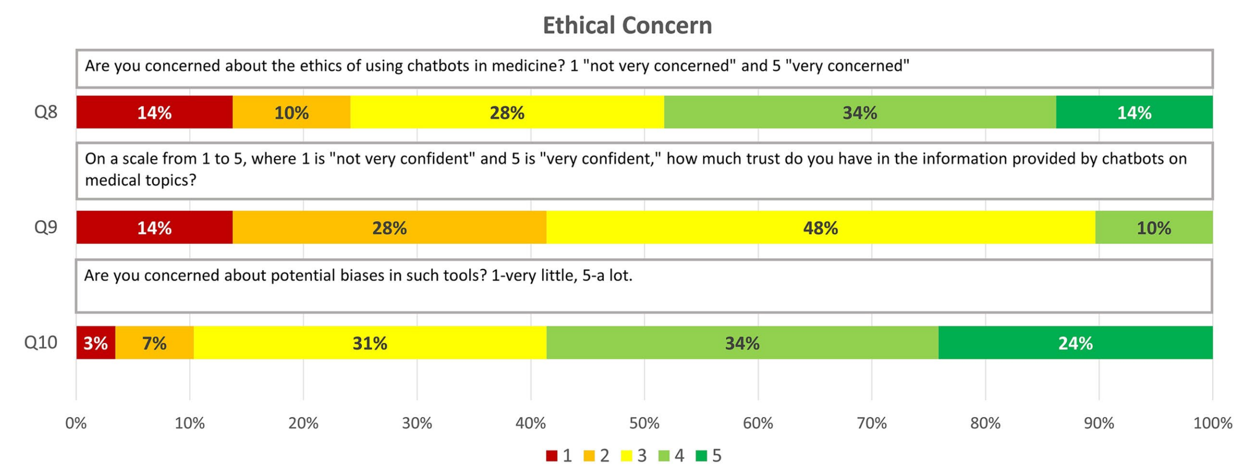


FIGURE 2
Graphic representation of the answers obtained for the questions on "Ethical concern" through relative frequencies (8–10).

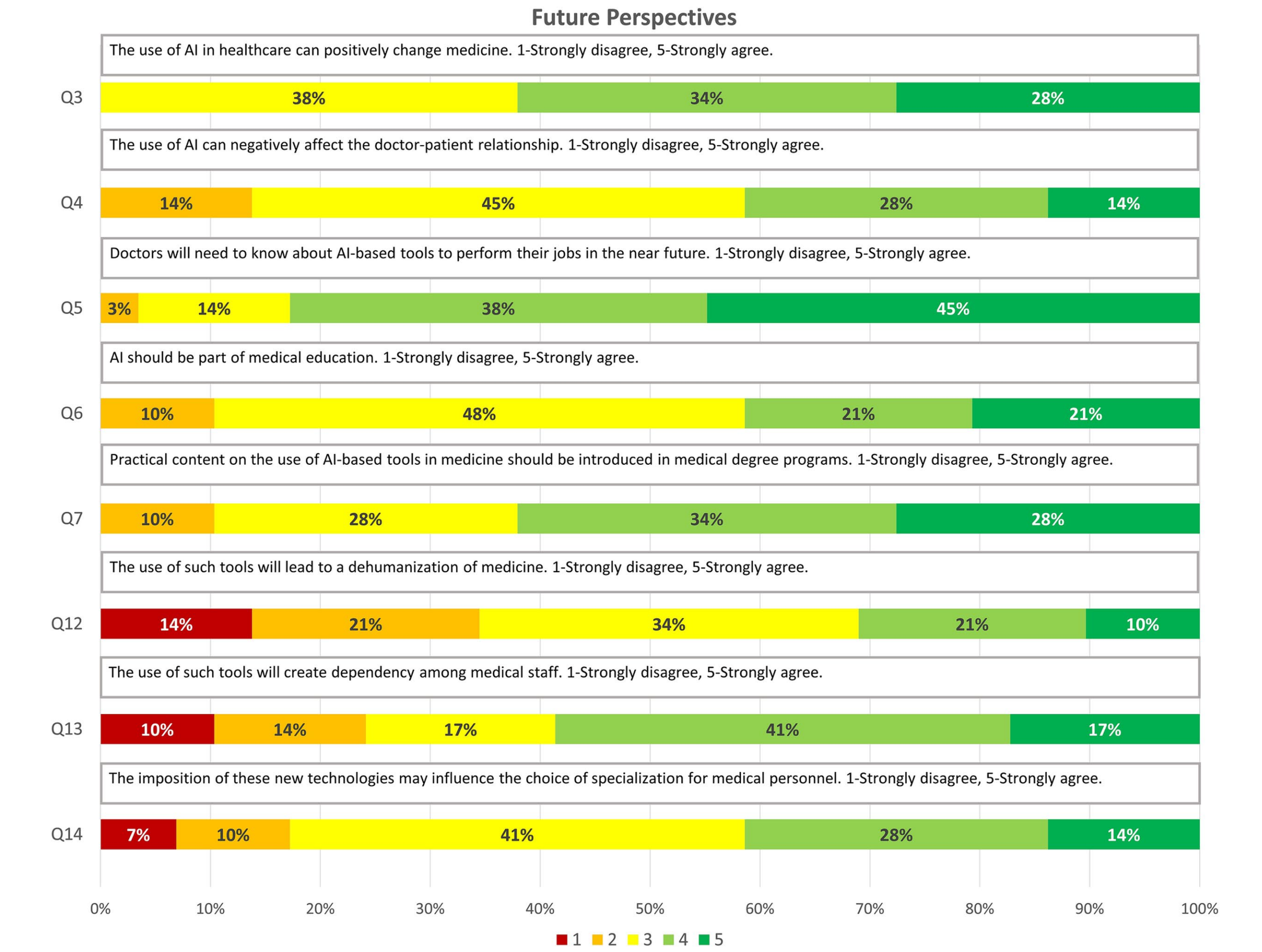


FIGURE 3
Graphic representation of the answers obtained for the questions on "Future perspectives" through relative frequencies (3–7, 12–14).

Concerns about the Doctor-Patient Relationship: 25 (86%) perceive that the use of AI could negatively affect this relationship (≥3), and 19 (65%) believe it could contribute to the dehumanization of healthcare (≥3). Additionally, 22 (76%) fear the development of dependence on these tools (≥3).

Influence on Specialty Choice: 24 (83%) consider that the imposition of new technologies, such as AI, could influence their future decisions regarding medical specialization (≥3).

Overall, these results demonstrate that while students have limited prior contact with AI tools, they show a growing interest in learning and integrating them. They recognize the potential of AI to transform medicine and medical education but remain cautious about the ethical and human implications of its implementation. These perceptions, aligned with the study's objective, provide an initial perspective on the educational needs, ethical concerns, and expectations of future healthcare professionals in the face of the increasing presence of AI in the health sector (Figure 3).

4 Discussion

The results obtained are consistent with the academic characteristics and formative stage of our sample of 29 medical students. These students, who have already completed Medical Ethics coursework and are concurrently engaging in clinical practices alongside theoretical subjects, represent an ideal profile for capturing how future healthcare professionals perceive the integration of Artificial Intelligence (AI) tools into their medical activities. Additionally, the non-mandatory nature of theoretical seminar attendance at this stage, combined with the documented absenteeism phenomenon in health sciences (20), reinforces the relevance of this sample as a study group.

This teaching improvement project aimed to explore the level of knowledge, ethical perceptions, and future perspectives of medical students regarding the use of AI tools in the healthcare field, specifically the employment of chatbots. Despite their limited direct experience with AI, the findings indicate that students are aware of the inherent ethical challenges of these technologies while recognizing the importance of acquiring competencies in this area for their future professional practice.

4.1 Attitudes and prior knowledge

The limited prior use of chatbots to obtain medical information aligns with trends described in the literature (21), indicating that these tools have not yet been widely incorporated into students' routine information-seeking practices. This lack of familiarity suggests the need for specific educational interventions that increase exposure to AI and enhance understanding of its applications (22). Nevertheless, the positive disposition towards learning these technologies reflects an open field for curricular development.

4.2 Ethical challenges

The identification of ethical concerns by the students constitutes one of the most significant findings of this study, highlighting an area that warrants deeper attention. Participants expressed concerns about the accuracy of information, the presence of biases, data confidentiality, and the moral implications of using chatbots in clinical practice. This sensitivity to ethical dilemmas aligns with literature that underscores the importance of addressing these issues in the integration of AI in healthcare (17, 23, 24).

Although students showed a certain degree of trust in the responses provided by chatbots, this trust is tempered by the previously mentioned ethical reservations. It is clear that the mere incorporation of AI tools is insufficient: it is imperative to establish solid ethical frameworks, well-defined guidelines, and training that goes beyond technical competencies. Including ethics modules focused on AI, case-based discussions, and dialogues with ethics and technology experts could foster a critical and responsible view of the use of these tools. In this way, future doctors can adopt balanced approaches, ensuring safe, equitable, and patient-centered applications.

4.3 Future perspectives

The students' perspectives suggest that AI could facilitate collaboration between healthcare professionals and chatbots, potentially optimizing care in an increasingly complex clinical environment (1). The nearly unanimous conviction that knowledge of these tools will be essential in their careers underscores the need to reform medical curricula, incorporating technological skills that prepare future professionals for a rapidly transforming care scenario (7, 25).

Furthermore, concerns about the risk of dehumanizing care, potential technological dependence, or the influence of AI on specialty choice should not be overlooked. These warnings highlight the importance of balancing technological literacy with the development of humanistic, ethical, and communication competencies. Extending these training strategies to other health science degrees will promote teamwork and a comprehensive approach to AI usage.

4.4 Limitations

Although this project provides valuable preliminary findings, it is important to acknowledge several limitations that affect the generalizability and robustness of the results. Firstly, the sample size was small, and participation was not mandatory, which not only impedes the representativeness of the general population of medical students but also introduces a non-response bias: those students who chose not to participate might hold different perceptions or attitudes regarding AI in education. Secondly, the study was conducted within the specific context of a seminar focused on AI, so the perceptions gathered could be influenced by the educational intervention itself, generating a potential acquiescence bias toward the presented environment.

Additionally, although the questionnaire used underwent a second iteration following a pilot with 14 participants and was agreed upon with expert educators, it lacks a formal psychometric validation process. The absence of objective questions that assess the actual level of knowledge limits the ability to contrast subjective perceptions with more direct indicators, and the simplicity of the instrument may not capture the real complexity of the perceptions, attitudes, and contextual factors that influence the use of AI in medical training environments.

To address these limitations, future research should consider using larger, more diverse samples with higher response rates to enhance representativeness and statistical power. It would also be advisable to evaluate the effectiveness of AI educational initiatives in different training contexts and over longer periods, as well as to refine and validate the questionnaire through rigorous psychometric analyses, incorporate objective questions, and encompass broader contextual factors. In this way, the conclusions drawn would be more robust, applicable, and generalizable to a wider range of medical education settings.

5 Conclusion

This teaching improvement project, aimed at describing the knowledge, attitudes, and perspectives of medical students regarding the application of AI and the use of chatbots in the healthcare field, revealed that participants are not significantly exposed to these tools nor are they a regular part of their academic or clinical routines. Despite this limited familiarity, they demonstrated a moderate awareness of the ethical challenges involved in incorporating AI into medical practice, reflecting an emerging sensitivity to the moral and deontological implications of these technologies.

At the same time, a marked optimism regarding the future adoption of AI-based tools was evident, as all students recognized the need to acquire knowledge in this area to perform effectively as healthcare professionals. This combination of ethical concerns and positive expectations underscores the importance of integrating specific AI-related educational content into medical education, enabling future doctors to use these tools effectively, thoughtfully, and responsibly.

Ultimately, the need to strengthen AI training within the medical curriculum not only responds to the growing presence of these technologies in healthcare delivery but also addresses the urgency of preparing tomorrow's physicians to leverage the opportunities offered by AI while resolving the complex ethical implications associated with its implementation.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

This project was conducted in accordance with current ethical guidelines and focused exclusively on the collection of anonymous data regarding students' perceptions and attitudes toward the described topic. No identifiable personal data or sensitive information related to participants' health or academic performance were collected. Participation was entirely voluntary, with no economic or academic incentives offered. Therefore, no Ethics Board approbation or informed consent was required.

Author contributions

JG-G: Conceptualization, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. LB-M: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. MB: Supervision, Writing – review & editing. AV: Writing – review & editing. IT-R: Conceptualization, Investigation, Methodology, Supervision, Writing – review & editing. AP: Conceptualization, Investigation, Methodology, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We thank all the students who attended the seminar and actively participated in the questionnaire.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. Generative AI tools have been used to translate and correct writing errors.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. (2019) 25:44–56. doi: 10.1038/s41591-018-0300-7

2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. (2022) 28:31–8. doi: 10.1038/s41591-021-01614-0

3. Garcia-Vidal C, Sanjuan G, Puerta-Alcalde P, Moreno-García E, Soriano A. Artificial intelligence to support clinical decision-making processes. *EBioMedicine*. (2019) 46:27–9. doi: 10.1016/j.ebiom.2019.07.019

4. Gomez C, Smith BL, Zayas A, Unberath M, Canares T. Explainable AI decision support improves accuracy during telehealth strep throat screening. *Commun Med*. (2024, 2024) 4:1, 149–111. doi: 10.1038/s43856-024-00568-x

5. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc*. (2018) 25:1248–58. doi: 10.1093/JAMIA/OCY072

6. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. *J Med Internet Res*. (2020) 22:e20346. doi: 10.2196/20346

7. Masters K. Artificial intelligence in medical education. *Med Teach*. (2019) 41:976–80. doi: 10.1080/0142159X.2019.1595557

8. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence–based Chatbots for promoting health behavioral changes: systematic review. *J Med Internet Res*. (2023) 25:e40789. doi: 10.2196/40789

9. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. *Big Data Soc*. (2016) 3:679. doi: 10.1177/2053951716679679

10. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics*. (2020) 46:205–11. doi: 10.1136/MEDETHICS-2019-105586

11. Rigby MJ. Ethical dimensions of using artificial intelligence in health care. *AMA J Ethics*. (2019) 21:E121–4. doi: 10.1001/AMAJETHICS.2019.121

12. Wartman SA, Donald CC. Medical education must move from the information age to the age of artificial intelligence. *Acad Med*. (2018) 93:1107–9. doi: 10.1097/ACM.0000000000002044

13. Sit C, Srinivasan R, Amlani A, Muthuswamy K, Azam A, Monzon L, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights. *Imaging*. (2020) 11:830. doi: 10.1186/S13244-019-0830-7

14. Paranjape K, Schinkel M, Panday RN, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR. *Med Educ*. (2019) 5:e16048. doi: 10.2196/16048

15. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. *JMIR Med Educ*. (2019) 5:e13930. doi: 10.2196/13930

16. Kolachalama VB, Garg PS. Machine learning and medical education. *NPJ Digit Med*. (2018) 1:54. doi: 10.1038/S41746-018-0061-1

17. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. *Soc Sci Med*. (2020) 260:113172. doi: 10.1016/j.socscimed.2020.113172

18. Floridi L, Cowls J. A unified framework of five principles for AI in society. *Harv Data Sci Rev*. (2019) 1:550. doi: 10.1162/99608F92.8CD550D1

19. Rosic A. Legal implications of artificial intelligence in health care. *Clin Dermatol*. (2024) 42:451–9. doi: 10.1016/J.CLINDERMATOL.2024.06.014

20. Babakhanian Z, Khorashadi RR, Vakili R, Abbasi MA, Akhavan H, Saeidi M. Analyzing the factors affecting students' absenteeism in university classrooms; a systematic review. *Syst Rev*. (2022) 3:563–75. doi: 10.22034/MEB.2022.353007.1064

21. Buabbas AJ, Miskin B, Alnaqi AA, Ayed AK, Shehab AA, Syed-Abdul S, et al. Investigating students' perceptions towards artificial intelligence in medical education. *Healthcare*. (2023) 11:1298. doi: 10.3390/healthcare11091298

22. Cherrez-Ojeda I, Gallardo-Bastidas JC, Robles-Velasco K, Osorio MF, Velez Leon EM, Leon Velastegui M, et al. Understanding health care students' perceptions, beliefs, and attitudes toward AI-powered language models: cross-sectional study. *JMIR Med Educ*. (2024) 10:e51757. doi: 10.2196/51757

23. Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell*. (2019) 1:501–7. doi: 10.1038/s42256-019-0114-4

24. Nundy S, Montgomery T, Wachter RM. Promoting trust between patients and physicians in the era of artificial intelligence. *JAMA*. (2019) 322:497–8. doi: 10.1001/jama.2018.20563

25. Jackson P, Ponath Sukumaran G, Babu C, Tony MC, Jack DS, Reshma VR, et al. Artificial intelligence in medical education - perception among medical students. *BMC Med Educ*. (2024) 24:1–9. doi: 10.1186/S12909-024-05760-0/TABLES/2

frontiers | Frontiers in Medicine

TYPE Perspective
PUBLISHED 23 January 2025
DOI 10.3389/fmed.2025.1527864

OPEN ACCESS

EDITED BY
Ariel Soares Teles,
Science and Technology of Maranhão, Brazil

REVIEWED BY
Radhika Devraj,
Southern Illinois University Edwardsville,
United States

*CORRESPONDENCE
Konstantinos Margetis
✉ Konstantinos.Margetis@mountsinai.org

RECEIVED 13 November 2024
ACCEPTED 09 January 2025
PUBLISHED 23 January 2025

CITATION
Aydin S, Karabacak M, Vlachos V and Margetis K (2025) Navigating the potential and pitfalls of large language models in patient-centered medication guidance and self-decision support.
Front. Med. 12:1527864.
doi: 10.3389/fmed.2025.1527864

Navigating the potential and pitfalls of large language models in patient-centered medication guidance and self-decision support

Serhat Aydin[1], Mert Karabacak[2], Victoria Vlachos[3] and Konstantinos Margetis[2]*

[1]School of Medicine, Koç University, Istanbul, Türkiye, [2]Department of Neurosurgery, Mount Sinai Health System, New York, NY, United States, [3]College of Human Ecology, Cornell University, Ithaca, NY, United States

Large Language Models (LLMs) are transforming patient education in medication management by providing accessible information to support healthcare decision-making. Building on our recent scoping review of LLMs in patient education, this perspective examines their specific role in medication guidance. These artificial intelligence (AI)-driven tools can generate comprehensive responses about drug interactions, side effects, and emergency care protocols, potentially enhancing patient autonomy in medication decisions. However, significant challenges exist, including the risk of misinformation and the complexity of providing accurate drug information without access to individual patient data. Safety concerns are particularly acute when patients rely solely on AI-generated advice for self-medication decisions. This perspective analyzes current capabilities, examines critical limitations, and raises questions regarding the possible integration of LLMs in medication guidance. We emphasize the need for regulatory oversight to ensure these tools serve as supplements to, rather than replacements for, professional healthcare guidance.

KEYWORDS

Large Language Models, ChatGPT, patient education, self-medication, artificial intelligence, machine learning, deep learning

KEY ASPECTS

- LLMs are transforming patient education by offering easily accessible and user-friendly guidance on medication use, improving patient understanding and self-management.
- These models may empower patients in remote or underserved areas by providing immediate, reliable information on health conditions and self-care, especially where healthcare access is limited.
- However, challenges remain in ensuring accuracy, particularly in complex cases due to the current limitations in accessing real-time data and personalized patient information.
- There are ethical concerns regarding the use of LLMs for self-medication guidance without healthcare oversight, which may lead to unintended health risks.
- To improve safety, future efforts should focus on integrating real-time medical databases and establishing clear regulations for the use of LLMs in healthcare contexts.

10.3389/fmed.2025.1527864

1 Introduction

The Large Language Models (LLMs) represent a significant advancement in patient education, particularly in personalized health and medication counseling. Leading examples such as OpenAI's ChatGPT (1), and Google's Gemini (2) can process extensive datasets and engage in conversational interactions. These artificial intelligence (AI) applications are increasingly being explored in healthcare to provide drug information, help patients navigate complex medication regimens, and guide initial responses to medical situations. By generating information of variable reliability, the extent to which LLMs can effectively influence patient autonomy in self-medication decisions and healthcare choices remains an open question.

The appeal of LLMs in healthcare stems from their accessibility and ease of use. Patients can readily access information about medication dosages, interactions, side effects, and alternatives without waiting to consult a healthcare provider. These models can enhance health literacy by translating medical jargon into plain language, helping patients make informed decisions about over-the-counter medications and some prescribed treatments. For example, studies show that LLMs can provide basic guidance for immediate-response situations, such as initial management of snakebites or other common conditions requiring urgent attention (3).

However, significant challenges exist in safely integrating LLMs into patient self-care decisions. A primary concern is the reliability of LLM-generated information, particularly regarding complex drug interactions or rare conditions. Cases of AI systems providing incorrect or misleading information have been documented, notably in sensitive areas with significant health and ethical implications, such as self-managed medication abortion (4).

Building upon our recent scoping review that identified six major themes in LLM applications for patient education (5), this article examines one critical theme: the role of LLMs in patient-centered medication guidance and self-decision support. We assess both the potential of LLMs to enhance autonomous medication use and the risks associated with their misuse or misunderstanding. This perspective article reviews recent advances, identifies key challenges, and proposes future directions for LLM implementation that balance patient autonomy with healthcare safety and ethical standards. By examining this specific theme in detail, we aim to contribute targeted insights into the responsible integration of LLM technology in medication guidance while addressing critical questions about patient safety and ethical implementation.

2 Current advances in LLMs for customized medication use and self-decision

2.1 LLMs as informational aids for drug interactions and side effects

LLMs show promise as informational resources for medication guidance, particularly in explaining drug interactions, potential side effects, and usage instructions. These models can translate complex pharmacological information into accessible language for patients with limited medical knowledge. This capability helps patients better understand their medication regimens and may reduce drug-drug interactions caused by misunderstandings (6, 7).

A recent study by Iqbal et al. examined ChatGPT's reliability as a secondary opinion source for dermatological treatments (8). While dermatologists approved 98.87% of the model's medication suggestions, they identified limitations such as incorrect Anatomical Therapeutic Chemical codes and errors in drug route specifications. These findings suggest that while ChatGPT shows promise for general treatment guidance, it requires further refinement for precise clinical applications.

LLMs also demonstrate potential in helping patients manage complex medication regimens, particularly in cases of polypharmacy where drug-drug interactions pose significant risks. Research shows that these models can effectively identify and explain risks associated with specific drug combinations, including interactions between over-the-counter medications and treatments for chronic conditions (9). This capability could help prevent medication errors and resulting hospitalizations from adverse drug reactions.

Recent research also explores LLMs' potential in helping healthcare professionals screen for drug interactions. A comparative analysis of ChatGPT, Google Bard, and Bing AI found that while these tools do not yet match the accuracy of specialized clinical software, they can effectively identify relevant drug interactions in real-time. Among the tested models, Bing AI demonstrated the highest accuracy and specificity, while ChatGPT-4 showed improvements over its predecessor (6). These findings highlight the need for further development of LLM capabilities, indicating that while they show potential, they are not yet ready for reliable use in clinical settings but may be in the future.

2.2 Facilitating self-decision in self-administered treatments

LLMs show potential in guiding patients through self-administered treatments, particularly in situations requiring immediate action. For example, studies have evaluated ChatGPT's ability to provide first-aid advice for venomous snakebites while emphasizing the need for urgent medical care (3). This capability could be particularly valuable in remote areas with limited healthcare access, offering patients guidance to take appropriate immediate actions while awaiting professional care. Infrastructural challenges, such as unreliable internet connectivity, may hinder its implementation in such settings, though its potential remains promising. However, researchers found that while ChatGPT-3.5 provided reliable general guidance, it should not replace professional medical consultation, especially in critical situations. The study emphasized the need for continued improvements to enhance AI's reliability in high-stakes medical scenarios.

Roosan et al. evaluated ChatGPT's effectiveness in Medication Therapy Management, focusing on drug interaction identification and therapeutic adjustments (10). While ChatGPT-4 demonstrated high accuracy with simple and moderately complex cases, it showed limitations when handling complex scenarios requiring patient-specific considerations. The model proved capable of identifying common drug-drug interactions but struggled with personalized dosage adjustments, highlighting the continued need for human oversight in clinical decision-making.

3 Challenges and limitations in LLMs for medication guidance and self-decision

3.1 Inaccuracy and misleading information

A critical challenge in using LLMs for medication guidance is their potential to generate inaccurate or misleading information. While these models can process large datasets, they lack access to real-time, continuously updated medical databases, potentially leading to outdated or incorrect advice. For example, studies have found that ChatGPT-3.5 provided inaccurate information about self-managed medication abortion, exaggerating risks despite evidence supporting its safety when properly administered (4). Such misinformation can increase patient anxiety, perpetuate stigma, and discourage evidence-based healthcare decisions.

Research by Sheikh et al. compared ChatGPT-3.5 and ChatGPT-4's ability to assess the safety of non-prescription medications and supplements for patients with kidney disease (11). While ChatGPT-4 showed improvement over its predecessor (81.4% vs 64.5% concordance with Micromedex), neither matched the reliability of established drug information resources. Both models particularly struggled with supplement safety assessments, often defaulting to "unknown toxicity" classifications due to limited data.

Rao et al. (9) assessed ChatGPT-3.5's role in managing polypharmacy in geriatric patients, finding its deprescribing recommendations aligned with guidelines for patients without cardiovascular disease but lacked accuracy when factoring in functional impairments and cardiovascular history. Notably, it often recommended deprescribing pain medications without considering older adults' pain management needs. Similarly, in cases of renal dysfunction, ChatGPT achieved only 16.7% accuracy in dose adjustments incorporating patient-specific variables such as renal markers and comorbidities (12). These findings highlight the limitations of LLMs in complex scenarios requiring personalized clinical expertise, emphasizing their role as supplementary tools rather than replacements for professional judgment. This low accuracy poses significant risks in clinical settings where precise dosing is crucial, demonstrating that while LLMs may support preliminary decision-making, they cannot reliably replace clinical expertise in complex medical situations.

3.2 Ethical and safety concerns in self-decision support

The use of LLMs for self-medication guidance raises significant ethical concerns, particularly when patients use these tools without healthcare professional oversight. A primary risk is that LLMs may provide seemingly authoritative advice that lacks clinical nuance, potentially encouraging unsafe medical decisions. This risk is heightened in regions with limited healthcare access, where patients might rely on AI as their primary medical information source.

Hsu et al. examined ChatGPT's ability to handle medication consultations and drug-herb interaction questions (13). While the model effectively addressed basic public inquiries, it performed poorly on complex questions from healthcare providers. The study revealed particular limitations in analyzing interactions between traditional Chinese and Western medicines, often providing vague or incomplete information. These findings indicate that while ChatGPT can help with basic medication questions, it currently lacks the sophistication needed for reliable guidance in specialized clinical contexts.

Ethical concerns also emerge in managing sensitive medical conditions, such as cancer. When evaluated for cancer symptom management guidance, ChatGPT's recommendations showed notable discrepancies from National Comprehensive Cancer Network (NCCN) guidelines. The model tended to provide generalized advice that failed to address the complex symptom burdens typical of cancer patients (14). This gap between AI-generated recommendations and evidence-based guidelines underscores the risks of relying on LLMs for critical health decisions.

Privacy constraints prevent LLMs from accessing individual medical records, limiting their ability to provide personalized recommendations. This limitation is particularly problematic for high-risk populations, including elderly patients and those with chronic illnesses, who require carefully tailored treatment plans. Without access to patient-specific data, LLMs default to generalized advice that may be inappropriate or unsafe for complex medical conditions. As demonstrated in previous research, ChatGPT's inability to consider specific renal function metrics led to incorrect dosing recommendations for patients with kidney disease, illustrating the potential safety risks of such limitations (12).

These limitations highlight the critical need for a structured ethical framework governing LLM deployment in healthcare. The integration of AI into patient self-decision support requires a balanced approach that positions these tools as supplements to, not replacements for, professional medical expertise. A collaborative model combining AI capabilities with clinical oversight could optimize the benefits of LLMs while minimizing risks. The development of robust regulatory guidelines will be essential to harness LLM potential while maintaining patient safety and ethical standards.

4 Future directions and recommendations

4.1 Improving accuracy and reliability of LLMs for medication-related information

Enhancing LLM reliability for medication guidance requires integration with real-time medical databases and continuous content updates. Connecting these models to current pharmacological databases would enable access to the latest drug interaction guidelines, side effect profiles, and dosage recommendations. Such integration could help align AI systems with evolving healthcare information while improving response accuracy for patient inquiries. Development of frameworks allowing LLMs to access validated sources such as PubMed, FDA databases, and regional repositories would strengthen the clinical relevance of their recommendations.

Specialized training protocols represent another key avenue for improvement, particularly in enhancing LLMs' contextual understanding of patient inquiries. Targeted training in medical ethics and patient safety could reduce risks in high-stakes areas such as mental health, reproductive health, and complex medication management. Collaboration between healthcare professionals and AI developers is crucial for ensuring these models meet clinical standards. By involving medical experts in model refinement, especially for context-specific information and decision-making guidance, developers can better align AI outputs with the nuanced requirements of personalized medicine. Strategic partnerships between AI companies and medical institutions could facilitate ongoing model validation and improvement.

4.2 Balancing autonomy with safety: ethical and regulatory perspectives

The growing role of LLMs in medication guidance necessitates an ethical framework balancing patient autonomy with safety. Our previous scoping review highlighted that while LLMs effectively simplify medical terminology, they often lack reliability in critical, high-stakes scenarios (5). This finding underscores the need for comprehensive regulatory standards ensuring transparency in AI recommendations, including clear disclaimers about the importance of professional medical consultation. Such guidelines would help users understand that AI-generated advice supplements, rather than replaces, clinical expertise.

Looking forward, establishing medical AI ethical review boards, similar to institutional review boards for clinical research, could provide structured oversight of LLM implementation. These boards could evaluate training data, assess response biases, and monitor AI applications in patient education and self-care. This framework would ensure AI development aligns with patient safety priorities and evolving healthcare policies.

5 Discussion

LLMs show promise in supporting patient self-decision making for medication use, providing accessible, on-demand resources for drug-related information. These tools help patients explore questions about drug interactions, side effects, and medication schedules, potentially enhancing health literacy and informed decision-making. However, significant limitations and risks exist. The inability of LLMs to incorporate individual patient data, including medical histories and current medications, creates a fundamental barrier to personalized advice. this limitation, combined with potential inaccuracies in AI-generated responses, necessitates careful integration of LLMs into healthcare, particularly in sensitive areas such as reproductive and mental health.

In environments where access to healthcare professionals is limited or communication systems are disrupted, such as remote areas or disaster zones, LLMs can provide support for patient self-care. These AI tools can deliver immediate, situation-specific advice for managing medical concerns when professional help is unavailable. This immediate guidance can be life-saving in cases where there are no healthcare facilities nearby, offering a sense of empowerment and structured steps for non-professionals facing medical emergencies. Nevertheless, while LLMs can provide a valuable bridge until medical assistance is available, they cannot replace the expertise of healthcare professionals in complex or high-stakes situations. As such, their recommendations should emphasize the provisional nature of AI guidance in austere environments, ideally directing individuals to seek professional care as soon as circumstances allow.

In addition to emergencies, LLMs can be used to support patients in everyday medication decisions, particularly with over-the-counter (OTC) drugs. Many individuals may not fully understand the risks of combining OTC medications with prescription drugs or specific medical conditions, often due to the complex and lengthy drug information provided on packaging. Patients may also assume OTC medications are inherently safe or may avoid consulting healthcare professionals for minor issues. In such cases, LLMs can assist by analyzing drug information and identifying potential interactions or contraindications based on a patient's reported medications and medical conditions. This guidance can help patients make safer choices, promoting informed self-care in routine health decisions. However, the accuracy and safety of these recommendations depend on LLMs being continuously updated with the latest clinical data. The potential for adverse outcomes highlights the need for rigorous oversight, ensuring that LLM-driven advice is a safe, supplementary resource in patient-centered healthcare.

While LLMs can empower patients with information, the risks of misinformation or oversimplified guidance are substantial, especially if patients bypass professional medical consultation in favor of AI recommendations. Future developments must address both accuracy and ethical considerations. Key improvements should include integrating validated medical databases and increased collaboration with healthcare professionals. Additionally, regulatory oversight must establish clear boundaries for LLM use, ensuring these tools serve as supportive rather than standalone resources. Clear disclaimers and transparent communication about AI limitations can help position LLMs as supplements to professional healthcare guidance.

LLMs represent a transformative development in patient education, potentially reshaping how patients approach self-medication and health decisions. Their successful implementation depends on addressing current limitations in probabilistic data synthesis, personalization capabilities, and ethical considerations in sensitive healthcare areas. The path forward requires balancing AI's informational capabilities with professional medical guidance while maintaining focus on patient safety and autonomy. This balanced approach will be crucial for realizing the full potential of LLMs in patient-centered healthcare.

Data availability statement

The original contributions presented in this study are included in this article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SA: Conceptualization, Data curation, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing. MK: Conceptualization, Data curation, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review and editing. VV: Conceptualization, Data curation, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing. KM: Conceptualization, Data curation, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. OpenAI. Available online at: https://openai.com/index/hello-gpt-4o (accessed May 13, 2024). (2024).

2. Google. Available at: https://gemini.google.com/ (accessed 2024). (2023).

3. Altamimi I, Altamimi A, Alhumimidi A, Altamimi A, Temsah M. Snakebite advice and counseling from Artificial intelligence: An acute venomous snakebite consultation with chatgpt. *Cureus.* (2023) 15(6):e40351. doi: 10.7759/cureus.40351

4. McMahon H, McMahon B. Automating untruths: Chatgpt, self-managed medication abortion, and the threat of misinformation in a post-roe world. *Front Digit Health.* (2024) 6:1287186. doi: 10.3389/fdgth.2024.1287186

5. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: A scoping review of applications in medicine. *Front Med.* (2024) 11:1477898. doi: 10.3389/fmed.2024.1477898

6. Al-Ashwal F, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar A. Evaluating the sensitivity, specificity, and accuracy of chatgpt-3.5, chatgpt-4, bing ai, and bard against conventional drug-drug interactions clinical tools. *Drug Healthc Patient Saf.* (2023) 15:137–47. doi: 10.2147/DHPS.S425858

7. Juhi A, Pipil N, Santra S, Mondal S, Behera J, Mondal H. The capability of chatgpt in predicting and explaining common drug-drug interactions. *Cureus.* (2023) 15(3):e36272. doi: 10.7759/cureus.36272

8. Iqbal U, Lee L, Rahmanti A, Celi L, Li Y. Can large language models provide secondary reliable opinion on treatment options for dermatological diseases? *J Am Med Inform Assoc.* (2024) 31(6):1341–7. doi: 10.1093/jamia/ocae067

9. Rao A, Kim J, Lie W, Pang M, Fuh L, Dreyer K, et al. Proactive polypharmacy management using large language models: Opportunities to enhance geriatric care. *J Med Syst.* (2024) 48(1):41. doi: 10.1007/s10916-024-02058-y

10. Roosan D, Padua P, Khan R, Khan H, Verzosa C, Wu Y. Effectiveness of chatgpt in clinical pharmacy and the role of artificial intelligence in medication therapy management. *J Am Pharm Assoc.* (2024) 64(2):422-8 e8. doi: 10.1016/j.japh.2023.11.023.

11. Sheikh M, Barreto E, Miao J, Thongprayoon C, Gregoire J, Dreesman B, et al. Evaluating Chatgpt's efficacy in assessing the safety of non-prescription medications and supplements in patients with kidney disease. *Digit Health.* (2024) 10. doi: 10.1177/20552076241248082

12. van Nuland M, Snoep J, Egberts T, Erdogan A, Wassink R, van der Linden P. Poor performance of chatgpt in clinical rule-guided dose interventions in hospitalized patients with renal dysfunction. *Eur J Clin Pharmacol.* (2024) 80(8):1133–40. doi: 10.1007/s00228-024-03687-5

13. Hsu H, Hsu K, Hou S, Wu C, Hsieh Y, Cheng Y. Examining real-world medication consultations and drug-herb interactions: Chatgpt performance evaluation. *JMIR Med Educ.* (2023) 9:e48433. doi: 10.2196/48433

14. Lazris D, Schenker Y, Thomas T. Ai-generated content in cancer symptom management: A comparative analysis between Chatgpt and Nccn. *J Pain Symptom Manage.* (2024) 68(4):e303–11. doi: 10.1016/j.jpainsymman.2024.06.019

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 13 March 2025
DOI 10.3389/fmed.2025.1529335

OPEN ACCESS

EDITED BY
Alaa Abd-alrazaq,
Weill Cornell Medicine-Qatar, Qatar

REVIEWED BY
Balachandran Manavalan,
Sungkyunkwan University, Republic of Korea
Watshara Shoombuatong,
Mahidol University, Thailand

*CORRESPONDENCE
Yang Zhang
✉ yangzhang@cdutcm.edu.cn
Hua Tang
✉ huatang@swmu.edu.cn
Hao Lin
✉ hlin@uestc.edu.cn

RECEIVED 16 November 2024
ACCEPTED 27 February 2025
PUBLISHED 13 March 2025

CITATION
Zhang H-Q, Arif M, Thafar MA, Albaradei S, Cai P, Zhang Y, Tang H and Lin H (2025) PMPred-AE: a computational model for the detection and interpretation of pathological myopia based on artificial intelligence. *Front. Med.* 12:1529335.
doi: 10.3389/fmed.2025.1529335

PMPred-AE: a computational model for the detection and interpretation of pathological myopia based on artificial intelligence

Hong-Qi Zhang[1], Muhammad Arif[2], Maha A. Thafar[3], Somayah Albaradei[4], Peiling Cai[5], Yang Zhang[6]*, Hua Tang[7,8]* and Hao Lin[1]*

[1]The Clinical Hospital of Chengdu Brain Science Institute, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China, [2]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar, [3]Computer Science Department, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia, [4]Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia, [5]School of Basic Medical Sciences, Chengdu University, Chengdu, China, [6]Innovative Institute of Chinese Medicine and Pharmacy, Academy for Interdiscipline, Chengdu University of Traditional Chinese Medicine, Chengdu, China, [7]School of Basic Medical Sciences, Southwest Medical University, Luzhou, China, [8]Central Nervous System Drug Key Laboratory of Sichuan Province, Luzhou, China

Introduction: Pathological myopia (PM) is a serious visual impairment that may lead to irreversible visual damage or even blindness. Timely diagnosis and effective management of PM are of great significance. Given the increasing number of myopia cases worldwide, there is an urgent need to develop an automated, accurate, and highly interpretable PM diagnostic technology.

Methods: We proposed a computational model called PMPred-AE based on EfficientNetV2-L with attention mechanism optimization. In addition, Gradient-weighted class activation mapping (Grad-CAM) technology was used to provide an intuitive and visual interpretation for the model's decision-making process.

Results: The experimental results demonstrated that PMPred-AE achieved excellent performance in automatically detecting PM, with accuracies of 98.50, 98.25, and 97.25% in the training, validation, and test datasets, respectively. In addition, PMPred-AE can focus on specific areas of PM image when making detection decisions.

Discussion: The developed PMPred-AE model is capable of reliably providing accurate PM detection. In addition, the Grad-CAM technology was also used to provide an intuitive and visual interpretation for the decision-making process of the model. This approach provides healthcare professionals with an effective tool for interpretable AI decision-making process.

KEYWORDS

myopia, pathological myopia, deep learning, EfficientNetv2, Grad-CAM

1 Introduction

Pathological myopia (PM) is a serious visual disease that can lead to irreversible visual damage or even blindness (1–3). In recent years, PM has become one of the main causes of visual impairment and permanent blindness worldwide, especially in Asian countries. According to the research by Holden et al. (4), by 2050, nearly half of the global population will be affected by myopia, with approximately 10% suffering from high myopia, which will also become the leading cause of permanent blindness. In addition, retinopathy and complications related to myopia may also increase the risk of visual damage (5–7). Therefore, timely diagnosis and early detection of PM are crucial. Currently, develop an automated, accurate, and non-invasive method PM diagnosis method is an urgent task.

With the development of artificial intelligence (AI) and the accumulation of myopia data, a variety of computational methods have been developed (8–10). For example, Liu et al. (10) introduced a method using texture features and Support Vector Machine (SVM) (11–13) to automatically detect PM. This method processed retinal fundus images by extracting region of interest (ROI) and detecting the optic nerve head. Subsequently, texture-based metrics were generated, categorized and grouped into zones for context-based generation of features. Finally, SVM was used to detect PM based on these features, achieving an accuracy (ACC) of 87.5% (14). Zhang et al. (15) proposed an automatic detection method for PM based on max-relevance and min-redundancy (mRMR). This method built a feature space from information extracted from fundus images and medical screening data, created a ranked feature library using mRMR, searched for the most compact feature set with a forward selection wrapper, and then used SVM for detection. As a result, they achieved an ACC of 89.3% for the right eye and 88.5% for the left eye (15). Xu et al. (16) developed a detection method for PM based on bag-of-feature and sparse learning. During the training phase, the codebook for the bag-of-feature model and the classification model were learned, and the top related visual features were discovered through sparse learning.

In the detection phase, local features were first extracted from a given retinal fundus image, quantified using the learned codebook to obtain global features. Finally, the classification model was used to determine the presence of PM, achieving an ACC of 90.6% (16). Zhang et al. (17) also developed an automatic diagnostic method for PM based on heterogeneous biomedical data, integrating data from various sources including imaging data, demographic/clinical data, and genotyping data, and ultimately using a multiple kernel learning (MKL) approach to accurately detect PM, achieving an average Area Under Curve (AUC) of 0.888. Chen et al. (18) introduced a deep learning architecture for automating the diagnosis of glaucoma. This method used a convolutional neural networks (CNN) (19, 20) model with four convolutional layers and two fully connected layers, combined with dropout and data augmentation strategies to enhance diagnostic performance. The method achieved AUC values of 0.831 and 0.887 on the ORIGA and SCES datasets, respectively (18). Xu et al. (21) proposed an automated detection method for tessellated fundus based on texture features, color features and SVM. The method could achieve an ACC of 98%. Xu et al. (22) proposed a method for detecting ocular disease based on multiple informatics domains. This method combined pre-learned SVM classifiers effectively merging personal demographic data, genome information, and visual information from retinal fundus images. The final model obtained an AUCs of 0.935 for glaucoma, 0.822 for age-related macular degeneration (AMD), and 0.946 for PM (22). Septiarini et al. (23) introduced a method based on statistical features to automatically detect peripapillary atrophy in retinal fundus images. This method involved four steps: optic nerve head (ONH) localization, ONH segmentation, preprocessing, and features extraction. Through these steps, three key features were extracted: standard deviation (σ), smoothness (S), and third moment (μ3). By using a backpropagation neural network (BPNN), they achieved an ACC of 95% (23). Rauf et al. (24) proposed a CNN-based method for PM detection and obtained an ACC of 95%. Although these studies have achieved positive results, there are still several challenges: (1) Many advanced deep learning methods are emerging, but in the field of PM detection, these advanced technologies have not yet been applied. (2) Due to the uniqueness of the medical industry and the high requirements for model accuracy, model performance still needs to be improved. (3) Due to the differences in actual medical facilities, the efficiency of these models in poorly equipment medical environments is an important problem that needs to be overcome. (4) As an auxiliary diagnosis method, the interpretability of models was an important task, but current research in this area is still insufficient (25–28).

To address the aforementioned challenges, this study designed an improved model named PMPred-AE based on EfficientNetV2-L to automatically identify and diagnose PM. This study further enhanced the model's ability to identify key features in the retina images by introducing the attention mechanism, thereby improving the accuracy of the diagnosis of PM. In order to provide visual explanations for the decision-making process of the model, we also adopted the Gradient-weighted class activation mapping (Grad-CAM) technique. Our study provides an efficient, accurate, and explainable model for the detection of PM.

2 Materials and methods

2.1 Dataset construction

The study utilized the PALM Challenge dataset, comprising training images, verification images and test images. The training dataset contains 187 non-PM and 213 PM. Similarity, the verification set consists of 400 images, with 189 labeled as non-PM and 211 as PM. Additionally, test set includes 400 images with corresponding labels: 187 categorized as non-PM and 213 as PM (29). This dataset configuration enabled rigorous evaluation and validation of the proposed methodologies.

2.2 Model design

The PMPred-AE architecture consists of two core components: a feature extractor and a classifier. In the feature extraction stage, we chose EfficientNetV2-L, an advanced CNN model aimed at accelerating image processing and improving its performance. As an upgraded version of the EfficientNet series, EfficientNetV2-L underwent pre-trained on a massive ImageNet dataset that covers millions of images and thousands of categories. Through its scalable architecture, EfficientNetV2-L cleverly balances the network depth, width, and resolution to achieve optimal performance and efficiency.

EfficientNetV2-L is an upgraded version of the EfficientNet series. It optimizes the balance of network depth, width, and resolution to achieve high efficiency and accuracy in image processing tasks. Compared to advanced vision transformer (ViT) series' ViT-L/16, EfficientNetV2-L achieves higher accuracy. Meanwhile, the training speed could increase by 7 times (30). In particular, the model utilizes lightweight depthwise separable convolution techniques, significantly reducing computational burden and model size while maintaining efficient feature extraction capabilities. Therefore, in the context of PM-detection, EfficientNetV2-L could efficiently identify key features in images and provide accurate data input for classifiers, significantly improving the performance of the model. Moreover, its superior computing speed and efficiency made it very suitable for application in medical environments with rudimentary equipment, providing strong technical support for early diagnosis and treatment. In the classification stage, we used an improved fully-connected neural network based on the attention mechanism. The core function of this improvement is to enhance the model's attention to the most important parts of the input features. By assigning different weights to the input features, the attention mechanism allows the model to prioritize the features that contribute the most to the final classification decision, rather than treating all input features equally. This dynamic weight allocation method not only improves the model's understanding of the data, but also increases the adaptability and flexibility of the model, enabling it to automatically focus on the most critical information. Specifically, we used a linear layer to transform all the features into a one-dimensional space, and then map them to a value between 0 and 1 using the Softmax function. Finally, this weight is multiplied by the original input features to emphasize the features that contribute the most to the classification result. This improvement was particularly important for the detection of PM. It allows the model to pay special attention to the areas that revealed the pathological features of myopia. Through this mechanism, our model provided an efficient tool for the early diagnosis and treatment of PM.

2.3 Grad-CAM

In order to visually explain the decision-making process of CNN in PM detection tasks, we used Grad-CAM technique to generate a heatmap. Through Grad-CAM, we can clearly see which areas are given more attention when the model makes detection. This approach relies on the gradient information of the model, particularly focusing on the gradients of the feature layers from the last convolutional layer, to highlight the regions that contribute most to the model predictions. The working principle of Grad-CAM can be briefly described by the following mathematical expression.

First, for each channel in the feature layer A, the global average pooling of these slopes is calculated to obtain the weight coefficient (Equation 1):

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

where, y^c is the output score of the model for category c, A_{ij} is the activation value of the feature layer at position (i,j), k is the k-th channel in the feature layer A, and Z is the total number of units in the feature layer.

Then, the weight coefficient is multiplied by the activation value of the feature layer and then accumulated. The final heatmap is generated by filtering through the *ReLU* function (Equation 2):

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{2}$$

This process ensures that only features that have positive impact on l prediction category c of the mode were visualized, thereby enhancing the clarity and interpretability of the model's decision. By applying Grad-CAM to the PMPred-AE model, the heatmap clearly reveals that the model focuses on the location of key pathological changes in the retina image when identifying PM. The heatmap provided by Grad-CAM not only demonstrates the reason behind the model's high performance, but also proves its focusing ability, which is crucial to improve the reliability and trust of the model in practical medical applications. Through this way, Grad-CAM provides healthcare professionals with an intuitive tool to better understand and explain the decision-making process of the PMPred-AE, especially in medical diagnosis and treatment planning.

2.4 Parameter setting

The learning rate is set to 0.0001, the batch size is 8, the number of epochs is 50, and the optimizer is AdamW.

2.5 Evaluation index

Several widely used evaluation indicators (31–37), including precision (Pre) (Equation 3), recall (Rec) (Equation 4), accuracy (ACC) (Equation 5), F1-score (F1) (Equation 6), and Matthew's coefficient of association (MCC) (Equation 7), were utilized to evaluate model's performance, defined as follows:

$$Pre = \frac{TP}{TP + FP} \tag{3}$$

$$Rec = \frac{TP}{TP + FN} \tag{4}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$F1 = \frac{2PreRec}{Pre + Rec} \tag{6}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \tag{7}$$

where *TP*, *TN*, *FP*, and *FN* represented the true positive, true negative, false positive, and false negative of the sample, respectively. We also drew the receiver operating characteristic curve (ROC) and precise recall curve (PRC), and obtained the area under the curve (AUC, AUPRC) (27, 38–41).

3 Results

3.1 Overview of experiment

In our experiment, we first adopted data augmentation techniques to enrich and expand the original data set, and created more diverse training samples. Data enhancement included operations such as image rotation, resizing, and cropping. It was designed to simulate different shooting conditions and perspectives to improve the model's generalization and robustness. The data-enhanced dataset was used to train our PMPred-AE model, which was based on the EfficientNetV2-L architecture and optimized to meet the specific requirements of PM-detection. EfficientNetV2-L is the foundation of our model. It has been pre-trained on the ImageNet data set, and therefore has strong feature extraction capabilities (42, 43). In order to further improve the performance of the model, we introduced an attention mechanism in the fully connected layer of the model. This mechanism enables the model to focus more on the key areas related to PM diagnosis in the image, thereby improving the accuracy of diagnosis. During the model training process, the model parameters were adjusted based on the performance on the verification set to achieve the optimal configuration. After training, we visualized the output of the model at different levels (shallow, middle, and deep). This step helped us understand how the model gradually extracted and utilized image features. In addition, we also used Grad-CAM technology to generate a heatmap that highlight the areas that the model focuses on when making predictions. In this way, we can not only verify the decision-making process of the model, but also provide intuitive visual explanations for doctors to help them better understand the basis of the model. Overall, our experiment combined data augmentation, attention mechanisms, and advanced model architecture and explanatory techniques to develop an efficient, accurate, and explainable model for the detection of PM (Figure 1).

3.2 Data augmentation

Due to the difficulty of collecting and annotating pathological images, only a small number of data samples could be collected under normal circumstances. Therefore, data augmentation was a very necessary task. It can effectively reduce the over-fitting degree of the model, and allow the model to learn more general knowledge instead of focusing too much on noise and some unique features, thereby improve the generalization and robustness of the model (44–46). In this study, we employed a combination approach for sample augmentation. The detailed procedure included initially resizing the images to 256×256 pixels. Subsequently, they are randomly cropped to 224×224 pixels. Then anti-aliasing techniques were applied to ensure image quality. In addition, to increase visual variety, the probability of horizontal and vertical flipping was set to 50%. This method also incorporated subtle random affine transformations, including rotations between −10 to 10 degrees, translations of 10% of the image width or height, and scaling between 90 to 110%. Furthermore, random erasure is applied with a 50% probability, randomly covering a small portion of the image, enhancing the model's ability to handle image occlusion (Figure 2). Finally, the images were converted into tensors and normalized according to a specific mean and standard deviation to suit the needs of model training. We mainly used these methods to address the following issues: by randomly cropping and resizing, we simulated the scene where doctors observe the eyes from different distances and angles, and random rotation and affine transformation helped the model identify pathological features from multiple angles. Random erasure simulates potential occlusions during actual medical image acquisition. Normalization ensures consistency of image data during training, while anti-aliasing maintains the clarity of image details, which is crucial for identifying pathological features. By introducing various visual perturbations, this comprehensive data augmentation strategy facilitates the model in extracting valuable features from diverse image transformations, thereby enhancing performance and robustness in real-world application scenarios.

3.3 Model validation

A series of experiments have shown that the PMPred-AE model exhibits excellent performance in PM classification tasks. Firstly, the model is trained on the training set to ensure that it has sufficient learning foundation and can capture the key features and patterns in the data (47–50). Then, the validation set was used to adjust the parameters of the model, which further improved its performance and ensured its generalization ability on unseen data (51, 52). The experimental results showed that PMPred-AE performed well on the test set, and all evaluation indicators reached a very high level, such as ACC, F1, Pre, Rec and MCC with values of 0.9725, 0.9744, 0.9676, 0.9812 and 0.9448, respectively. This indicates that PMPred-AE has excellent ability to effectively distinguish PM from non-PM (Figure 3A, Table 1). In addition, by plotting ROC and PRC, we observed that the PMPred-AE model had good AUC and AUPRC under both conditions, with values of 0.9955 and 0.9962, respectively. This further demonstrated the efficiency of PMPred-AE model in feature extraction and capability in recognizing PM (Figures 3B,C). Finally, we used t-SNE technology to visualize the output of the model (Figure 3D) (53). The results showed that PM and non-PM can be clearly distinguished in a low-dimensional space, indicating that the model can effectively represent their features in a low-dimensional space and capture the complex patterns and structural differences between them. This further suggests that the PMPred-AE model has broad application prospect in clinical practice.

3.4 Model explanations

To further confirm that PMPred-AE could effectively extract features, we visualized the output of the model's shallow, middle, and deep layers. It can be clearly observed that as the depth of the model increases, the model can extract more abstract and higher-level features. This proves that the hierarchical structure of

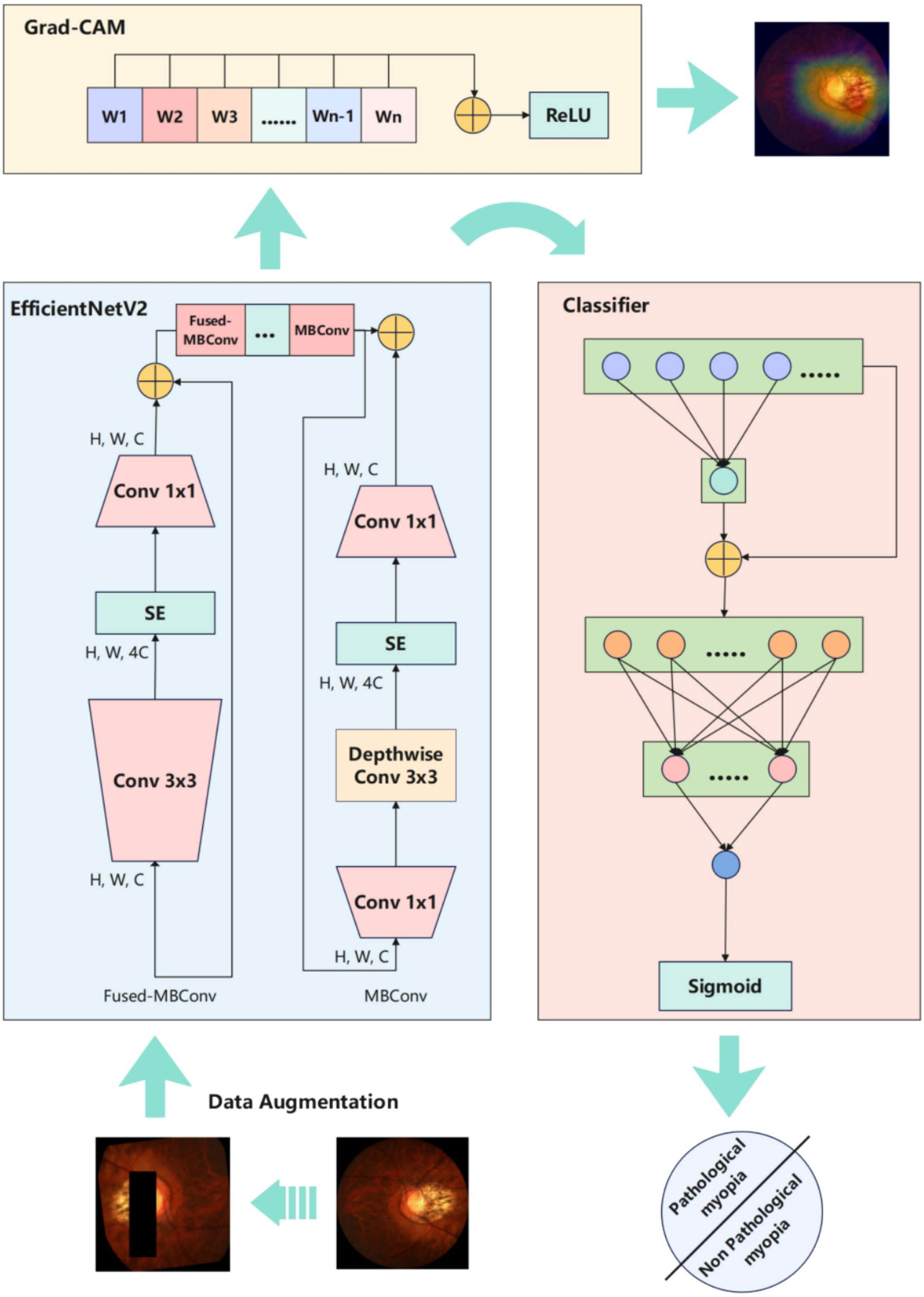


FIGURE 1
Experimental workflow overview diagram.

PMPred-AE model effectively promoted the gradual extraction and refinement of features (Figure 4A). Later, in order to further investigate why PMPred-AE could efficiently distinguish PM and non-PM, we used the Grad-CAM technology to generate a heatmap that could reveal the areas that the model focused on when making predictions, thus providing an explanation for the model's decision-making process (Figure 4B). The heatmap revealed that the PMPred-AE model could effectively focus on the location of the key pathological changes in the image when identifying PM. These positions were often the key for distinguishing between PM and non-PM, which explained why the model could achieve high accuracy. This focusing ability not only improved the prediction

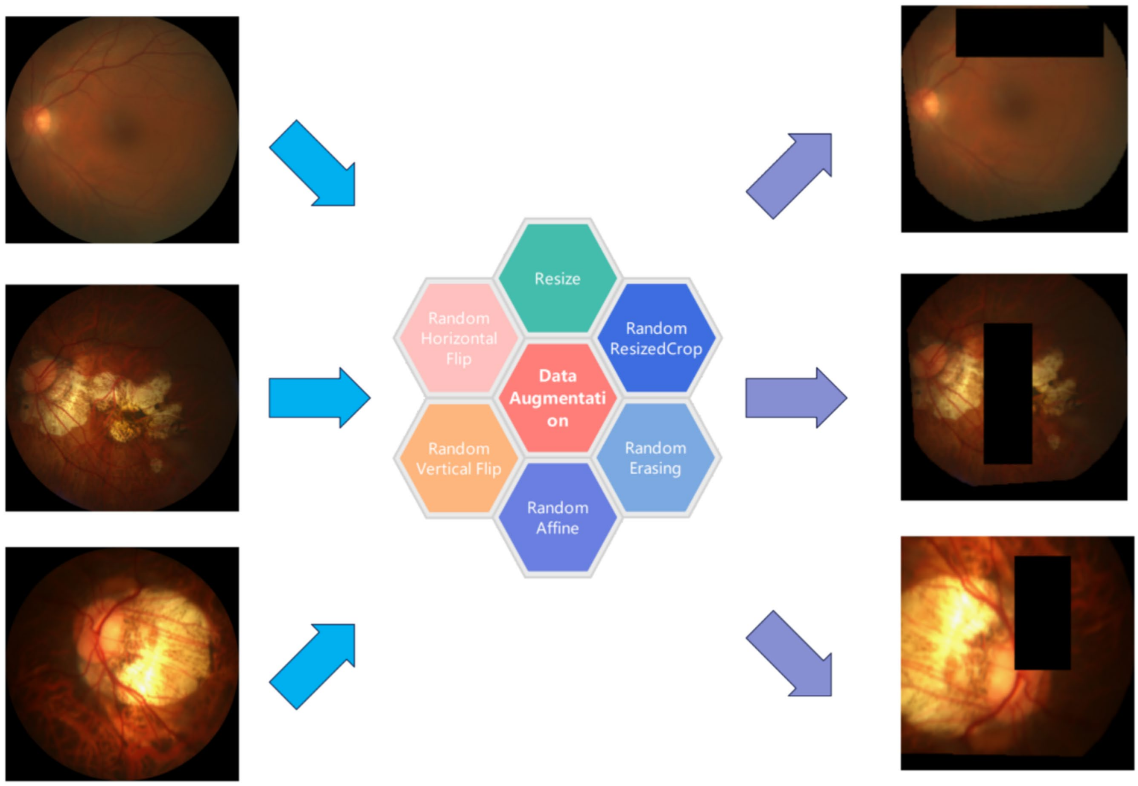


FIGURE 2
Data augmentation result diagram.

performance of the model, but also increased its reliability and credibility in practical applications, especially in medical diagnosis and treatment planning.

3.5 Comparisons with existed works

To further demonstrate the performance of PMPred-AE in detecting PM, we should compare the proposed model with existed studies. However, those studies we mentioned earlier did not share their source code and used different datasets, making it impossible for use to make a fair comparison. Fortunately, we could use the PALM's benchmark data from 2023 (Base-2023) (29). The experiment results showed that among all evaluation metrics, PMPred-AE is superior to Base-2023 (Figure 5, Table 2). By comparing with Base-2023, we further consolidated the validation of the PMPred-AE model and provided more reliable support for its application in clinical practice.

4 Discussion

In this study, we designed an improved EfficientNetV2-L model based on the attention mechanism (PMPred-AE) for the automatic detection of PM. By using EfficientNetV2-L as the basic architecture for feature extraction and introducing improvements based on the attention-based mechanism in the classification stage, the PMPred-AE model could efficiently identify key features in eye image and significantly improve the prediction performance of the model. In the research, data augmentation techniques were used to expand the training samples, including image rotation, resizing, and cropping to improve the model's generalization ability and reliability. In addition, Grad-CAM technology was introduced during the model training process to generate heatmaps, which provided a visual means to explain the decision process of the PMPred-AE in the identification of PM. The heatmap generated by Grad-CAM can clearly show the areas that the model focused on when making predictions, thereby enhancing the clarity and interpretability of the model's decisions. Compared with existing work, PMPred-AE had a significant improvement in ACC, Rec, ROC, and F1. This confirmed its leading position in the field of PM-detection and provided strong support for its application in clinical practice.

The PMPred-AE model demonstrates significant potential and scalability in the field of medical image analysis. In addition to effectively detecting PM, PMPred-AE is also applicable to various medical imaging tasks, including the analysis of tumors, brain diseases, and lung diseases. Despite the unique characteristics of different medical images, PMPred-AE offers an efficient and interpretable framework that can be applied across diverse medical scenarios, showcasing substantial clinical application potential. The clinical value of PMPred-AE lies not only in its high accuracy and efficiency but also in its seamless integration with existing healthcare systems. The model can directly process images generated by standard medical devices without requiring additional workflows. Furthermore, PMPred-AE uses Grad-CAM technology to generate heatmaps that visualize the regions the model focuses on, helping physicians make more precise clinical decisions. The model's lightweight design ensures efficient operation even in resource-constrained environments,

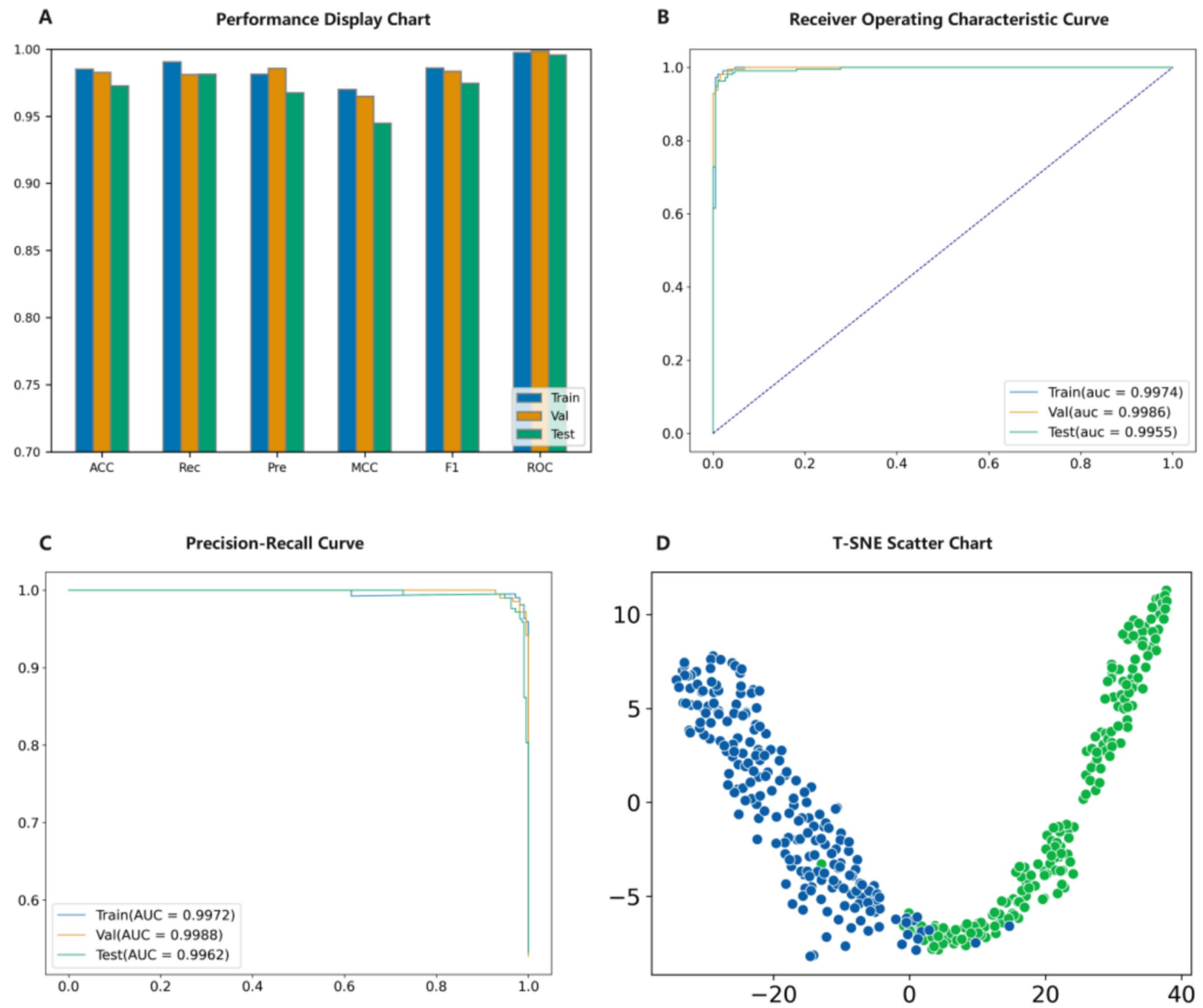


FIGURE 3
Model validation result diagram. **(A)** Evaluation results of the model. **(B)** ROC results of the model. **(C)** PRC results of the model. **(D)** t-SNE visualization the model.

TABLE 1 The performance evaluation of model.

Method	ACC	Pre	Rec	F1	ROC	MCC
Train	0.9850	0.9814	0.9906	0.9860	0.9974	0.9699
Val	0.9825	0.9857	0.9810	0.9834	0.9986	0.9649
Test	0.9725	0.9676	0.9812	0.9744	0.9955	0.9448

making it particularly suitable for regions with limited healthcare resources. However, there are several challenges to be addressed in the deployment of PMPred-AE in practice. First, the quality and diversity of fundus images may vary due to differences in imaging devices and conditions, potentially affecting model performance. To address this, we can enhance the model's generalization ability by expanding the training dataset and incorporating data augmentation techniques. Second, although the model employs an efficient network architecture, inference speed and computational resource requirements could become limiting factors in resource-constrained environments. To mitigate this, we plan to deploy the model on the cloud, leveraging cloud computing resources for inference to reduce the local computational burden. In summary, while the deployment of PMPred-AE faces several challenges, improvements in data quality, optimization of computational resources, and enhanced model robustness can effectively address these issues, ensuring the successful application of the model in clinical practice.

In summary, this research successfully developed an efficient, accurate, and explainable model for the detection of PM by combining advanced model architecture, attention mechanism, and explanatory techniques. This comprehensive method not only improved the performance of the model, but also provided a valuable reference for clinical diagnosis, demonstrating the great potential of deep learning in the field of medical image analysis. In the future, with the

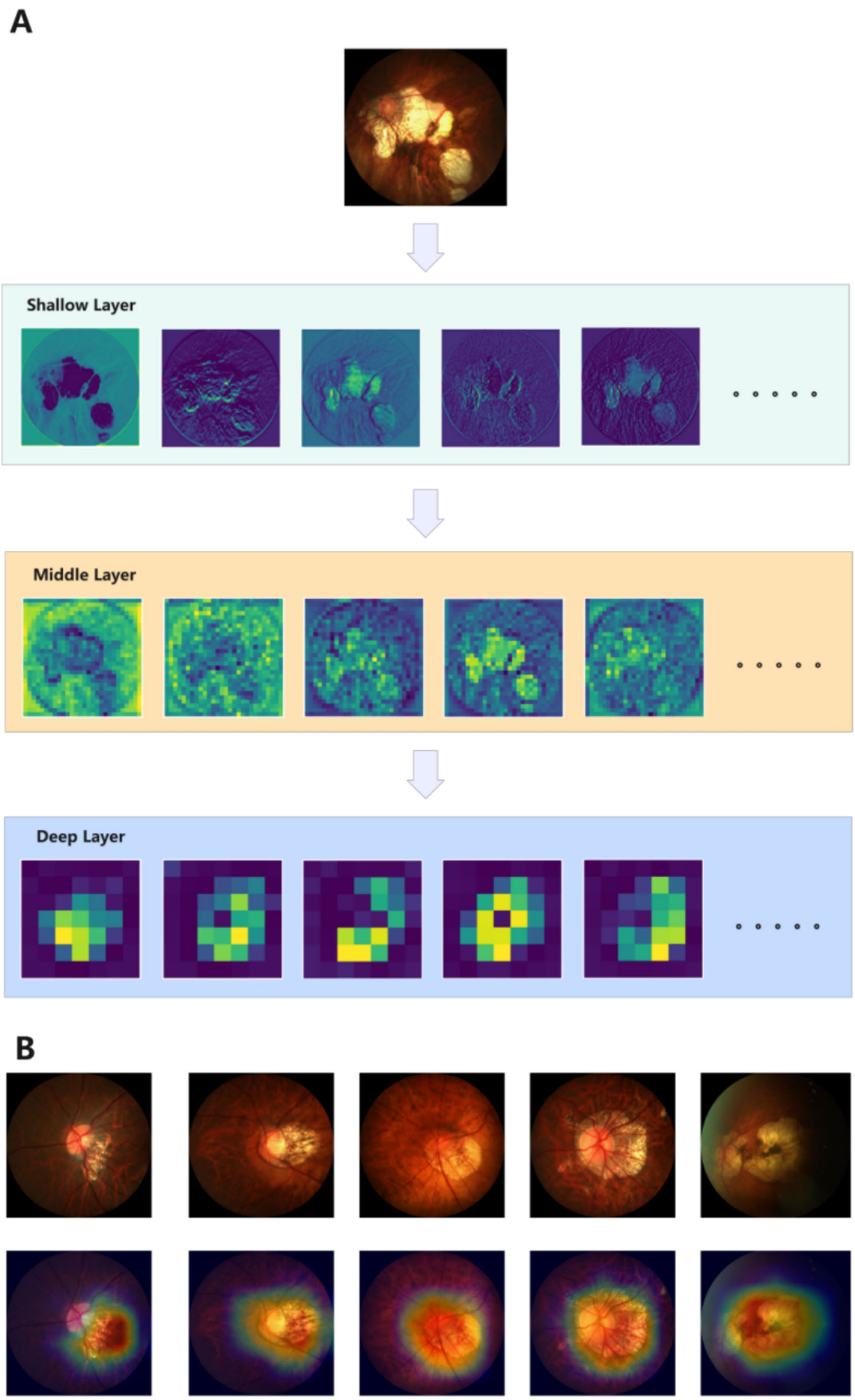


FIGURE 4
Model explanation display diagram. **(A)** Visualize the output results of shallow, middle, and deep layers of the model. **(B)** Visualization results of Grad-CAM.

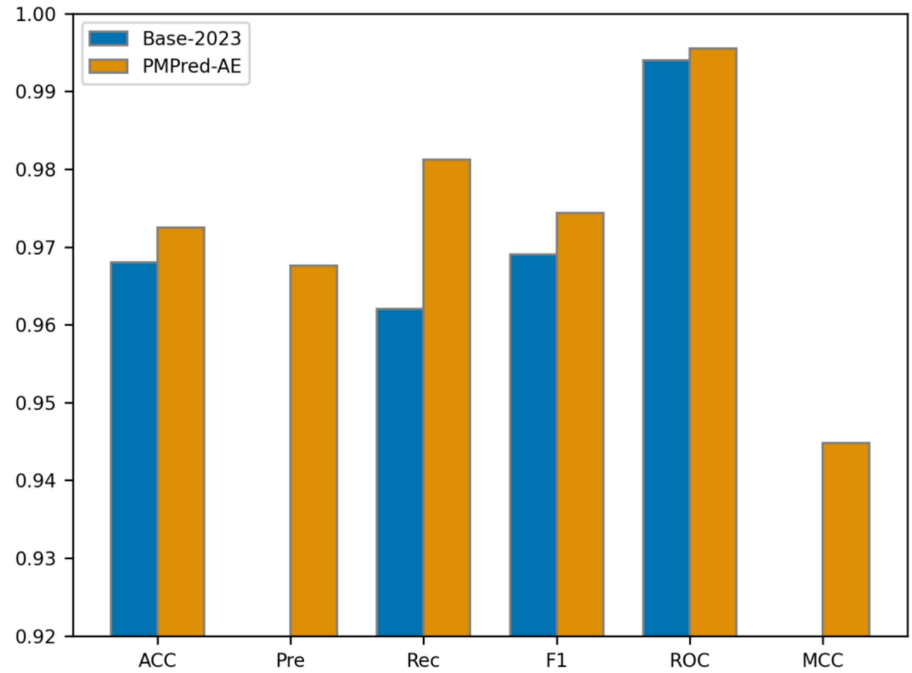


FIGURE 5
Comparison diagram of Base-2023.

TABLE 2 Comparison with published results.

Method	ACC	Pre	Rec	F1	ROC	MCC
Base-2023	0.968	/	0.962	0.969	0.994	/
PMPred-AE	**0.9725**	**0.9676**	**0.9812**	**0.9744**	**0.9955**	**0.9448**

The bold font indicates the classifiers that work best.

continuous advancement of algorithms and technology, such models are expected to play a greater role in improving the efficiency and accuracy of PM diagnosis. The source code has been uploaded to GitHub and can be accessed at: https://github.com/ZhangHongqi215/PMPred-AE.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

H-QZ: Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Writing – original draft. MA: Investigation, Methodology, Validation, Writing – original draft. MT: Validation, Writing – original draft. SA: Methodology, Validation, Writing – original draft. PC: Formal analysis, Investigation, Writing – original draft. YZ: Conceptualization, Funding acquisition, Project administration, Writing – review & editing. HT: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing. HL: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the National Natural Science Foundation of China (62250028, 62172343), the Sichuan Science and Technology Program (Grant No. 2022YFS0614).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Morgan IG, French AN, Ashby RS, Guo X, Ding X, He M, et al. The epidemics of myopia: aetiology and prevention. *Prog Retin Eye Res*. (2018) 62:134–49. doi: 10.1016/j.preteyeres.2017.09.004

2. Hemelings R, Elen B, Blaschko MB, Jacob J, Stalmans I, De Boever P. Pathological myopia classification with simultaneous lesion segmentation using deep learning. *Comput Methods Prog Biomed*. (2021) 199:105920. doi: 10.1016/j.cmpb.2020.105920

3. Saw SM, Gazzard G, Shih-Yen EC, Chua WH. Myopia and associated pathological complications. *Ophthalmic Physiol Opt*. (2005) 25:381–91. doi: 10.1111/j.1475-1313.2005.00298.x

4. Holden BA, Fricke TR, Wilson DA, Jong M, Naidoo KS, Sankaridurg P, et al. Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050. *Ophthalmology*. (2016) 123:1036–42. doi: 10.1016/j.ophtha.2016.01.006

5. Yang J, Ouyang X, Fu H, Hou X, Liu Y, Xie Y, et al. Advances in biomedical study of the myopia-related signaling pathways and mechanisms. *Biomed Pharmacother*. (2022) 145:112472. doi: 10.1016/j.biopha.2021.112472

6. Jonas JB, Jonas RA, Bikbov MM, Wang YX, Panda-Jonas S. Myopia: histology, clinical features, and potential implications for the etiology of axial elongation. *Prog Retin Eye Res*. (2023) 96:101156. doi: 10.1016/j.preteyeres.2022.101156

7. Agyekum S, Chan PP, Zhang Y, Huo Z, Yip BHK, Ip P, et al. Cost-effectiveness analysis of myopia management: a systematic review. *Front Public Health*. (2023) 11:1093836. doi: 10.3389/fpubh.2023.1093836

8. Wei L, He W, Malik A, Su R, Cui L, Manavalan B. Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. *Brief Bioinform*. (2020) 22:bbaa275. doi: 10.1093/bib/bbaa275

9. Liu T, Qiao H, Wang Z, Yang X, Pan X, Yang Y, et al. CodLncScape provides a self-enriching framework for the systematic collection and exploration of coding LncRNAs. *Adv Sci*. (2024) 11:e2400009. doi: 10.1002/advs.202400009

10. Alhatemi RAJ, Savaş S. A weighted ensemble approach with multiple pre-trained deep learning models for classification of stroke. *Medinformatics*. (2023) 1:10–9. doi: 10.47852/bonviewMEDIN32021963

11. Wang Y, Zhai Y., Ding Y., Zou Q. SBSM-Pro: support bio-sequence machine for proteins. *arXiv* [Preprint]. *arXiv:2308.10275* (2023).

12. Wang Y, Zhang W, Yang Y, Sun J, Wang L. Survival prediction of esophageal squamous cell carcinoma based on the prognostic index and sparrow search algorithm-support vector machine. *Curr Bioinforma*. (2023) 18:598–609. doi: 10.2174/1574893618666230419084754

13. Liu B. BioSeq-analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief Bioinform*. (2019) 20:1280–94. doi: 10.1093/bib/bbx165

14. Liu J, Wong DW, Lim JH, Tan NM, Zhang Z, Li H, et al. Detection of pathological myopia by PAMELA with texture-based features through an SVM approach. *J Healthc Eng*. (2010) 1:1–11. doi: 10.1260/2040-2295.1.1.1

15. Zhang Z, Cheng J, Liu J, Sheri YCM, Kong CC, Mei SS, editors. Pathological myopia detection from selective fundus image features. 2012 7th IEEE conference on industrial electronics and applications (ICIEA). IEEE; (2012)

16. Xu Y, Liu J, Zhang Z, Tan NM, Wong DWK, Saw SM, et al., editors. Learn to recognize pathological myopia in fundus images using bag-of-feature and sparse learning approach. 2013 IEEE 10th International Symposium on Biomedical Imaging. IEEE; (2013)

17. Zhang Z, Xu Y, Liu J, Wong DWK, Kwoh CK, Saw S-M, et al. Automatic diagnosis of pathological myopia from heterogeneous biomedical data. *PLoS One*. (2013) 8:e65736. doi: 10.1371/journal.pone.0065736

18. Chen X, Xu Y, Wong DWK, Wong TY, Liu J, editors. Glaucoma detection based on deep convolutional neural network. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; (2015)

19. Luo X, Wang Y, Zou Q, Xu L. Recall DNA methylation levels at low coverage sites using a CNN model in WGBs. *PLoS Comput Biol*. (2023) 19:e1011205. doi: 10.1371/journal.pcbi.1011205

20. Dou LJ, Zhang ZL, Xu L, Zou Q. iKcr_CNN: a novel computational tool for imbalance classification of human nonhistone crotonylation sites based on convolutional neural networks with focal loss. *Comput Struct Biotechnol J*. (2022) 20:3268–79. doi: 10.1016/j.csbj.2022.06.032

21. Xu M, Cheng J, Wong DWK, Cheng C-Y, Saw SM, Wong TY, editors. Automated tessellated fundus detection in color fundus images. Proceedings of the Ophthalmic Medical Image Analysis International Workshop, University of Iowa; (2016)

22. Xu Y, Duan L, Fu H, Zhang Z, Zhao W, You T, et al., editors. Ocular disease detection from multiple informatics domains. 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE; (2018).

23. Septiarini A, Harjoko A, Pulungan R, Ekantini R. Automatic detection of peripapillary atrophy in retinal fundus images using statistical features. *Biomed Signal Process Control*. (2018) 45:151–9. doi: 10.1016/j.bspc.2018.05.028

24. Rauf N, Gilani SO, Waris A. Automatic detection of pathological myopia using machine learning. *Sci Rep*. (2021) 11:16570. doi: 10.1038/s41598-021-95205-1

25. Wei L, Tang J, Zou Q. Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inf Sci*. (2017) 384:135–44. doi: 10.1016/j.ins.2016.06.026

26. Zhang Y, Liu C, Liu M, Liu T, Lin H, Huang C-B, et al. Attention is all You need: utilizing attention in AI-enabled drug discovery. *Brief Bioinform*. (2024) 25:bbad467. doi: 10.1093/bib/bbad467

27. Liu T, Huang J, Luo D, Ren L, Ning L, Huang J, et al. Cm-siRPred: predicting chemically modified SiRNA efficiency based on multi-view learning strategy. *Int J Biol Macromol*. (2024) 264:130638. doi: 10.1016/j.ijbiomac.2024.130638

28. Xu Y, Liu T, Yang Y, Kang J, Ren L, Ding H, et al. Acvpred: enhanced prediction of anti-coronavirus peptides by transfer learning combined with data augmentation. *Futur Gener Comput Syst*. (2024) 160:305–15. doi: 10.1016/j.future.2024.06.008

29. Fang H, Li F, Wu J, Fu H, Sun X, Orlando JI, et al. PALM: open fundus photograph dataset with pathologic myopia recognition and anatomical structure annotation. arXiv [Preprint]. *arXiv:230507816* (2023).

30. Tan M, Le Q, (ed.) EfficientNetv2: smaller models and faster training. International conference on machine learning. PMLR; (2021)

31. Zou X, Ren L, Cai P, Zhang Y, Ding H, Deng K, et al. Accurately identifying hemagglutinin using sequence information and machine learning methods. *Front Med (Lausanne)*. (2023) 10:1281880. doi: 10.3389/fmed.2023.1281880

32. Zulfiqar H, Guo Z, Ahmad RM, Ahmed Z, Cai P, Chen X, et al. Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings. *Front Med*. (2024) 10:10. doi: 10.3389/fmed.2023.1291352

33. Zhu H, Hao H, Yu L. Identifying disease-related microbes based on multi-scale variational graph autoencoder embedding Wasserstein distance. *BMC Biol*. (2023) 21:294. doi: 10.1186/s12915-023-01796-8

34. Liu X, Yang H, Ai C, Ding Y, Guo F, Tang J. Mvml-Mpi: Multi-view multi-label learning for metabolic pathway inference. *Brief Bioinform*. (2023) 24:bbad393. doi: 10.1093/bib/bbad393

35. Liang C, Wang L, Liu L, Zhang H, Guo F. Multi-view unsupervised feature selection with tensor robust principal component analysis and consensus graph learning. *Pattern Recogn*. (2023) 141:109632. doi: 10.1016/j.patcog.2023.109632

36. Li H, Pang Y, Liu B. BioSeq-BLM: a platform for analyzing DNA, RNA, and protein sequences based on biological language models. *Nucleic Acids Res*. (2021) 49:e129. doi: 10.1093/nar/gkab829

37. Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res*. (2019) 47:e127. doi: 10.1093/nar/gkz740

38. Zhang ZY, Zhang Z, Ye X, Sakurai T, Lin H. A Bert-based model for the prediction of lncRNA subcellular localization in *Homo sapiens*. *Int J Biol Macromol*. (2024) 265:130659. doi: 10.1016/j.ijbiomac.2024.130659

39. Dou M, Tang J, Tiwari P, Ding Y, Guo F. Drug-drug interaction relation extraction based on deep learning: a review. *ACM Comput Surv*. (2024) 56:1–33. doi: 10.1145/3645089

40. Charoenkwan P, Schaduangrat N, Lio P, Moni MA, Shoombuatong W, Manavalan B. Computational prediction and interpretation of druggable proteins using a stacked ensemble-learning framework. *iScience*. (2022) 25:104883. doi: 10.1016/j.isci.2022.104883

41. Bupi N, Sangaraju VK, Phan LT, Lal A, Vo TTB, Ho PT, et al. An effective integrated machine learning framework for identifying severity of tomato yellow leaf curl virus and their experimental validation. *Research*. (2023) 6:0016. doi: 10.34133/research.0016

42. Cadrin-Chenevert A. Moving from imagenet to radimagenet for improved transfer learning and generalizability. *Radiol Artif Intell*. (2022) 4:e220126. doi: 10.1148/ryai.220126

43. Li X, Cen M, Xu J, Zhang H, Xu XS. Improving feature extraction from histopathological images through a fine-tuning imagenet model. *J Pathol Inform*. (2022) 13:100115. doi: 10.1016/j.jpi.2022.100115

44. Charoenkwan P, Schaduangrat N, Manavalan B, Shoombuatong W. M3S-ALG: improved and robust prediction of allergenicity of chemical compounds by using a novel multi-step stacking strategy. *Futur Gener Comput Syst*. (2025) 162:107455. doi: 10.1016/j.future.2024.07.033

45. Liu M, Li C, Chen R, Cao D, Zeng X. Geometric deep learning for drug discovery. *Expert Syst Appl*. (2023) 240:122498. doi: 10.1016/j.eswa.2023.122498

46. Zeng X, Wang F, Luo Y, Kang S-g, Tang J, Lightstone FC, et al. Deep generative molecular design reshapes drug discovery. *Cell Rep Med*. (2022) 3:100794. doi: 10.1016/j.xcrm.2022.100794

47. Hasan MM, Tsukiyama S, Cho JY, Kurata H, Alam MA, Liu X, et al. Deepm5C: a deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy. *Mol Ther*. (2022) 30:2856–67. doi: 10.1016/j.ymthe.2022.05.001

48. Shoombuatong W, Meewan I, Mookdarsanit L, Schaduangrat N. Stack-HDAC3i: a high-precision identification of HDAC3 inhibitors by exploiting a stacked ensemble-learning framework. *Methods*. (2024) 230:147–57. doi: 10.1016/j.ymeth.2024.08.003

49. Manavalan B, Lee J. FRTpred: a novel approach for accurate prediction of protein folding rate and type. *Comput Biol Med*. (2022) 149:105911. doi: 10.1016/j.compbiomed.2022.105911

50. Pham NT, Rakkiyapan R, Park J, Malik A, Manavalan B. H_2Opred: a robust and efficient hybrid deep learning model for predicting 2'-O-methylation sites in human RNA. *Brief Bioinform*. (2023) 25:bbad476. doi: 10.1093/bib/bbad476

51. Manavalan B, Patra MC. MLCPP 2.0: an updated cell-penetrating peptides and their uptake efficiency predictor. *J Mol Biol*. (2022) 434:167604. doi: 10.1016/j.jmb.2022.167604

52. Charoenkwan P, Chumnanpuen P, Schaduangrat N, Shoombuatong W. Stack-AVP: a stacked ensemble predictor based on multi-view information for fast and accurate discovery of antiviral peptides. *J Mol Biol*. (2024):168853. doi: 10.1016/j.jmb.2024.168853

53. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. (2008) 9:2579–2605.

frontiers | Frontiers in Pediatrics

TYPE Original Research
PUBLISHED 25 April 2025
DOI 10.3389/fped.2025.1461026

OPEN ACCESS

EDITED BY
Thomas F. Heston,
University of Washington, United States

REVIEWED BY
Masaomi Motegi,
Jikei University School of Medicine, Japan
Dalal Alabdulmohsen,
King Faisal University, Saudi Arabia
Runhan Shi,
Fudan University, China

*CORRESPONDENCE
Dan Xie
✉ xie_dan1980@163.com
Bin Dong
✉ dongbin@scmc.com.cn
Jiajun Yuan
✉ yuanjiajun@scmc.com.cn

†These authors have contributed equally to this work

RECEIVED 10 July 2024
ACCEPTED 07 April 2025
PUBLISHED 25 April 2025

CITATION
Yin Y, Zeng M, Wang H, Yang H, Zhou C, Jiang F, Wu S, Huang T, Yuan S, Lin J, Tang M, Chen J, Dong B, Yuan J and Xie D (2025) A clinician-based comparative study of large language models in answering medical questions: the case of asthma.
Front. Pediatr. 13:1461026.
doi: 10.3389/fped.2025.1461026

A clinician-based comparative study of large language models in answering medical questions: the case of asthma

Yong Yin[1,2,3,4†], Mei Zeng[2†], Hansong Wang[5†], Haibo Yang[1], Caijing Zhou[1], Feng Jiang[1], Shufan Wu[1], Tingyue Huang[1], Shuahua Yuan[2], Jilei Lin[2,3,4], Mingyu Tang[2], Jiande Chen[2], Bin Dong[3,4,6]*, Jiajun Yuan[3,4,7]* and Dan Xie[1]*

[1]Department of Respiratory Medicine, Hainan Branch, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Sanya, China, [2]Department of Respiratory Medicine, Shanghai Children's Medical Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [3]Pediatric AI Clinical Application and Research Center, Shanghai Children's Medical Center, Shanghai, China, [4]Shanghai Engineering Research Center of Intelligence Pediatrics (SERCIP), Shanghai, China, [5]Department of Performance, Shanghai Children's Medical Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [6]Department of Discipline Inspection and Supervision, Shanghai Children's Medical Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [7]Medical Department of Shanghai Children's Medical Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Objective: This study aims to evaluate and compare the performance of four major large language models (GPT-3.5, GPT-4.0, YouChat, and Perplexity) in answering 32 common asthma-related questions.

Materials and methods: Seventy-five clinicians from various tertiary hospitals participated in this study. Each clinician was tasked with evaluating the responses generated by the four large language models (LLMs) to 32 common clinical questions related to pediatric asthma. Based on predefined criteria, participants subjectively assessed the accuracy, correctness, completeness, and practicality of the LLMs' answers. The participants provided precise scores to determine the performance of each language model in answering pediatric asthma-related questions.

Results: GPT-4.0 performed the best across all dimensions, while YouChat performed the worst in all dimensions. Both GPT-3.5 and GPT-4.0 outperformed the other two models, but there was no significant difference in performance between GPT-3.5 and GPT-4.0 or between YouChat and Perplexity.

Conclusion: GPT and other large language models can answer medical questions with a certain degree of completeness and accuracy. However, clinical physicians should critically assess internet information, distinguishing between true and false data, and should not blindly accept the outputs of these models. With advancements in key technologies, LLMs may one day become a safe option for doctors seeking information.

KEYWORDS

large language models (LLMs), pediatric asthma, medical question answering, clinical decision support, AI in healthcare

1 Introduction

Asthma is a major chronic respiratory disease worldwide, affecting the health and quality of life of millions of people. In a multinational, multicenter study involving 453,473 subjects, it was found that 6.3% of children, 7.9% of adolescents, and 3.4% of adults were diagnosed with asthma by a doctor. Moreover, in middle-to-low-income countries, many individuals with severe asthma symptoms were not using inhaled corticosteroids (1). In China, 15.5% of asthma patients reported at least one emergency room visit, and 7.2% of patients reported at least one hospitalization due to worsening respiratory symptoms (2). Despite receiving high-intensity treatment, most children with poorly controlled symptoms can achieve improved asthma control when they adhere to the basic principles of asthma management (3). Frequent and severe asthma attacks can be fatal, and effective asthma management and treatment require close cooperation between patients, doctors, and caregivers. Therefore, improving the provision of accurate health information and personalized counseling is crucial for the self-management of asthma patients.

A survey of online health behaviors of Americans revealed that more than one-third of Americans turn to the Internet to diagnose health problems (4). Large Language Models (LLMs), such as GPT, are AI tools designed to process and generate text. They have been widely applied to various tasks and have demonstrated excellent performance in the medical field (5). LLMs will increasingly be used for information retrieval, automated summarization of literature notes, answering medical questions, and even as interactive tools in medical education (6, 7). This not only helps patients access important disease-related information more quickly but also supports the decision-making process of healthcare professional (8). However, Information errors, privacy issues, and ethical challenges and potential harm to patient care remain significant challenges (9). Ethical issues, including data privacy and breaches, must be addressed. In both medical and non-medical education, students are vulnerable to misinformation, hindering the development of critical thinking skills. The lack of mechanisms to ensure the accuracy of LLM outputs limits their use in clinical settings, where misinformation can have fatal consequences (7).

In this study, we aim to evaluate and compare the performance of four selected Large Language Models (GPT-3.5, GPT-4.0, YouChat, and Perplexity) in answering clinical questions related to pediatric asthma. The evaluation includes four dimensions: accuracy, precision, completeness, and practicality, combined with insights from professionals for a comprehensive assessment. Our findings may provide valuable insights into the clinical application of LLMs as medical auxiliary tools and promote clinical decision-making.

2 Article type

This study is an Original Research Article that evaluates and compares the performance of four major large language models (GPT-3.5, GPT-4.0, YouChat, and Perplexity) in answering 32 common asthma-related questions.

3 Material and methods

3.1 Model selection

Based on previous research, user volume, and training methodologies, this study selected four models for investigation: ChatGPT 3.5, ChatGPT 4.0, YouChat, and Perplexity. ChatGPT 3.5 and ChatGPT 4.0 were trained on predefined datasets and did not connect to the internet after their launch. ChatGPT 4.0 utilizes a more extensive and diverse pre-training dataset compared to ChatGPT 3.5, along with advanced training techniques such as more effective model optimization algorithms and smarter parameter initialization methods. The version of YouChat used in this study is the basic version, which extends ChatGPT 3.5 by integrating an internet search function. Similarly, the Perplexity version used is the basic one, functioning as an AI-powered search engine that combines proprietary language models with real-time web retrieval to generate responses.

3.2 Question selection and answering with large language models

The equations should be inserted in editable format from the equation editor. We selected 32 common asthma-related questions from the article “One hundred key issues on Chinese Children’s Asthma Action Plan” published in the Chinese Journal of Practical Pediatrics to test the model (10). On the one hand, these questions were selected after consultation with three pediatric respiratory asthma experts and reflected the main aspects of asthma management, such as diagnosis, treatment, prevention and follow-up. On the other hand, the selection process was designed to cover essential topics related to the concerns of clinicians and patients and their families in the clinical setting. All questions were posed and recorded in Chinese, and we translated them into English for presentation (see Table 1). The prompt for all models was set as: “Assume you are an expert in the field of pediatrics, and the following questions are all related to pediatrics. Please answer the following questions in less than 500 words.” The questions were inputted in the exact same order and content for all models. To ensure consistency and eliminate potential influence on clinician ratings, we manually removed all hyperlinks, quotation marks, and web-related formatting from all model responses. All answers were presented in a uniform plain text format and model identities were anonymized. This standardization ensured that assessments were based solely on the accuracy, correctness, completeness, and utility of the content, and not on the presence or absence of supporting links or reference formats. To evaluate the internal stability of the models, we created five dialogues using the same input method. The project team members jointly assessed the stability of the five responses, and the results were recorded on a ten-point scale, with a minimum of 1 and a maximum of 10.

TABLE 1 Questions used to test the performance of LLMs.

32 Questions Related to Childhood Asthma	
Question 1	What is asthma?
Question 2	Is asthma hereditary?
Question 3	What are the differences and similarities between asthma in children and adult asthma?
Question 4	What are the clinical features of asthma in children?
Question 5	How is bronchial asthma in children diagnosed?
Question 6	Can recurrent wheezing in infancy develop into asthma?
Question 7	What are the comorbid conditions of asthma?
Question 8	What impact does allergic rhinitis (AR) have on asthma?
Question 9	What are the common tests for childhood asthma?
Question 10	Can childhood asthma be cured?
Question 11	Does long-term ICS treatment affect the growth and development of children?
Question 12	Which children with asthma are eligible for allergen specific immune therapy (AIT)?
Question 13	Which children with asthma are eligible for biological treatments such as monoclonal antibodies?
Question 14	Why is it important to manage asthma in children?
Question 15	Why should children with asthma have regular follow-up visits to the hospital? How often should these visits occur?
Question 16	What are the main components of follow-up visits for children with asthma?
Question 17	What are the early preventive measures for asthma?
Question 18	What are common allergens? Why do children with asthma need allergen testing?
Question 19	What are dust mites? How can dust mite allergies be prevented?
Question 20	Which pet dander is likely to cause allergies?
Question 21	How can pollen allergies be managed?
Question 22	Can children with asthma receive vaccinations?
Question 23	What is the relationship between asthma attacks and upper respiratory infections?
Question 24	Can children with asthma exercise? How should they exercise?
Question 25	Can exercise induce asthma attacks? How can exercise-induced asthma attacks be prevented?
Question 26	What climate changes are likely to trigger asthma attacks? How can these be prevented?
Question 27	What are the adverse effects of cigarette smoke exposure on children with asthma? How can this be prevented?
Question 28	What factors are likely to cause acute asthma attacks during outdoor activities or travel?
Question 29	What signs can predict an acute attack of asthma in children?
Question 30	How can the severity of an acute asthma attack in children be assessed?
Question 31	How can severe acute asthma attacks be prevented?
Question 32	What emergency medications should be readily available at home or nearby for children with asthma?

3.3 Model evaluation dimensions

This study designed the questionnaire from the perspective of doctors. The questionnaire evaluates the responses of different models based on four dimensions: "accuracy," "correctness," "completeness," and "practicality." "Accuracy" is defined as the degree to which the model's answer is relevant to the question, reflecting the model's ability to understand the user's query. "Correctness" refers to the extent to which the model's answer aligns with the clinical experience and guidelines of the respondents. "Completeness" is defined as the thoroughness of the model's answer compared to clinical experience and guidelines. "Practicality" refers to the extent to which the model's answer is applicable in daily clinical practice, reflecting the model's ability to solve real-world problems. The results are recorded on a ten-point scale, with "unable to answer" responses scored as 0 and other answers scored between 1 and 10. The definitions of the four evaluation dimensions are placed on the first page of the questionnaire to clearly inform the respondents and facilitate accurate evaluation.

3.4 Questionnaire design

Each questionnaire contained thirty-two questions, arranged in the same order, with answers generated by different large language models. Participants were instructed to provide clear and unambiguous answers based on existing clinical guidelines. The four model-generated answers for each question were presented in random order, and participants were not informed which model corresponded to each answer. To improve the quality of questionnaire completion, we set a time limit for answering the questions. The questionnaires were then distributed in paper form to 75 clinicians and collected uniformly. This study was conducted from January to May 2024.

3.5 Participant inclusion

The evaluators in this study met the following criteria: (1) Hold a Master's degree in medicine or higher; (2) be under 60 years of age; (3) Have worked in the pediatric department of a tertiary hospital.

3.6 Questionnaire quality control

We implemented quality control for the questionnaires based on the following criteria: (1) Assigning a high score to responses with obvious errors/deficiencies was considered one quality control anomaly; (2) Completing the questionnaire in less than 2 h was counted as one quality control anomaly; (3) Having three responses with clearly outlier scores was counted as one quality control anomaly. If there were fewer than three such scores, it was counted as three instances. A sample was deemed to have failed quality control if it exhibited five instances of quality control anomalies. Only samples that passed quality control were included in the analysis.

3.7 Inter-rater reliability analysis

To assess the consistency of raters in rating different models, we conducted an inter-rater reliability analysis using the Intraclass Correlation Coefficient (ICC). The ICC is a widely used metric to measure the level of agreement between raters when rating continuous data. In this study, ICC values were calculated for four rating aspects—accuracy, completeness, correctness and practicality—using different models. Higher ICC values indicate better agreement between raters. The final results are shown in Supplementary Figure S1, where it can be observed that Perplexity and YouChat provided the most consistent ratings, with ICC values ranging from 0.85–0.91 across all aspects, indicating a high level of inter-rater agreement. In contrast, GPT-4.0 showed the greatest variation in raters' scores, particularly for Correctness and Practicality.

3.8 Statistical analysis

All data analysis was conducted using R 4.3.3. To comprehensively understand the responses of the four major language models to asthma-related clinical questions, we calculated the average score for each question answered by each model and presented the results through bar charts. Next, we calculated the average score for each model across all evaluative dimensions per question to examine the distinct responses provided by each model. Sankey diagrams were used to describe the commonalities and differences in cumulative scores for the top five and bottom five questions among the four models. To assess differences between the models, we first determined the average score for each question across different models and then performed hypothesis testing using Tukey's *post hoc* test. We then used Tukey's *post hoc* test to compare the performance of the four models across various dimensions. Finally, we utilized Tukey's *post hoc* test to evaluate the significance of differences within each model across different dimensions.

4 Results

4.1 Questionnaire distribution and recall

The research distributed a total of 75 questionnaires, all of which were returned and passed quality control, yielding a qualification rate of 100%.

4.2 Evaluation of LLMs' performance

Table 1 lists all the questions included in the 32 questionnaires. Figure 1 shows the flowchart of the study. Figure 2 shows the responses of the large language models (LLMs) to all questions. In the questionnaires, the median score for all questions answered by the LLMs was 7.9, with the highest scores for questions 26, 14, 2, 16, and 18, and the lowest scores for questions 6, 12, 22, 13, and 4. This indicates that the LLMs performed excellently in addressing the genetic causes, management strategies, and prevention of childhood asthma, but showed some weaknesses in addressing the clinical characteristics, early diagnosis, and specific treatments (such as allergen-specific immunotherapy and monoclonal antibody treatments) for childhood asthma.

Figure 3 displays the scores of different models on each question. ChatGPT 3.5 and ChatGPT 4.0 had higher median scores, both at 8.1, while Perplexity and YouChat had lower median scores, at 7.7 and 7.6, respectively.

Figure 4 illustrates the differences and similarities between the top five and bottom five questions answered by the various models. Our findings indicate that multiple models demonstrated proficiency in answering questions 2, 14, 18, and 26, suggesting that LLMs are more adept at addressing questions related to genetic causes, management measures, and the prevention of childhood asthma. The GPT 4.0 demonstrated particular proficiency in responding to the questions with the highest scores. However, in the case of the questions with the lowest scores, multiple models exhibited less impressive performance on questions 6, 12, 22, and 32. This indicates that the LLMs (even with GPT 4.0) were less adept at answering questions pertaining to early identification and prevention of childhood asthma, personalized treatment, prevention, and emergency care management.

4.3 Comparison in different dimensions of each model

Figure 5 illustrates the average scores of different models across all questions. GPT 3.5 and GPT 4.0 significantly outperformed Perplexity and You Chat, exhibiting more stable and higher scores. There was no significant difference in performance between GPT 3.5 and GPT 4.0, with their median scores being nearly identical. Similarly, there was no significant difference between Perplexity and You Chat, with their median scores being close to each other.

Figure 6 shows that the GPT-4.0 performed better on all four assessment dimensions, although statistical analyses showed no significant difference between the GPT-4.0 and GPT-3.5. Conversely, YouChat had the lowest performance in all aspects, putting it at a disadvantage compared to the other three models.

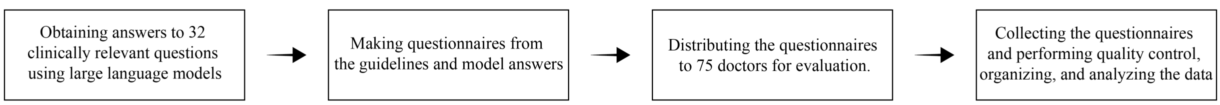


FIGURE 1
Flowchart.

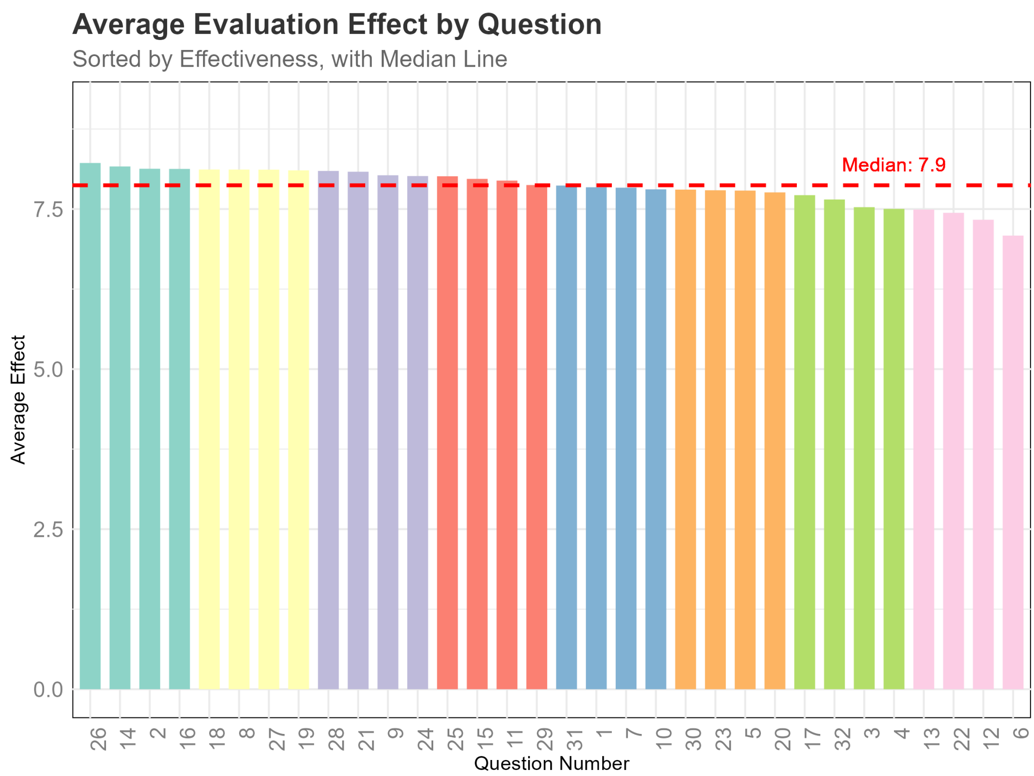


FIGURE 2
The average score of each question for all models.

5 Discussions

Artificial intelligence is increasingly being applied in various medical projects, including radiological image analysis (11), aiding diagnosis in complex cases (12), personalized treatment (13), anesthesia depth monitoring and control (14), and drug development and utilization (15). A study evaluated ChatGPT's performance on the United States Medical Licensing Examination (USMLE), and the results showed that ChatGPT met or nearly met the passing threshold without any specialized training or reinforcement (16). The LLM demonstrated strong performance in making final diagnoses across 36 clinical cases, achieving an accuracy rate of 76.9% (17). Importantly, compared to other decision support tools, LLMs not only incorporate more patient-specific information to generate more targeted recommendations but also encourage brainstorming, prompting doctors to consider diagnoses and treatments they might otherwise overlook. These results suggest that large language models may have the potential to aid in medical education and assist in clinical decision-making.

In this study, all the major language models performed well in answering a range of clinically relevant questions, with particular excellence in the areas of asthma causes, treatment and prevention. This is probably because these topics are of greater public interest and there are more sources of information available, resulting in more training data and consequently higher scores. For asthma diagnosis and new treatments, the LLMs showed less stable performance, indicating a need for more recent data training in these areas.

GPT and other large language models can answer medical questions with a certain degree of completeness and accuracy. Our results indicate that while GPT-4.0 demonstrated the highest scores across all dimensions, the statistical analysis revealed no significant difference between GPT-4.0 and GPT-3.5. This suggests that both models perform comparably in medical question answering, and the choice between them may depend on factors beyond numerical scores. Despite this, we still recommend GPT-4.0 due to its qualitative advantages over GPT-3.5, including a larger database, more advanced training data, improved model architecture, and better integration with clinical guidelines. These factors enable GPT-4.0 to understand and generate more accurate and effective information. Additionally, qualitative feedback from clinicians suggests that GPT-4.0 provides smoother and more contextually relevant responses, making it more reliable in real-world medical scenarios. In the top five questions (Question 5: Is asthma hereditary? Question 8: What is the impact of allergic rhinitis (AR) on asthma? Question 25: Can exercise induce asthma attacks? How to prevent exercise-induced asthma attacks? Question 26: What climate changes can trigger asthma attacks? How to prevent them? Question 27: What adverse effects does cigarette exposure have on children with asthma? How to prevent them?), GPT-4.0 did an excellent job of answering questions about asthma heredity, triggers, and preventive measures. However, GPT-4.0 showed weaker capabilities in handling questions related to asthma management

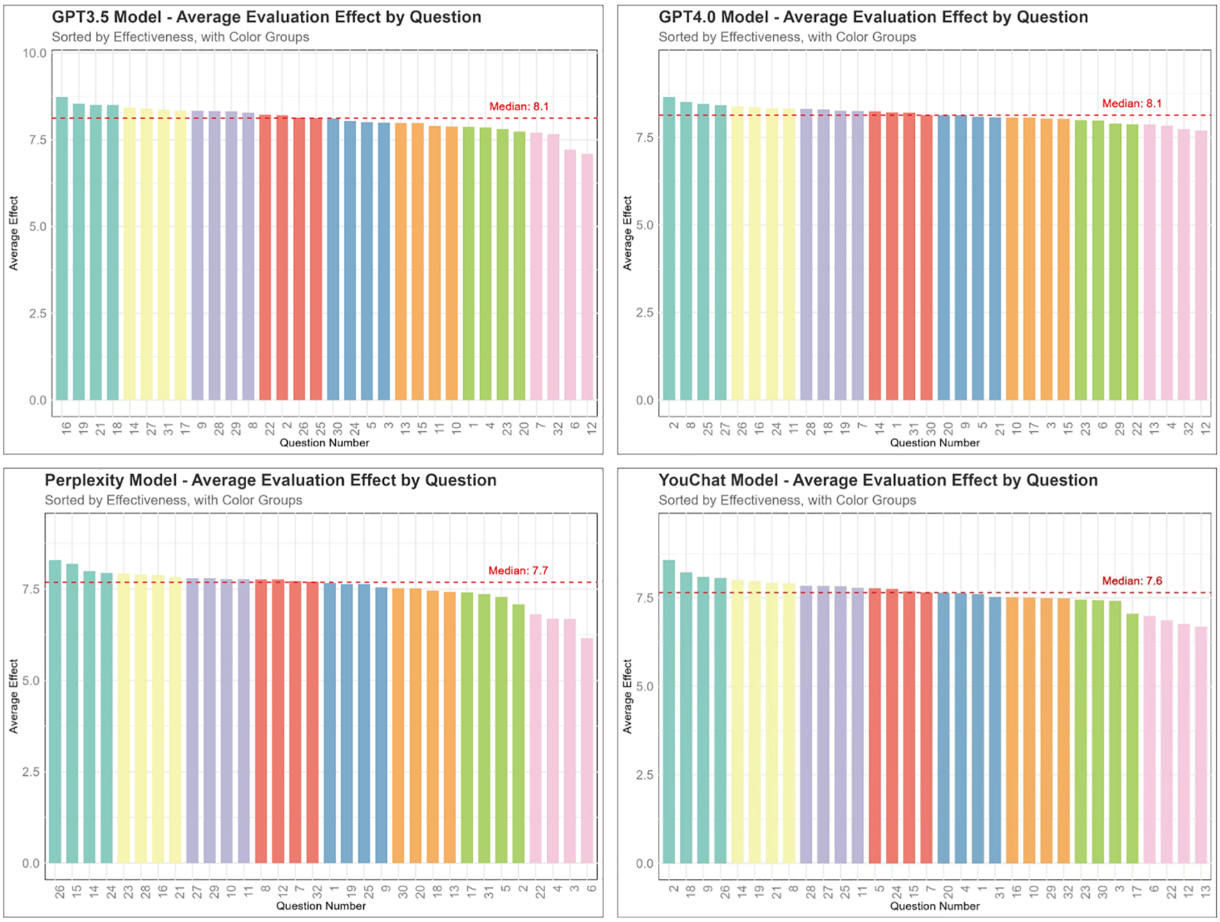


FIGURE 3
The average score of each question for different models.

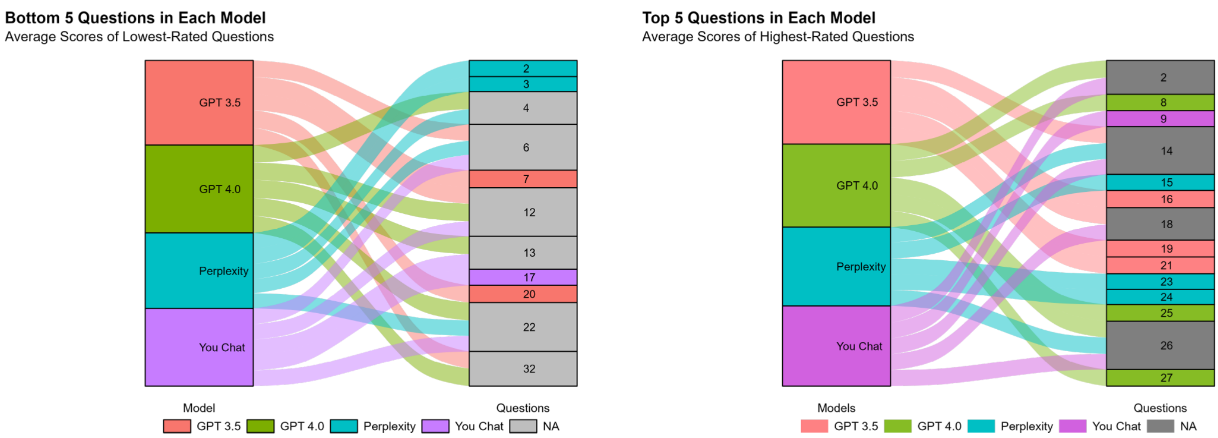


FIGURE 4
Sankey diagram of the questionnaire.

(vaccination) and treatment strategies (including emergency, immunotherapy, or biologic treatments). For some new asthma treatments, such as desensitization therapy and monoclonal antibody therapies, future model training should emphasize updating the database in these areas. If LLMs could be trained by reliable experts, it could rapidly improve and transform the dissemination of medical knowledge. Providing more and more disease information through LLMs could help address the growing prevalence of asthma.

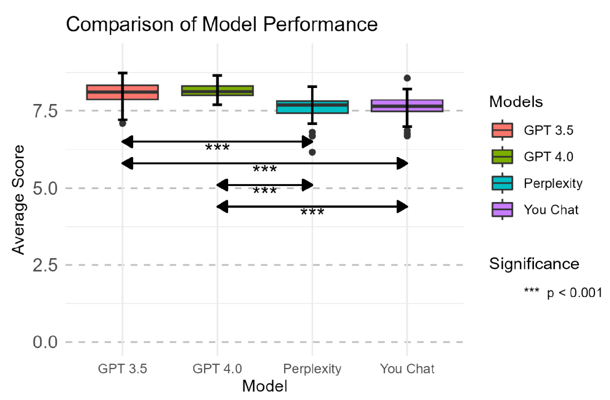


FIGURE 5
Comparison of average scores across all dimensions between models.

Although YouChat performed the worst of all models, it significantly outperformed the other three models in answering questions about accurately diagnosing asthma (Question 9: What are common tests for childhood asthma?) and identifying and managing allergens (Question 18: What are common allergens? Why do children with asthma need allergen testing?). These interventions are complementary and form a systematic approach to comprehensive asthma management, demonstrating that each model has strengths in different aspects of disease management.

However, there are several limitations to this study. First, the sample size is relatively small (75 doctors), which may affect the generalizability of the results. Second, there may be biases in the questionnaire design, as the selection and phrasing of questions could influence the models' responses. Additionally, this study focuses solely on pediatric asthma questions, different medical domains might yield different results. Future research could expand the sample size and diversity of questions to improve the generalizability and reliability of the findings. It may also consider evaluating the models' performance in various medical fields (e.g., hypertension,

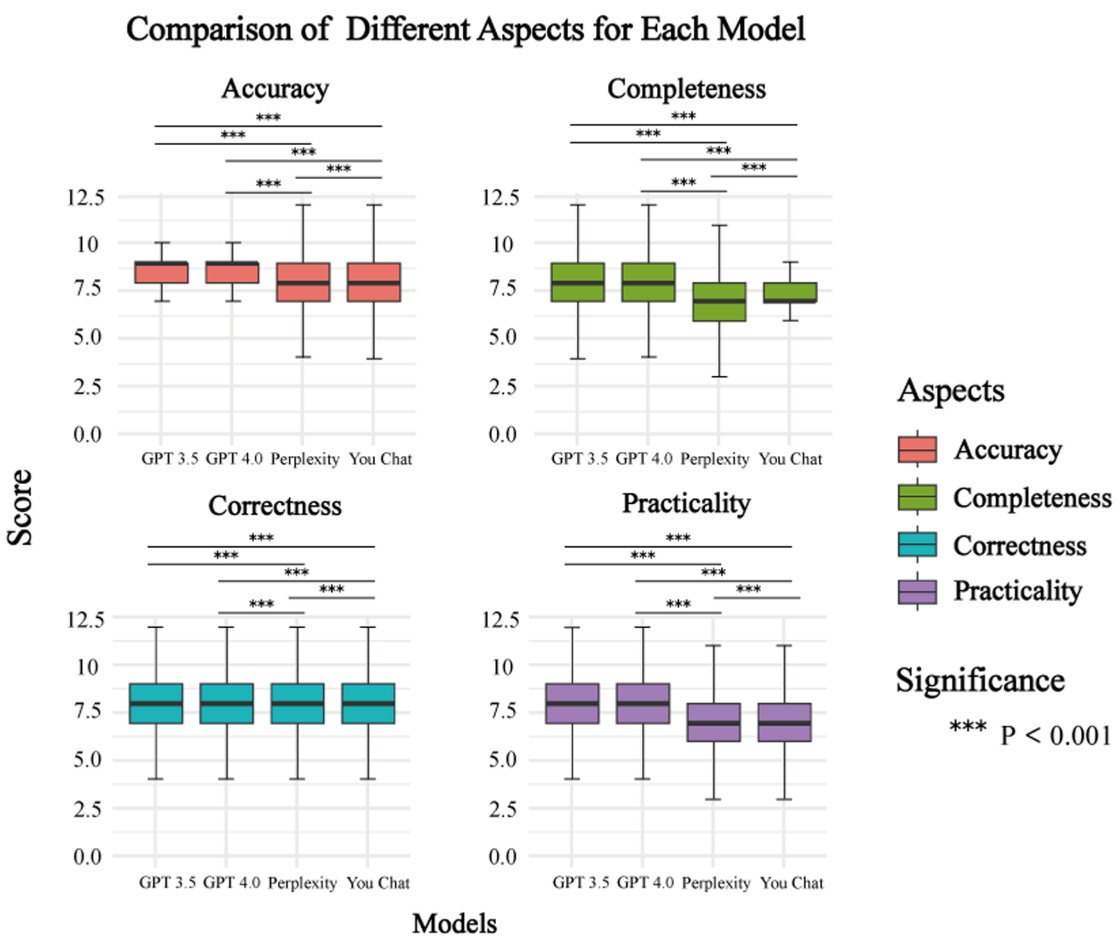


FIGURE 6
Comparison of average scores across different dimensions between models.

diabetes) to gain a comprehensive understanding of their potential applications in medicine. Furthermore, research could explore ways to further enhance the training data and model architecture to improve their performance in specialized fields. Although the models performed well in this study, in practice, LLMs may give incorrect responses when faced with prompts that do not have a single correct answer, and if they present these responses in a convincing manner, users might believe their accuracy (18). Therefore, in practical use, doctors should use LLMs as supplementary and enhanced support rather than relying solely on their responses (19).

While our findings suggest that large language models (LLMs) such as GPT-4.0 have great potential as tools for clinical decision support, it is important to recognize the ethical risks and challenges they pose—particularly the risk of misinformation. For example, if an LLM suggests the use of an outdated or contraindicated asthma medication without considering the clinical context, this could lead to harmful outcomes-especially if the recommendation is followed without expert review. From an ethical perspective, the use of LLMs also raises questions about responsibility and accountability. Unlike human clinicians, LLMs do not have intent, awareness, or professional responsibility, making it difficult to determine who is liable if AI-generated content causes harm. In addition, LLMs responses may reflect biases in their training data or generate information that sounds accurate but not to. To mitigate these risks, several strategies should be implemented: (1) Human oversight: All LLM-generated content should be reviewed by qualified healthcare professionals before being used in clinical practice. (2) Transparency and interpretability: Developers should improve how LLMs explain their answers and ensure that the system can flag low-confidence or uncertain answers. (3) User training: Clinicians and other users should be trained to understand the limitations of LLMs and to use their results critically. (4) Ongoing monitoring: The performance of LLMs should be regularly reviewed in real-world settings to ensure continued safety and accuracy.

Based on the above, doctors still need to receive proper education and continuously update their knowledge through various traditional evidence-based educational methods. It is crucial to apply critical thinking to the information provided by LLMs and regard it as a supplement to their clinical knowledge and experience. Otherwise, clinicians can be easily misled. Currently, whether in terms of data or training, large language models do not seem capable of replacing the unique intellectual abilities of humans. Clinicians need to be very vigilant and apply all evaluative and critical measures to the information provided before establishing such tools as support for clinical decision-making. In the future, with advancements in key technologies and the resolution of diagnostic blind spots and data privacy issues, large language models have the potential to become important tools for improving human healthcare.

6 Conclusion

GPT and other large language models can answer medical questions with a certain degree of completeness and accuracy. However, clinical physicians should critically assess internet information, distinguishing between true and false data, and should not blindly accept the outputs of these models. With advancements in key technologies, LLMs may one day become a safe option for doctors seeking information.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

YY: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal analysis, Software, Visualization. MZ: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal analysis, Software, Visualization. HW: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal analysis, Software, Visualization. HY: Investigation, Writing – review & editing, Data curation, Formal analysis, Supervision. CZ: Investigation, Writing – review & editing, Data curation, Formal analysis, Software. FJ: Investigation, Supervision, Writing – review & editing, Data curation, Formal analysis. SW: Investigation, Supervision, Writing – review & editing, Data curation, Formal analysis. TH: Data curation, Formal analysis, Investigation, Supervision, Writing – review & editing. SY: Data curation, Formal analysis, Investigation, Supervision, Writing – review & editing. JL: Data curation, Formal analysis, Investigation, Supervision, Writing – review & editing. MT: Data curation, Formal analysis, Investigation, Supervision, Writing – review & editing. JC: Data curation, Formal analysis, Investigation, Supervision, Writing – review & editing. BD: Conceptualization, Data curation, Validation, Supervision, Writing – review & editing. JY: Data curation, Validation, Conceptualization, Supervision, Writing – review & editing. DX: Conceptualization, Data curation, Validation, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by Fundamental Research Funds for the Central Universities (YG2024ZD20), Fundamental Research Funds for the Central Universities (YG2024QNB25), Fundamental Research Funds for the Central Universities (YG2022QN095), Shanghai Municipal Health Commission Healthcare Industry Clinical Research Special Project (20224Y0180); Shanghai Municipal Science and Technology Commission 'Yangfan Plan' (23YF1425100); Pudong New Area Science, Technology Development Fund Livelihood Research Special Fund Healthcare Project (PKJ2023-Y50), Shanghai Pudong New Area Science and Technology Development Fund Livelihood Research Special Fund Healthcare Project (PKJ2023-Y49) and Three-year action plan for strengthening the construction of the public health system in Shanghai (GWVI- 11.2- YQ58).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fped.2025.1461026/full#supplementary-material

References

1. García-Marcos L, Chiang C-Y, Asher MI, Marks GB, El Sony A, Masekela R, et al. Asthma management and control in children, adolescents, and adults in 25 countries: a global asthma network phase I cross-sectional study. *Lancet Glob Health*. (2023) 11(2):e218–e28. doi: 10.1016/S2214-109X(22)00506-X

2. Huang K, Yang T, Xu J, Yang L, Zhao J, Zhang X, et al. Prevalence, risk factors, and management of asthma in China: a national cross-sectional study. *Lancet*. (2019) 394(10196):407–18. doi: 10.1016/S0140-6736(19)31147-X

3. Pike KC, Levy ML, Moreiras J, Fleming L. Managing problematic severe asthma: beyond the guidelines. *Arch Dis Child*. (2018) 103(4):392–7. doi: 10.1136/archdischild-2016-311368

4. Kuehn BM. More than one-third of US individuals use the internet to self-diagnose. *JAMA*. (2013) 309(8):756–7. doi: 10.1001/jama.2013.629

5. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. *Lancet*. (2020) 395(10236):1579–86. doi: 10.1016/S0140-6736(20)30226-9

6. Tian S, Jin Q, Yeganova L, Lai P-T, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief Bioinform*. (2023) 25(1):bbad493. doi: 10.1093/bib/bbad493

7. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ*. (2024) 58(11):1276–85. doi: 10.1111/medu.15402

8. Liu S, McCoy AB, Wright AP, Carew B, Genkins JZ, Huang SS, et al. Leveraging large language models for generating responses to patient messages-a subjective analysis. *J Am Med Inform Assoc*. (2024) 31(6):1367–79. doi: 10.1093/jamia/ocae052

9. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J-N, Laleh NG, et al. The future landscape of large language models in medicine. *Commun Med (Lond)*. (2023) 3(1):141. doi: 10.1038/s43856-023-00370-1

10. Diseases CNCRCfR, Cooperative Group of Asthma tSGoR, the Society of Pediatrics, Chinese Medical Association, Pediatrics CMEACo, Pediatrics CMDACoR, Pediatrics CRHACo, Pediatrics CN-GMIACo, et al. One hundred key issues on Chinese children's asthma action plan. *Chin J Pract Pediatr*. (2021) 36(7):491–513.

11. Savadjiev P, Chong J, Dohan A, Vakalopoulou M, Reinhold C, Paragios N, et al. Demystification of AI-driven medical image interpretation: past, present and future. *Eur Radiol*. (2019) 29(3):1616–24. doi: 10.1007/s00330-018-5674-x

12. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*. (2023) 330(1):78–80. doi: 10.1001/jama.2023.8288

13. Bhinder B, Gilvary C, Madhukar NS, Elemento O. Artificial intelligence in cancer research and precision medicine. *Cancer Discov*. (2021) 11(4):900–15.

14. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. *Anesthesiology*. (2020) 132(2):379–94. doi: 10.1097/ALN.0000000000002960

15. Liu Z, Roberts RA, Lal-Nag M, Chen X, Huang R, Tong W. AI-based language models powering drug discovery and development. *Drug Discov Today*. (2021) 26(11):2593–607. doi: 10.1016/j.drudis.2021.06.009

16. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. (2023) 2(2):e0000198.

17. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res*. (2023) 25:e48659. doi: 10.2196/48659

18. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med*. (2023) 388(13):1233–9. doi: 10.1056/NEJMsr2214184

19. Mello MM, Guha N. ChatGPT and physicians' malpractice risk. *JAMA Health Forum*. (2023) 4(5):e231938. doi: 10.1001/jamahealthforum.2023.1938

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 12 May 2025
DOI 10.3389/fmed.2025.1512870

Check for updates

OPEN ACCESS

EDITED BY
Thomas F. Heston,
University of Washington, United States

REVIEWED BY
Wojciech Lesiński,
University of Białystok, Poland
Zuheng Wang,
Guangxi Medical University, China
Bhumandeep Kour,
Lovely Professional University, India

*CORRESPONDENCE
Duxian Liu
✉ ldx849756917@qq.com

RECEIVED 17 October 2024
ACCEPTED 21 April 2025
PUBLISHED 12 May 2025

CITATION
Peng C, Gong C, Zhang X and Liu D (2025) A prognostic model for highly aggressive prostate cancer using interpretable machine learning techniques.
Front. Med. 12:1512870.
doi: 10.3389/fmed.2025.1512870

A prognostic model for highly aggressive prostate cancer using interpretable machine learning techniques

Cong Peng, Cheng Gong, Xiaoya Zhang and Duxian Liu*

Department of Pathology, The Second Hospital of Nanjing, Affiliated to Nanjing University of Chinese Medicine, Nanjing, Jiangsu, China

Background: Extremely aggressive prostate cancer, including subtypes like small cell carcinoma and neuroendocrine carcinoma, is associated with poor prognosis and limited treatment options. This study sought to create a robust, interpretable machine learning-based model that predicts 1-, 3-, and 5-year survival in patients with extremely aggressive prostate cancer. Additionally, we sought to pinpoint key prognostic factors and their clinical implications through an innovative method.

Materials and methods: This study retrospectively analyzed data from 1,620 patients with extremely aggressive prostate cancer in the SEER database (2000–2020). Feature selection was performed using the Boruta algorithm, and survival predictions were made using nine machine learning algorithms, including XGBoost, logistic regression (LR), support vector machine (SVM), random forest (RF), k-nearest neighbor (KNN), decision tree (DT), elastic network (Enet), multilayer perceptron (MLP) and lightGBM. Model performance was evaluated using metrics such as AUC, accuracy (F1 score), confusion matrix, and decision curve analysis. Additionally, Shapley Additive Explanations (SHAP) were applied to interpret feature importance within the model, revealing the clinical factors that influence survival predictions.

Results: Among the nine models, the lightGBM model exhibited the best performance, with an AUC and F1 score of (0.8, 0.809) for 1-year survival prediction, (0.809, 0.751) for 3-year survival prediction, and (0.773, 0.611) for 5-year survival prediction. SHAP analysis revealed that M stage was the most important feature for predicting 1- and 3-year survival, while PSA level had the greatest impact on 5-year survival predictions. The model demonstrated good clinical utility and predictive accuracy through decision curve analysis and confusion matrix.

Conclusion: The lightGBM model has good predictive power for survival in patients with extremely aggressive prostate cancer. By identifying key clinical factors and providing actionable predictions, the model has the potential to enhance prognostic accuracy and improve patient outcomes.

KEYWORDS

prostate cancer, survival, machine, predictive analytics, Boruta algorithm

Introduction

According to the Cancer Statistics 2024 published by the American Cancer Society, the United States is expected to diagnose approximately 299,010 new cases of prostate cancer in 2024, accounting for 14.9 percent of all new cancer cases. In addition, about 35,250 men are expected to die from prostate cancer in 2024, making it the second leading cause of cancer deaths among men in the United States (1). Extremely aggressive prostate adenocarcinoma, a rare subtype of prostate cancer, represents 5 to 10% of all prostate cancer cases (2). This category includes subtypes such as small cell carcinoma, squamous cell carcinoma, and neuroendocrine carcinoma, which are associated with higher metastatic rates and a worse prognosis (3, 4). In contrast to typical prostate adenocarcinomas, these aggressive forms are often resistant to standard hormonal therapies and present with widespread metastases at the time of diagnosis, leading to significantly reduced survival times (5, 6). Once metastasis occurs, the median survival for these patients is typically reported to be less than one year, and current treatment options show limited effectiveness (7, 8).

In recent years, machine learning, a burgeoning tool within the realm of artificial intelligence, has found extensive application in the medical field (9–11). By leveraging large-scale clinical datasets, machine learning can automatically detect and learn complex patterns, thereby enhancing the accuracy of disease prognosis predictions (9, 12). The latest review highlights how machine learning models are redefining the diagnosis and management of prostate cancer (13, 14).

Several previous studies have focused on developing machine-learning-based risk prediction models for prostate cancer. For example, Changhee et al. used machine learning to predict cancer-specific mortality in patients with non-metastatic prostate cancer. While Peng et al. developed a machine-learning-based prognostic model for patients with lymph node-positive prostate cancer. However, there is a lack of clinical tools for prognostic assessment of extremely aggressive prostate cancer patients with poor prognosis. Although traditional statistical models can provide some prognostic prediction, their ability to mine high-dimensional nonlinear data is limited and cannot fully reveal the relationship between complex biological features and prognostic outcomes (15, 16). Therefore, a novel predictive tool is needed to improve model performance and provide guidance for individualized treatment decisions. The innovation of this study is to combine Shap (Shapley Additive Explanations) with traditional machine learning, which breaks through the limitation of "black-boxing" of traditional machine learning models, and provides the importance scores of clinical variables for each prediction. This enables the model to not only provide highly accurate predictions but also quantify the specific impact of clinical variables on patient prognosis. This feature significantly improves the clinical usability of the model, and our study provides innovative ideas for the prognostic management of patients with extremely aggressive subtypes of prostate cancer.

Methods

Data source and patient selection

Patient information on extremely aggressive prostate cancer was obtained from the Surveillance, Epidemiology, and End Results (SEER) database, which covers approximately 30% of the U.S. population and is publicly accessible. We selected patients diagnosed between 2000 and 2020 with prostate cancer (ICD-O-3 code C61.9) who had pathological subtypes such as small cell carcinoma, large cell carcinoma, neuroendocrine carcinoma, squamous cell carcinoma, and aggressive ductal adenocarcinoma. Data extraction was performed using SEER*Stat software.

The exclusion criteria were as follows: (1) mismatched pathological type; (2) patients with multiple primary tumors; and (3) patients with incomplete clinical information, such as missing data on race, survival, TNM stage, PSA level, Gleason score, or other key clinical variables. The inclusion and exclusion process are depicted in Figure 1.

Study variables and feature selection

Data pertaining to demographics and clinical characteristics of prostate cancer patients were meticulously extracted from the SEER database. This encompassed variables such as age at diagnosis, race, gender, TNM stage as per the American Joint Committee on Cancer (AJCC) 7th edition, marital status, prostate-specific antigen (PSA) levels, Gleason score (GS), median household income, and various treatment modalities including surgery, radiotherapy, and chemotherapy. Following the categorization in previous studies (17, 18), age was divided into three groups: ≤60, 61–69, and ≥70 years. PSA levels were recorded as continuous variables, with values ≤0.1 ng/mL recorded as 0.1 ng/mL and values ≥98 ng/mL capped at 98 ng/mL, ranging from 0.1 to 98 ng/mL. Gleason scores were grouped into categories of ≤3 + 4, 4 + 3, 8, and ≥9. Missing data were addressed using the following strategies: for variables with missing rates below 20%, Random Forest Imputation was employed to estimate and fill in the missing values (19). Variables with more than 20% missing data were excluded from the analysis. In this study, all variables included in the analysis had missing rates below 20%. Among the variables included in the analysis, missing rates were as follows: Chemotherapy (4.2%), Marital status (6.8%), Income (3.1%), T stage (8.7%), N stage (7.3%) and M stage (4.1%). Random Forest Imputation (using the missForest package in R) was applied to ensure data completeness and consistency. For feature selection, we utilized the Boruta algorithm (20), which is a robust method for identifying the most significant features within a dataset. It determines feature importance by comparing the Z-scores of each actual feature against those of corresponding "shadow features." In this process, all genuine features are duplicated and shuffled to create shadow features, which are then evaluated using a Random Forest model to obtain their respective Z-scores. Additionally, the Z-scores of the shadow features are generated by randomly permuting the original features (21). A true feature is deemed "important" (indicated in green) and classified as an acceptable variable if its Z-score consistently surpasses the maximum Z-score of the shadow features across multiple independent tests. Conversely, if a true feature's Z-score does not significantly exceed that of the shadow features, it is labeled as "unimportant" (indicated in red) and classified as an unacceptable variable. Acceptable variables are retained during the feature selection process as they are considered to contribute positively to the

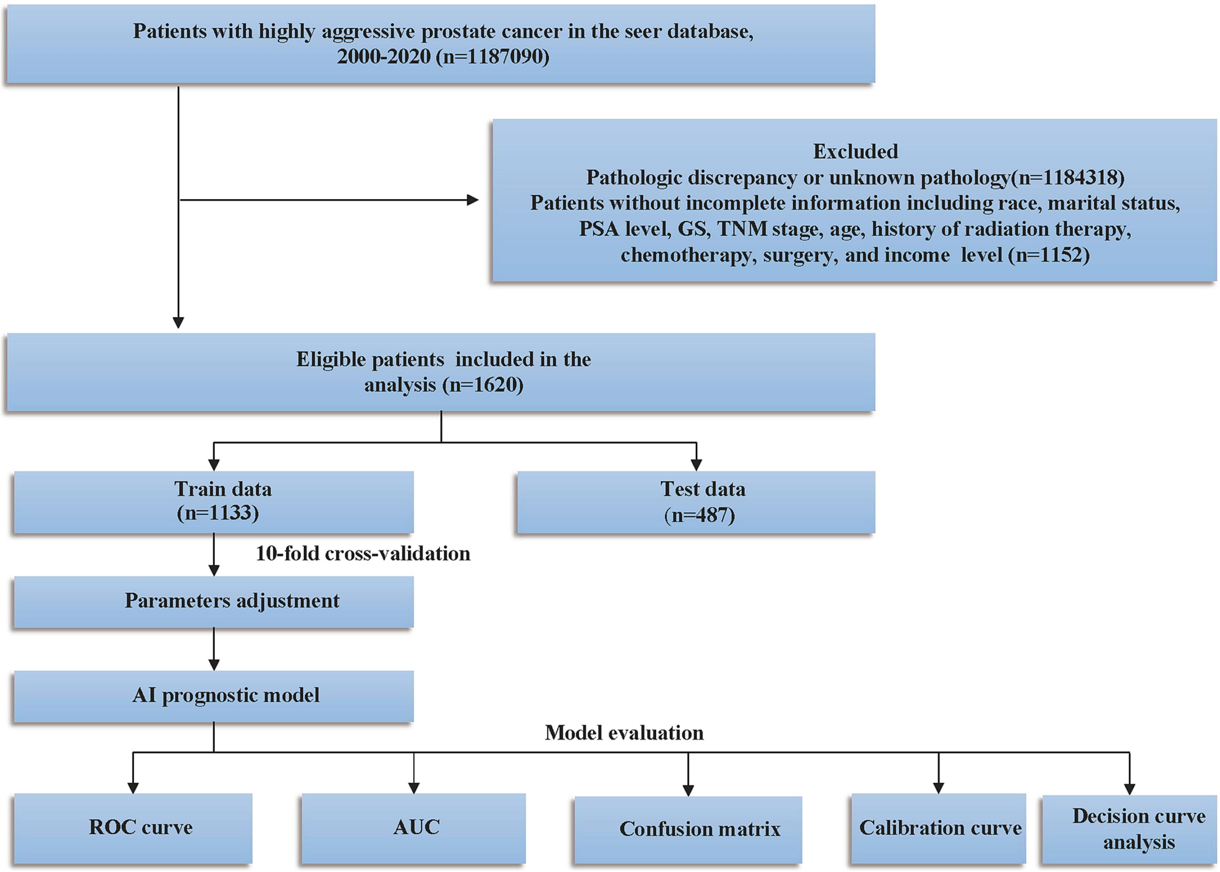


FIGURE 1
Study design and patient selection flowchart. PSA, prostate-specific antigen; GS, Gleason score; SEER, Surveillance, Epidemiology, and End Results; TNM, tumor lymph node metastasis; ROC, curve receiver operating characteristic curve; AUC, area under the curve.

model's performance. In contrast, unacceptable variables are excluded from the final feature set because they do not demonstrate sufficient predictive capability for the target variable during the feature selection process.

Model development

Prognostic models were constructed using nine machine learning algorithms: XGBoost, logistic regression (LR), support vector machine (SVM), random forest (RF), k-nearest neighbor (KNN), decision tree (DT), elastic network (Enet), multilayer perceptron (MLP), and lightGBM. To ensure model stability, the dataset was split into a 70:30 ratio for training and testing. Cross-validation was performed with 10-fold testing, and hyperparameters were tuned in the training set. Final validation was conducted on the test set. The objective was to develop models that could predict the overall survival of patients with extremely aggressive prostate cancer at 1, 3, and 5 years.

Statistical analysis

Categorical variables were analyzed using the χ^2 test and expressed as numbers (*n*) and percentages (%). Non-normally distributed continuous variables were assessed with the Kruskal-Wallis test and reported as medians with interquartile ranges (IQR). All statistical analyses and model development were conducted using R (version 4.0.5). A *p*-value of <0.05 was considered statistically significant.

Model performance evaluation

The performance of the nine machine learning models was evaluated using receiver operating characteristic (ROC) curve analysis and confusion matrices. The area under the curve (AUC) of the ROC curve measures the performance of the model, and F1 scores combining sensitivity and specificity are used to assess the robustness of the model (22). Additionally, calibration curves based on Bier scores and decision curve analysis (DCA) were applied to assess the models' prediction accuracy and clinical utility.

Model interpretation

SHAP (Shapley Additive Explanations) values were used to interpret the machine learning models. SHAP values, derived from game theory, provide insights into which features most significantly influence the model's predictions and how each feature affects the model's output.

10.3389/fmed.2025.1512870

Results

Patient characteristics

1,620 patients were included in this study, and the baseline characteristics of the training set and test set are shown in Table 1. There was no difference between the training set and validation set in the baseline data. There were 1,133 columns of patients assigned to the training set and 487 columns of patients assigned to the validation set. In the training set 631 patients died and 502 patients survived. In the validation set 277 patients died and 210 patients survived.

TABLE 1 Baseline characteristics of extremely aggressive prostate cancer patients.

Characteristics	Training cohort (*n* = 1,133)	Validation cohort (*n* = 487)	*P* value
Age, yr. *n* (%)			0.53
≤60	197 (17.39)	96 (19.71)	
61–69	363 (32.04)	153 (31.42)	
≥70	573 (50.57)	238 (48.87)	
Race, *n* (%)			0.09
White	915 (80.76)	386 (79.26)	
Black	114 (10.06)	65 (13.35)	
Other[a]	104 (9.18)	36 (7.39)	
Clinical T stage, *n* (%)			0.36
T1	426 (37.6)	211 (43.33)	
T2	312 (27.54)	125 (25.67)	
T3	204 (18.01)	82 (16.84)	
T4	191 (16.86)	69 (14.17)	
N, *n* (%)			0.86
N0	931 (82.17)	398 (81.72)	
N1	202 (17.83)	89 (18.28)	
M, *n* (%)			0.46
M0	761 (67.17)	337 (69.20)	
M1	372 (32.83)	150 (30.80)	
Surgery, *n* (%)			0.69
No/Unknown	549 (48.46)	230 (47.23)	
Yes	584 (51.54)	257 (52.77)	
Radiation, *n* (%)			0.56
Yes	353 (31.16)	144 (29.57)	
No/Unknown	780 (68.84)	343 (70.43)	
Chemotherapy, *n* (%)			0.70
Yes	284 (25.07)	117 (24.02)	
No/Unknown	849 (74.93)	370 (75.98)	
Survival status, *n* (%)			0.69
Dead	631 (55.69)	277 (56.88)	
Alive	502 (44.31)	210 (43.12)	
Marital status, *n* (%)			0.66
Married	791 (69.81)	334 (68.58)	
Unmarried[b]	342 (30.19)	153 (31.42)	
Income, *n* (%)			0.35
≤100,000	947 (83.58)	397 (81.52)	
>100,000	186 (16.42)	90 (18.48)	
PSA level (ng/ml)			0.86
Median [IQR]	8.900 [4.700, 19.582]	9.000 [4.300, 20.499]	

PSA, prostate specific antigen; IQR, interquartile range; Other[a]: Asian/Pacific Islander, American Indian/Alaska Native. Unmarried[b]: Widowed, Divorced, Separated, Single (never married).

Feature predictor selection

We use the same feature sets for our 1-, 3-, and 5-year prediction models. The Boruta algorithm identified unique feature sets for the 1-, 3-, and 5-year prediction models (Figure 2). The results showed that the feature variables included in the 1-year prognostic model were age, radiotherapy, N stage, surgery, PSA level, chemotherapy, and M stage (Figure 2A). Characteristic variables included in the 3-year prognostic model were T stage, radiotherapy, income level, N stage, age, PSA level, M stage and chemotherapy (Figure 2B). Characteristic variables included in the 5-year prognostic model were age, survival status, surgery, income, PSA level, chemotherapy, and M stage (Figure 2C).

Construction of machine learning predictive models

Considering survival months as the prognostic state, we integrate the features selected by the appeal-based Boruta algorithm into the variable training model. In the training set species, we used 10-fold cross-validation for iteration and optimization and finally determined that the lightGBM model performs best. We adjusted the parameter balance to avoid data overfitting and finally identified the key hyperparameters. The key parameters of lightGBM are as follows: tree_depth = 1, trees = 458, learn_rate = 0.0059, mtry = 5, min_n = 10, loss_reduction = 0.291. See Supplementary material 1 for hyperparameters of the nine machine learning models.

Evaluating machine learning prognostic models

Our analysis revealed that lightGBM demonstrated consistent efficacy in forecasting highly aggressive prostate cancer at 1, 3, and 5 years, as evidenced by the AUC values derived from the ROC curves of both the training and test sets. Data for 1 year (0.777 for the training set, 0.8 for the test set), 3 years (0.881 for the training set, 0.809 for the test set), and 5 years (0.888 for the training set, 0.773 for the test set) are presented in Figure 3 and Table 2.

See Table 2 for the best and most stable performance of lightGBM compared to the other 8 machine learning models. In addition, we evaluated the accuracy of the lightGBM model using

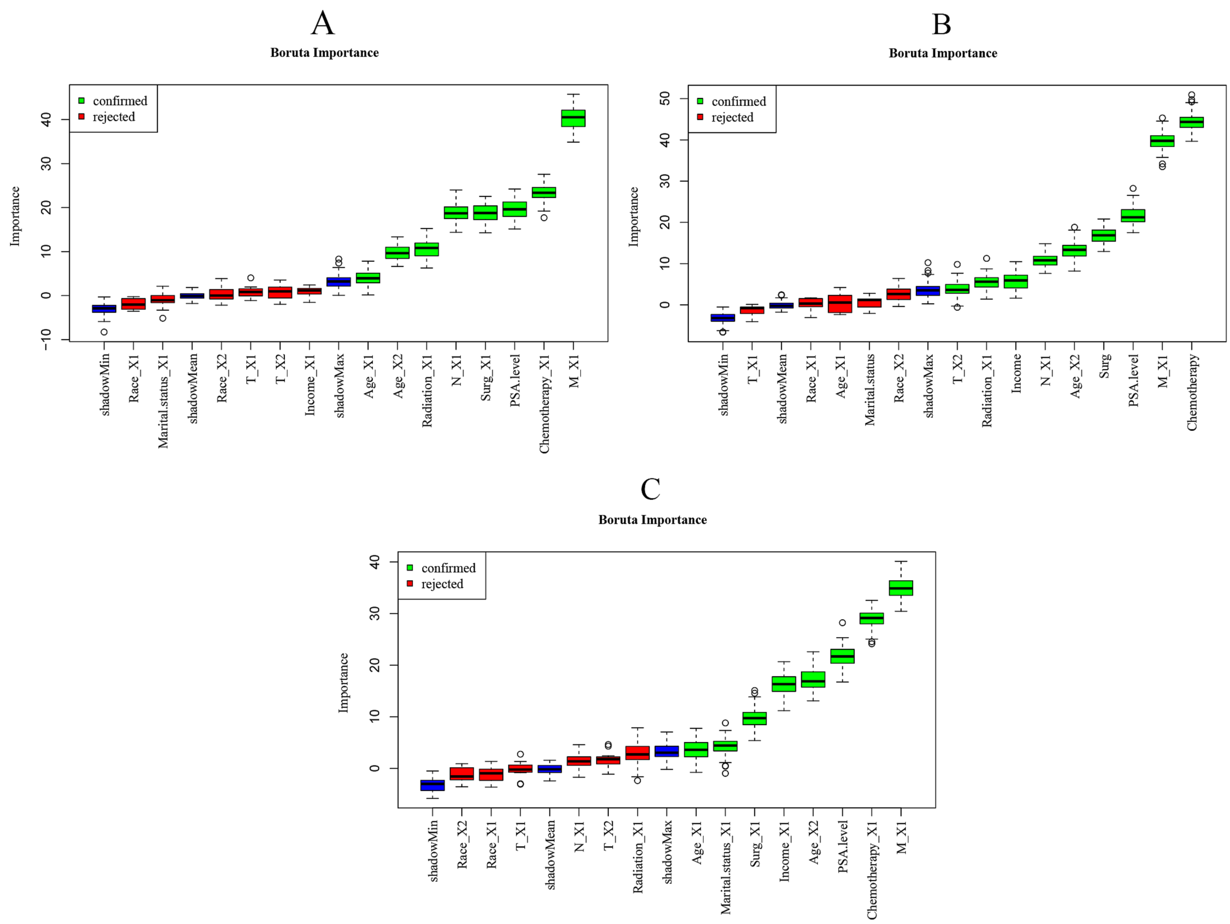


FIGURE 2
Importance of each feature in the predictive model based on Boruta's algorithm. **(A)** Importance of each feature in the 1-year prognostic model. **(B)** Importance of each feature in the 3-year prognostic model. **(C)** Importance of each feature in the 5-year prognostic model. The Boruta algorithm determines the importance of a feature by comparing the Z-score of each actual feature with the corresponding "shadow feature." A real feature is considered "important" (shown in green), whereas, if the Z-score of a real feature does not significantly exceed the Z-score of the shadow feature, it is marked as "not important" (shown in red) and classified as an unacceptable variable.

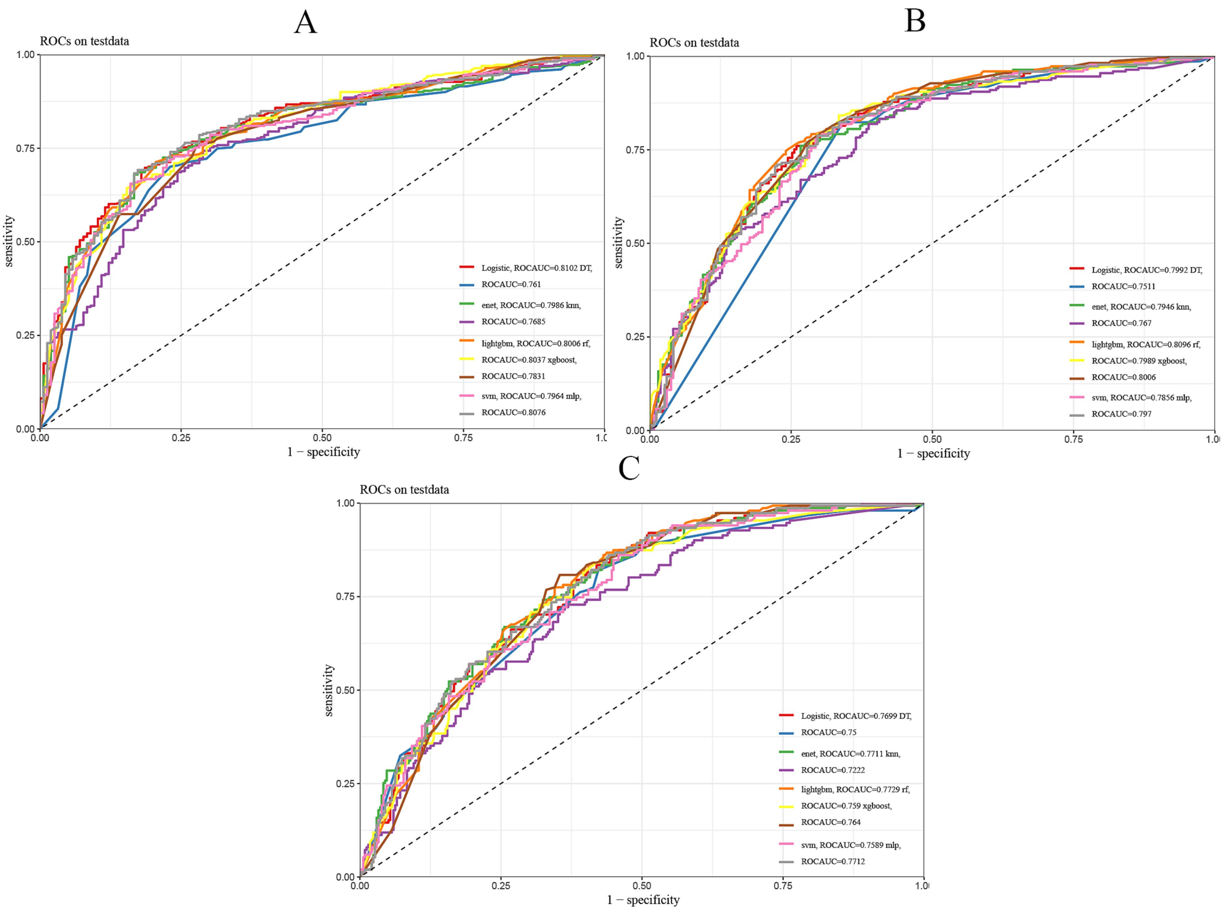


FIGURE 3
Nine machine learning models evaluated. **(A)** ROC curves of 1-year prognostic models in the test set. **(B)** ROC curves of 3-year prognostic models in the test set. **(C)** ROC curves of 5-year prognostic models in the test set. The plot presents the ROC curves for nine different machine learning models in a prediction task. The x-axis represents the false positive rate (FPR), and the y-axis represents the true positive rate (TPR). The area under the curve (AUC) reflects the overall performance of each model, with a larger AUC indicating better predictive ability.

a confusion matrix (Supplementary Figure 1). For 1-year, 3-year and 5-year survival predictions, F1 scores of lightGBM model validation set are 0.809, 0.751 and 0.611, respectively (Supplementary Table 1). Therefore, lightGBM model has the best predictive performance in 3-year and 5-year models. Although the one-year survival prediction is slightly lower than that of Logistic, MLP and RF models, the stability of LightGBM model is superior to these three models. In summary, we choose LightGBM model as the best model.

Finally, we used calibration curves based on Bier scores showing that the predictions of 1-, 3-, and 5-year survival probabilities in the train and test sets were also more consistent with the actual observations (Supplementary Figures 2, 3). Also, DCA decision curve analysis showed good clinical utility and positive net benefit of lightGBM in 1, 3, 5-year survival prediction (Figure 4).

Interpretation of models

These key features were ranked using a SHAP plot (Figure 5) showing the level of influence of the machine learning model for each feature. The SHAP plot showed that the largest factor influencing patient survival at 1 and 3 years was M stage and the largest factor influencing patient survival at 5 years was PSA level.

Application of model

To facilitate clinical adoption, we have uploaded the R code, dataset, and the completed model to Supplementary material 3. Additionally, we propose integrating this model into hospital electronic health records (EHRs) and clinical decision support systems (CDSS) to assist oncologists in real-time prognostic estimation.

Discussion

Patients with extremely aggressive prostate cancer, including small cell carcinoma, large cell carcinoma, squamous cell carcinoma, neuroendocrine carcinoma, undifferentiated carcinoma, aggressive ductal carcinoma, and ductal adenocarcinoma, often exhibit more aggressive biological behavior and have a poorer prognosis compared to other forms of prostate cancer (23–25). Accurate survival prediction for these patients is therefore clinically significant. However, current

TABLE 2 Performance of predictive models built by 9 machine learning algorithms in training and test sets (area under the ROC curve).

	1-year survival	3-year survival	5-year survival
Train set			
LightGBM	0.777	0.881	0.888
DT	0.856	0.782	0.853
ENET	0.768	0.782	0.853
KNN	0.909	0.788	0.805
Logistic	0.776	0.805	0.824
MLP	0.777	0.869	0.862
RF	0.852	0.796	0.819
SVM	0.779	0.802	0.807
XGBoost	0.763	0.799	0.808
Test set			
LightGBM	0.800	0.809	0.773
DT	0.761	0.751	0.75
ENET	0.798	0.795	0.771
KNN	0.769	0.767	0.722
Logistic	0.810	0.799	0.769
MLP	0.808	0.797	0.771
RF	0.804	0.798	0.759
SVM	0.796	0.786	0.758
XGBoost	0.783	0.800	0.764

DT, decision tree; ENET, Elastic Net; KNN, K-Nearest Neighbors; LightGBM, Light Gradient Boosting Machine; RF, Random Forest; XGBoost, Extreme Gradient Boosting; SVM, Support Vector Machine; MLP, Multi-Layer Perceptron.

clinical tools for prognostic prediction in extremely aggressive prostate cancer have substantial limitations, particularly the absence of reliable models that leverage artificial intelligence and machine learning.

This research involved the creation of nine models grounded in machine learning to forecast survival rates at 1, 3, and 5 years for the patient cohort in question. Among these, the lightGBM model showed the highest predictive performance, with AUCs of 0.77, 0.80, 0.88, and 0.81 for the training and test sets at 1, 3, and 5 years, respectively, demonstrating strong predictive ability. An AUC value of ≥0.7 is considered indicative of a model with sufficient predictive power (26).

In recent years, artificial intelligence has garnered increasing attention in the medical field, including in prostate cancer research (27–30). In contrast to conventional algorithms, machine learning models operate without the limitations imposed by non-proportionality, multicollinearity, or nonlinearity challenges (30). Thereby minimizing biases that can arise from conventional modeling. For example, Peng et al. used machine learning algorithms to develop a survival prognostic model for patients with lymph node-positive prostate cancer, achieving better predictive performance than traditional Cox regression models (31). Similarly, Dai et al. (32) demonstrated that machine learning models outperformed traditional algorithms in predicting survival for patients with confined prostate cancer.

In this study, we incorporated 12 key clinical characteristics of patients with extremely aggressive prostate cancer and used the Boruta algorithm, a feature selection method based on random forest classifiers, to select the most relevant features for survival prediction. The Boruta algorithm is designed to identify all variables that are important to the dependent variable, rather than the smallest set of features relevant to a particular model (33, 34). In contrast to the objective of a typical feature selection algorithm, the Boruta feature selection algorithm aims to identify the features that hold the greatest relevance to the dependent variable, rather than merely seeking the most compact set of features pertinent to a specific model (34). Our results identified factors such as age, PSA level, surgery, and radiotherapy as key risk factors for prognosis, with tumor metastasis (M stage) emerging as the most significant predictor of survival at 1 and 3 years, and PSA level as the strongest predictor at 5 years. These findings have important clinical implications. For example, the model highlights surgery and radiotherapy as influential factors, suggesting that multimodal treatment approaches may provide survival benefits in certain subgroups of patients with highly aggressive prostate cancer. This underscores the need for personalized treatment selection based on a patient's predicted prognosis and treatment response patterns.

A systematic review identified high Gleason scores as independent risk factors for early tumor progression, and multiple organ metastases were associated with reduced survival (35). In a separate investigation, the median overall survival for patients newly diagnosed with neuroendocrine prostate cancer was recorded at 16.8 months, significantly less than the 53.5 months noted in cases associated with treatment (36). Regarding treatment, platinum-based chemotherapy is commonly used for patients with small cell carcinoma. Combination regimens including cisplatin, etoposide, and doxorubicin have shown partial benefit, though they are not recommended for neuroendocrine prostate cancer patients due to the risk of severe neutropenia. For neuroendocrine prostate cancer, immune checkpoint inhibitors, such as atezolizumab combined with platinum-based chemotherapy (36) or second-line treatments such as natalizumab with ibritumomab may be considered (37).

Early detection of prostate cancer is critical. Various non-invasive imaging techniques have been studied for predicting metastasis (38–40). Multiparametric MRI (mpMRI) has shown enhanced sensitivity and specificity relative to conventional MRI in the identification of tumors and lymph nodes; however, it may experience signal loss or image distortion in DWI sequences (39). Similarly, PSMA PET/CT is extensively utilized for the detection of prostate cancer in both soft tissue and bone, yet its detection rate for lymph node metastases measuring 2–5 mm hovers around 60% (40, 41). Emerging imaging techniques, such as MR lymphography and targeted PET using superparamagnetic iron oxide (SPIO) nanoparticles, are under investigation, though their effectiveness in predicting lymph node metastasis remains uncertain (41–43). Furthermore, fluid-based diagnostics, exemplified by the FDA-approved Prostate Cancer Antigen 3 (PCA3), which is a urine-based, non-coding RNA biomarker, have demonstrated promise in informing decisions regarding repeat biopsies, with reported AUCs varying from 0.64 to 0.762 (43, 44). Other urine-based genomic assays, including multigene panels (e.g., PUR), exosome-based assays (e.g., ExoDx), DNA methylation markers (e.g., epiCaPture), and mRNA-based assays (e.g., SelectMDx), have also demonstrated prognostic value (44, 45). Lih et al. (46) identified urinary glycopeptides, such as ACPP, CLU, ORM1, and CD97, that may help differentiate between low- and high-risk prostate cancer,

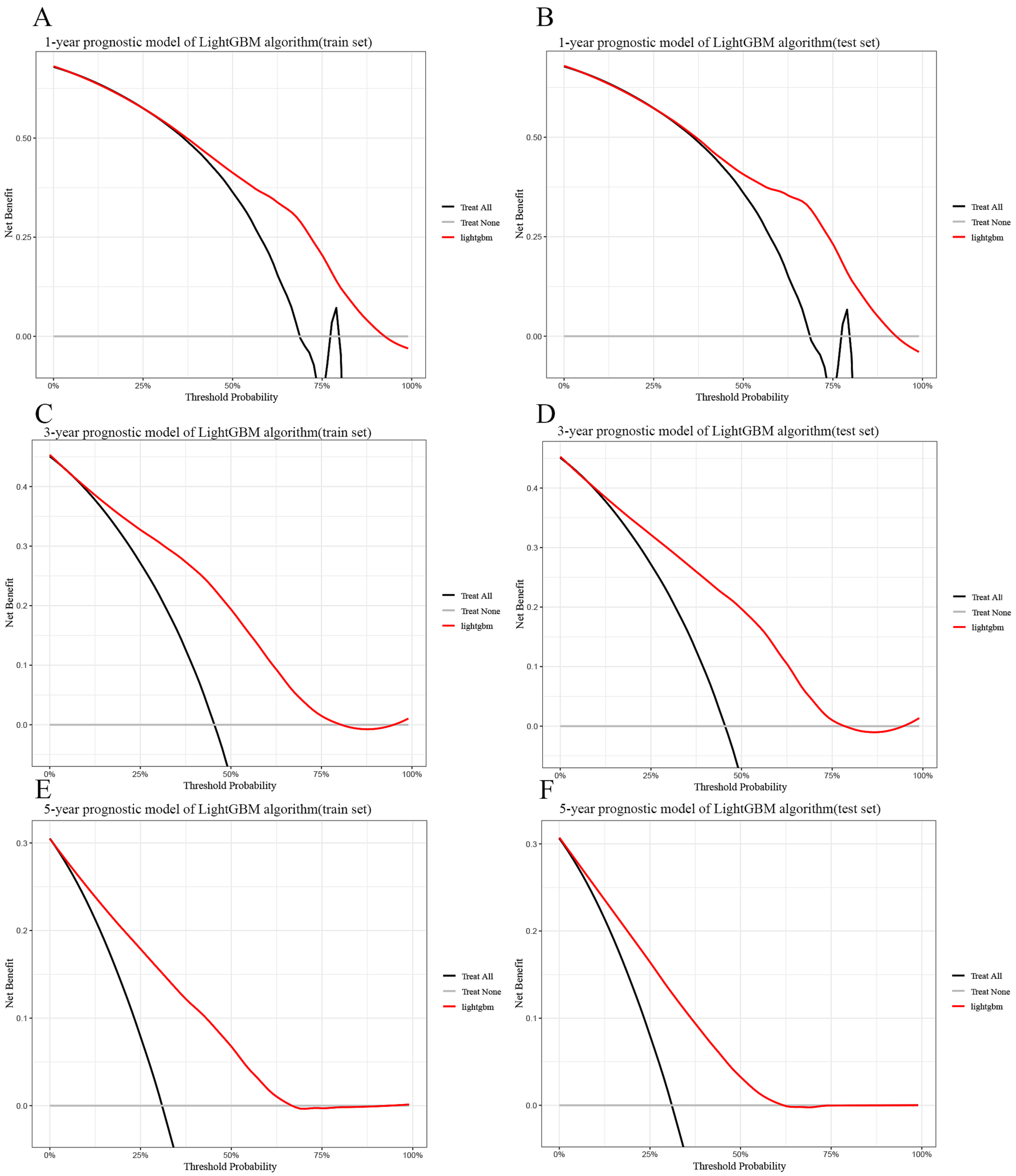


FIGURE 4
Decision curve analysis curves for the LightGBM model for the training and test sets. **(A)** 1-year train set. **(B)** 1-year test set. **(C)** 3-year train set. **(D)** 3-year test set. **(E)** 5-year train set. **(F)** 5-year test set. LightGBM: Light Gradient Boosting Machine. In the figure, the red curve represents the predicted performance of the GBM model, respectively. In addition, there are two lines, which represent two extreme cases. The gray vertical line indicates the assumption of survival for all patients. The black horizontal line indicates that there is no survival assumption. For example, in the 1-year training set, the survival probability is between 0.3 and 0.93. When using this GBM predictive model to make clinical decisions, survival probabilities can be distinguished.

showing potential for early identification of aggressive forms of the disease.

This study is the first to develop multiple machines learning prognostic models specifically for extremely aggressive prostate cancer. We incorporated 13 significant prognostic features and employed SHAP values to assess the contribution of each feature, revealing that metastasis, surgery, and PSA level were the most impactful variables.

However, this study has several limitations that should be acknowledged. First, as a retrospective study utilizing SEER data,

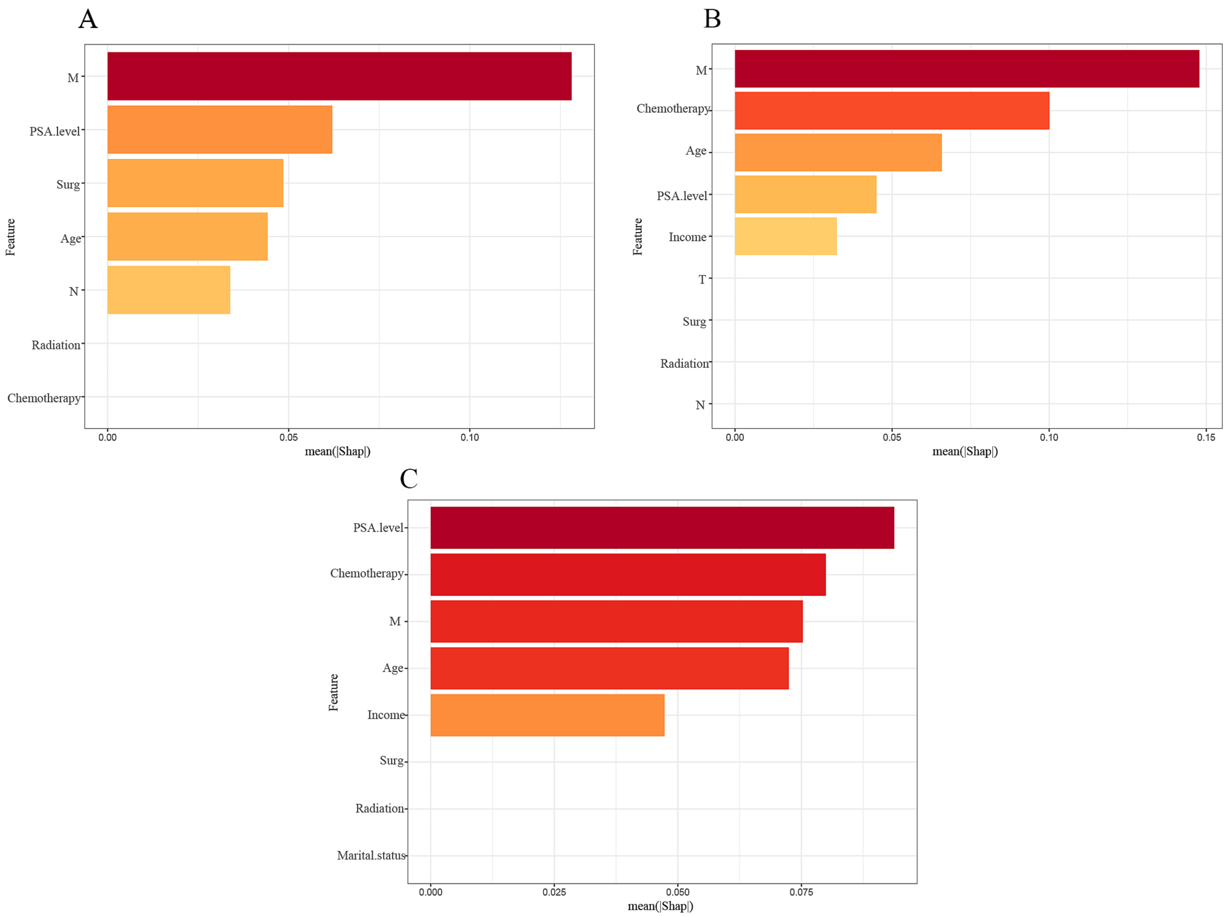


FIGURE 5
Importance ranking of features based on SHAP values in diagnostic models based on LightGBM algorithm. The features are ranked based on the sum of SHAP values of all the patients, and the distribution of the impact of each feature on the output of the model LightGBM is expressed in terms of the SHAP value. The x-axis represents the SHAP value's impact on the model's output. The higher the value of the x-axis, the greater the impact on the model. **(A)** 1-year model; **(B)** 3-year model; **(C)** 5-year model.

it may be subject to selection bias and incomplete case reporting, potentially affecting the generalizability of our findings. Second, the SEER database does not provide detailed molecular markers, genetic data, or treatment response information, which are critical for a more comprehensive prognostic assessment. The absence of these key clinical variables may limit the ability of our model to fully capture the biological heterogeneity of extremely aggressive prostate cancer. Future studies should aim to incorporate multi-omics data and real-world patient responses to further refine predictive accuracy. Additionally, while our model has demonstrated strong internal validation, external validation on independent datasets and prospective clinical trials are needed to ensure its applicability across diverse populations.

Overall, this study highlights the potential of machine learning models to guide clinical decisions and optimize treatment strategies for extremely aggressive prostate cancer. Specifically, our model can be used for risk stratification and treatment planning of patients, as well as monitoring and follow-up adjustments for patients at different risks, and finally, by integrating the model into EHRs and CDSS, can provide real-time survival predictions to help physicians make evidence-based treatment recommendations. With the accumulation of more clinical data and further optimization of algorithms, AI-based prognostic models could significantly improve treatment outcomes and survival for patients with extremely aggressive prostate cancer in the future.

Conclusion

In conclusion, we developed and evaluated nine machine learning models, incorporating SHAP values to enhance interpretability, for predicting survival in patients with extremely aggressive prostate cancer. Among them, the lightGBM model demonstrated the best predictive performance, offering a valuable clinical tool for personalized prognosis estimation. Future research should focus on external validation using independent cohorts, integrating molecular biomarkers, and exploring the incorporation of real-time patient data to further enhance the model's robustness and clinical utility.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://seer.cancer.gov/.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

CP: Conceptualization, Investigation, Software, Writing – original draft, Writing – review & editing. CG: Writing – original draft. XZ: Writing – original draft. DL: Conceptualization, Data curation, Investigation, Software, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

We thank the SEER database for open data access.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2025.1512870/full#supplementary-material

References

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*. (2024) 74:12–49. doi: 10.3322/caac.21820

2. Humphrey PA. Histological variants of prostatic carcinoma and their significance. *Histopathology*. (2012) 60:59–74. doi: 10.1111/j.1365-2559.2011.04039.x

3. Stein ME, Bernstein Z, Abacioglu U, Sengoz M, Miller RC, Meirovitz A, et al. Small cell (neuroendocrine) carcinoma of the prostate: etiology, diagnosis, prognosis, and therapeutic implications--a retrospective study of 30 patients from the rare cancer network. *Am J Med Sci*. (2008) 336:478–88. doi: 10.1097/MAJ.0b013e3181731e58

4. Epstein JI, Amin MB, Beltran H, Lotan TL, Mosquera J-M, Reuter VE, et al. Proposed morphologic classification of prostate cancer with neuroendocrine differentiation. *Am J Surg Pathol*. (2014) 38:756–67. doi: 10.1097/PAS.0000000000000208

5. Sheng Z-C, Dong J, Xu S. Clinically rare subtypes of prostate cancer: Progress in research. *Zhonghua Nan Ke Xue*. (2023) 29:264–8.

6. Abbott T, Ng K, Nobes J, Muehlschlegel P. Small-cell carcinoma of the prostate - challenges of diagnosis and treatment: a next of kin and physician perspective piece. *Oncol Ther*. (2023) 11:291–301. doi: 10.1007/s40487-023-00238-3

7. Taher A, Jensen CT, Yedururi S, Surasi DS, Faria SC, Bathala TK, et al. Imaging of neuroendocrine prostatic carcinoma. *Cancers (Basel)*. (2021) 13:5765. doi: 10.3390/cancers13225765

8. Wang Y, Wang Y, Ci X, Choi SYC, Crea F, Lin D, et al. Molecular events in neuroendocrine prostate cancer development. *Nat Rev Urol*. (2021) 18:581–96. doi: 10.1038/s41585-021-00490-0

9. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med*. (2021) 13:152. doi: 10.1186/s13073-021-00968-x

10. Nguyen TT, Ho CT, Bui HTT, Ho LK, Ta VT. Multidimensional machine learning for assessing parameters associated with COVID-19 in Vietnam: validation study. *JMIR Form Res*. (2023) 7:e42895. doi: 10.2196/42895

11. Sajda P. Machine learning for detection and diagnosis of disease. *Annu Rev Biomed Eng*. (2006) 8:537–65. doi: 10.1146/annurev.bioeng.8.061505.095802

12. Senders JT, Staples P, Mehrtash A, Cote DJ, Taphoorn MJB, Reardon DA, et al. An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning. *Neurosurgery*. (2020) 86:E184–92. doi: 10.1093/neuros/nyz403

13. Pak S, Park SG, Park J, Cho ST, Lee YG, Ahn H. Applications of artificial intelligence in urologic oncology. *Investig Clin Urol*. (2024) 65:202–16. doi: 10.4111/icu.20230435

14. Zhang B, Shi H, Wang H. Machine learning and AI in Cancer prognosis, prediction, and treatment selection: a critical approach. *J Multidiscip Healthc*. (2023) 16:1779–91. doi: 10.2147/JMDH.S410301

15. Spooner A, Chen E, Sowmya A, Sachdev P, Kochan NA, Trollor J, et al. A comparison of machine learning methods for survival analysis of high -dimensional clinical data for dementia prediction. *Sci Rep*. (2020) 10:20410. doi: 10.1038/s41598-020-77220-w

16. Hao L, Kim J, Kwon S, Ha ID. Deep learning-based survival analysis for high-dimensional survival data. *Mathematics*. (2021) 9:1244. doi: 10.3390/math9111244

17. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. (2019) 380:1347–58. doi: 10.1056/NEJMra1814259

18. Abdollah F, Karnes RJ, Suardi N, Cozzarini C, Gandaglia G, Fossati N, et al. Predicting survival of patients with node-positive prostate Cancer following multimodal treatment. *Eur Urol*. (2014) 65:554–62. doi: 10.1016/j.eururo.2013.09.025

19. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics*. (2012) 28:112–8. doi: 10.1093/bioinformatics/btr597

20. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw*. (2010) 36:1–13. doi: 10.18637/jss.v036.i11

21. Yue S, Li S, Huang X, Liu J, Hou X, Zhao Y, et al. Machine learning for the prediction of acute kidney injury in patients with sepsis. *J Transl Med*. (2022) 20:215. doi: 10.1186/s12967-022-03364-0

22. Davis J, Goadrich M. The relationship between precision-recall and ROC curves In: Proceedings of the 23rd international conference on machine learning. Pittsburgh, PA: Association for Computing Machinery (2006). 233–40.

23. Alabi BR, Liu S, Stoyanova T. Current and emerging therapies for neuroendocrine prostate cancer. *Pharmacol Ther*. (2022) 238:108255. doi: 10.1016/j.pharmthera.2022.108255

24. Spetsieris N, Boukovala M, Patsakis G, Alafis I, Efstathiou E. Neuroendocrine and aggressive-variant prostate cancer. *Cancers*. (2020) 12:3792. doi: 10.3390/cancers12123792

25. Tsaur I, Heidegger I, Kretschmer A, Borgmann H, Gandaglia G, Briganti A, et al. Aggressive variants of prostate cancer - are we ready to apply specific treatment right now? *Cancer Treat Rev*. (2019) 75:20–6. doi: 10.1016/j.ctrv.2019.03.001

26. Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Med*. (2003) 29:1043–51. doi: 10.1007/s00134-003-1761-8

27. Goldenberg SL, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. *Nat Rev Urol*. (2019) 16:391–403. doi: 10.1038/s41585-019-0193-3

28. Tătaru OS, Vartolomei MD, Rassweiler JJ, Virgil O, Lucarelli G, Porpiglia F, et al. Artificial intelligence and machine learning in prostate cancer patient management-current trends and future perspectives. *Diagnostics*. (2021) 11:354. doi: 10.3390/diagnostics11020354

29. Zhu L, Pan J, Mou W, Deng L, Zhu Y, Wang Y, et al. Harnessing artificial intelligence for prostate cancer management. *Cell Rep Med*. (2024) 5:101506. doi: 10.1016/j.xcrm.2024.101506

30. Du M, Haag DG, Lynch JW, Mittinty MN. Comparison of the tree-based machine learning algorithms to cox regression in predicting the survival of Oral and pharyngeal cancers: analyses based on SEER database. *Cancers*. (2020) 12:2802. doi: 10.3390/cancers12102802

31. Peng Z-H, Tian J-H, Chen B-H, Zhou H-B, Bi H, He M-X, et al. Development of machine learning prognostic models for overall survival of prostate cancer patients with lymph node-positive. *Sci Rep*. (2023) 13:18424. doi: 10.1038/s41598-023-45804-x

32. Dai X, Park JH, Yoo S, D'Imperio N, McMahon BH, Rentsch CT, et al. Survival analysis of localized prostate cancer with deep learning. *Sci Rep*. (2022) 12:17821. doi: 10.1038/s41598-022-22118-y

33. Wei L, Wan S, Guo J, Wong KK. A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif Intell Med*. (2017) 83:82–90. doi: 10.1016/j.artmed.2017.02.005

34. Zhou H, Xin Y, Li S. A diabetes prediction model based on Boruta feature selection and ensemble learning. *BMC Bioinformatics*. (2023) 24:224. doi: 10.1186/s12859-023-05300-5

35. Conteduca V, Oromendia C, Eng KW, Bareja R, Sigouros M, Molina A, et al. Clinical features of neuroendocrine prostate cancer. *Eur J Cancer*. (2019) 121:7–18. doi: 10.1016/j.ejca.2019.08.011

36. Horn L, Mansfield AS, Szczęsna A, Havel L, Krzakowski M, Hochmair MJ, et al. First-line Atezolizumab plus chemotherapy in extensive-stage small-cell lung cancer. *N Engl J Med*. (2018) 379:2220–9. doi: 10.1056/NEJMoa1809064

37. Antonia SJ, López-Martin JA, Bendell J, Ott PA, Taylor M, Eder JP, et al. Nivolumab alone and nivolumab plus ipilimumab in recurrent small-cell lung cancer (CheckMate 032): a multicentre, open-label, phase 1/2 trial. *Lancet Oncol*. (2016) 17:883–95. doi: 10.1016/S1470-2045(16)30098-5

38. Hofman MS, Hicks RJ, Maurer T, Eiber M. Prostate-specific membrane antigen PET: clinical utility in prostate cancer, normal patterns, pearls, and pitfalls. *Radiographics*. (2018) 38:200–17. doi: 10.1148/rg.2018170108

39. van Leeuwen PJ, Emmett L, Ho B, Delprado W, Ting F, Nguyen Q, et al. Prospective evaluation of 68Gallium-prostate-specific membrane antigen positron emission tomography/computed tomography for preoperative lymph node staging in prostate cancer. *BJU Int*. (2017) 119:209–15. doi: 10.1111/bju.13540

40. von Bodman C, Godoy G, Chade DC, Cronin A, Tafe LJ, Fine SW, et al. Predicting biochemical recurrence-free survival for patients with positive pelvic lymph nodes at radical prostatectomy. *J Urol*. (2010) 184:143–8. doi: 10.1016/j.juro.2010.03.039

41. Rittenhouse H, Blase A, Shamel B, Schalken J, Groskopf J. The long and winding road to FDA approval of a novel prostate cancer test: our story. *Clin Chem*. (2013) 59:32–4. doi: 10.1373/clinchem.2012.198739

42. Stephan C, Ralla B, Jung K. Prostate-specific antigen and other serum and urine markers in prostate cancer. *Biochim Biophys Acta*. (2014) 1846:99–112. doi: 10.1016/j.bbcan.2014.04.001

43. Muteganya R, Goldman S, Aoun F, Roumeguère T, Albisinni S. Current imaging techniques for lymph node staging in prostate cancer: a review. *Front Surg*. (2018) 5:74. doi: 10.3389/fsurg.2018.00074

44. Fujita K, Nonomura N. Urinary biomarkers of prostate cancer. *Int J Urol*. (2018) 25:770–9. doi: 10.1111/iju.13734

45. Kim Y, Jeon J, Mejia S, Yao CQ, Ignatchenko V, Nyalwidhe JO, et al. Targeted proteomics identifies liquid-biopsy signatures for extracapsular prostate cancer. *Nat Commun*. (2016) 7:11906. doi: 10.1038/ncomms11906

46. Lih T-SM, Dong M, Mangold L, Partin A, Zhang H. Urinary marker panels for aggressive prostate cancer detection. *Sci Rep*. (2022) 12:14837. doi: 10.1038/s41598-022-19134-3

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 14 May 2025
DOI 10.3389/fmed.2025.1545730

Check for updates

OPEN ACCESS

EDITED BY
Ariel Soares Teles,
Federal Institute of Education, Science and Technology of Maranhão, Brazil

REVIEWED BY
Manuel Corpas,
University of Westminster, United Kingdom
Shubashini Velu,
Prince Mohammad bin Fahd University, Saudi Arabia

*CORRESPONDENCE
Xiaoyang Li
✉ woodslee429@126.com
Rong Wang
✉ wangrong@ystt.org.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 15 December 2024
ACCEPTED 24 April 2025
PUBLISHED 14 May 2025

CITATION
Lu X, Gao X, Wang X, Gong Z, Cheng J, Hu W, Wu S, Wang R and Li X (2025) Comparison of medical history documentation efficiency and quality based on GPT-4o: a study on the comparison between residents and artificial intelligence. *Front. Med.* 12:1545730.
doi: 10.3389/fmed.2025.1545730

Comparison of medical history documentation efficiency and quality based on GPT-4o: a study on the comparison between residents and artificial intelligence

Xiaojing Lu[1†], Xinqi Gao[1†], Xinyi Wang[1], Zhenye Gong[1], Jie Cheng[1], Weiguo Hu[1], Shaun Wu[2], Rong Wang[3*] and Xiaoyang Li[1*]

[1]Department of Medical Education, Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China, [2]WORK Medical Technology Group LTD, Hangzhou, China, [3]Shanghai Resident Sandardized Training Center, Shanghai, China

Background: As medical technology advances, physicians' responsibilities in clinical practice continue to increase, with medical history documentation becoming an essential component. Artificial Intelligence (AI) technologies, particularly advances in Natural Language Processing (NLP), have introduced new possibilities for medical documentation. This study aims to evaluate the efficiency and quality of medical history documentation by ChatGPT-4o compared to resident physicians and explore the potential applications of AI in clinical documentation.

Methods: Using a non-inferiority design, this study compared the documentation time and quality scores between 5 resident physicians from the hematology department (with an average of 2.4 years of clinical experience) and ChatGPT-4o based on identical case materials. Medical history quality was evaluated by two attending physicians with over 10 years of clinical experience using ten case content criteria. Data were analyzed using paired t-tests and Wilcoxon signed-rank tests, with Kappa coefficients used to assess scoring consistency. Detailed scoring criteria included completeness (coverage of history elements), accuracy (correctness of information), logic (organization and coherence of content), and professionalism (appropriate use of medical terminology and format), each rated on a 10-point scale.

Results: In terms of medical history quality, ChatGPT-4o achieved an average score of 88.9, while resident physicians scored 89.6, with no statistically significant difference between the two ($p = 0.25$). The Kappa coefficient between the two evaluators was 0.82, indicating good consistency in scoring. Non-inferiority testing showed that ChatGPT-4o's quality scores fell within the preset non-inferiority margin (5 points), indicating that its documentation quality was not inferior to that of resident physicians. ChatGPT-4o's average documentation time was 40.1 s, significantly shorter than the resident physicians' average of 14.9 min ($p < 0.001$).

Conclusion: While maintaining quality comparable to resident physicians, ChatGPT-4o significantly reduced the time required for medical history documentation. Despite these positive results, practical considerations such as data preprocessing, data security, and privacy protection must be addressed in real-world applications. Future research should further explore ChatGPT-4o's capabilities in handling complex cases and its applicability across different clinical settings.

KEYWORDS

artificial intelligence, GPT-4o, medical history documentation, quality, efficiency

Introduction

With the continuous advancement of medical technology, physicians are shouldering increasingly greater responsibilities in clinical practice (1). The collection and documentation of medical history has become an indispensable part of daily work, particularly in the management of hospitalized patients. Medical history serves not only as a crucial basis for diagnosis and treatment but also as a key document for legal and insurance purposes (2). Therefore, accurate and comprehensive documentation is vital for patient outcomes and the quality of healthcare services (3).

However, in busy hospital environments, resident physicians often face tremendous time pressure (4). Particularly in China, they are required to complete high-quality medical history documentation within limited time frames, which undoubtedly presents a significant challenge. This situation may affect the quality of documentation, leading to reduced work efficiency and increased professional burnout among physicians.

In recent years, the application of Artificial Intelligence (AI) technology in healthcare has been expanding, bringing new possibilities for improving the quality and efficiency of healthcare delivery (5–7). Among these technologies, Natural Language Processing (NLP) has demonstrated remarkable potential in medical text generation and analysis (8). The emergence of large language models like GPT-4o, in particular, has made AI-assisted medical documentation possible, potentially transforming traditional documentation methods (9).

GPT-4o (10), through its analysis of vast amounts of language data, can generate structured and coherent text, establishing a solid foundation for its application in medical documentation (11). However, despite AI's promising prospects in healthcare, its effectiveness and reliability in actual clinical settings still require further validation (12). Particularly in generating critical medical documents such as medical histories, AI's performance needs thorough investigation.

This study hypothesizes that when provided with identical case materials, ChatGPT-4o can complete medical history documentation in less time while maintaining quality comparable to that of resident physicians. Through systematic comparison of documentation time and quality between the two, we aim to evaluate ChatGPT-4o's potential applications in actual clinical work and provide reference for AI's further development in healthcare.

The research findings may offer new insights into current medical documentation practices and provide novel solutions for optimizing resource allocation and improving work efficiency in healthcare institutions. Furthermore, this study will explore the limitations of AI applications in healthcare, providing direction for subsequent technological improvements and practical applications.

Methods

Study design

This study adopts a non-inferiority comparative design to evaluate the performance of ChatGPT-4o and residents in terms of medical record quality and efficiency. The study participants include five residents (3 males, 2 females) from the hematology department, a computer system equipped with ChatGPT-4o, and two attending physicians with more than 10 years of clinical experience, who will independently score the quality of medical records. Each resident and ChatGPT-4o will generate medical records based on the same case materials, and the attending physicians will score the quality of these records. The evaluation criteria include completeness, accuracy, logic, and professionalism, with clear and standardized scoring criteria to ensure consistency and objectivity in the assessment.

Participants

- Residents: five residents currently undergoing standardized training in hematology, each with at least 1 year of clinical experience (average experience 2.4 ± 0.9 years, ensuring they possess sufficient skills in medical record collection and documentation. The residents' abilities in record-keeping will be pre-assessed to minimize individual differences that may influence the results. Selection criteria for residents included: (1) currently undergoing standardized training; (2) having at least 1 year of clinical experience; and (3) having recorded at least 30 hematology cases in the past 2 months.
- ChatGPT-4o: The latest version of ChatGPT-4o will be used to generate medical records. To ensure comparability, the system configuration and usage will be standardized, including the setting of prompts and generation parameters. Detailed configuration is provided in Appendix A. The main prompt template used was: "Based on the following transcribed doctor-patient dialogue, please generate a standard hematology medical history record, including chief complaint, present illness, past medical history, personal history, family history, physical examination, auxiliary examination, and diagnosis. Please ensure the content is complete, accurate, logically clear, and meets professional standards."
- Attending Physicians: two experienced hematology attending physicians were responsible for scoring the medical records. Both had over 10 years of clinical experience and had been involved in resident training for the past 3 years. The scoring process was independent, with clear evaluation criteria to ensure consistency in the results.

Data collection

- Interview Transcription: the resident will record the entire interview process while taking the patient's medical history, and the recorded content will be transcribed by specialized software (iFlytek Medical Version 1.2.0) into text, which will serve as the basis for the medical record. All transcriptions will undergo quality checks to ensure accuracy. The transcription process included: (1) audio collection (resident-patient dialogue); (2) automatic transcription (using speech recognition software); (3) manual correction (linguistic experts checking and correcting errors in automatic transcription); and (4) quality review (attending physicians confirming medical accuracy of the transcription). Transcription quality was assessed by comparison with the original audio, achieving an average accuracy rate of over 95%.
- Medical Record Documentation: each resident will independently document the medical record based on the transcribed text, and the same materials will be input into the ChatGPT-4o system to generate a medical record. The time taken for each resident and ChatGPT-4o to complete the medical record will be recorded to ensure comparability of time differences.
- Quality Scoring: the two attending physicians will independently score the medical records based on completeness, accuracy, logic, and professionalism. The scoring used a 100-point scale, and the final score will be the average of the two attending physicians' scores. Detailed scoring criteria are presented in Table 1 and Appendix B.

Sample size calculation

The sample size calculation was based on a non-inferiority design. With an anticipated standard deviation of 10 points for quality scores, a non-inferiority margin (Δ) of 5 points (5% of the total score), a significance level (α) of 0.05, and a statistical power (β) of 0.80, we determined that each group required 63 cases. This 5-point margin was established through consultation with experienced attending physicians who considered a difference of <5% in overall quality score to be clinically insignificant. To account for potential issues such as transcription quality, we included a final total of 65 cases to enhance the study's reliability. It is important to note that while only 5 residents participated, the unit of analysis was the medical record, not the number of participants, which aligns with the requirements of non-inferiority study design (13–15). We acknowledge the limitations of this sampling strategy and discuss them in detail in the discussion section.

Evaluation indicators

- Medical Record Quality: scored by attending physicians, evaluating the completeness, accuracy, logic, and professionalism of the medical records.
- Documentation Time: the time taken by each resident and ChatGPT-4o to complete the medical record, measured in minutes.
- Medical Record Quality: scored by attending physicians, evaluating different aspects of the medical records across three main categories:

  General Items (11 points): including chief complaint (6 points) and overall requirements (5 points)
  Core Content (55 points): including present illness (30 points), past medical history (10 points), personal history (10 points), and family history (5 points)
  Examination and Diagnosis (34 points): including physical examination (20 points), auxiliary examination (10 points), and diagnosis (4 points).

Data preprocessing

To ensure that ChatGPT-4o could effectively process medical dialogues, we performed the following preprocessing on the transcribed text:

TABLE 1 Medical record quality scoring criteria.

Scoring category	Scoring item	Scoring criteria	Maximum points
General items	Chief complaint	Accurately extract main symptoms, concise and professional expression	6
	General requirements	Standardized format, complete content, clear structure	5
Core content	Present illness	Complete recording of onset time, triggers, clinical manifestations, medical visit process, treatment effects, etc.	30
	Past medical history	Accurate recording of all past diseases, surgeries, blood transfusions, allergies, etc.	10
	Personal history	Comprehensive recording of lifestyle habits, occupational exposure, social psychological factors, etc.	10
	Family history	Complete recording of family members' relevant disease history	5
Examination and diagnosis	Physical examination	Systematic and comprehensive physical findings, accurate description of abnormalities	20
	Auxiliary examination	Accurate recording of all examination results with important results highlighted	10
	Diagnosis	Diagnosis consistent with clinical manifestations, reasonable logical reasoning	4
Total			**100**

- Removal of filler words and repetitive content
- Standardization of medical terminology and abbreviations
- Organization of question-answer pairs in chronological order
- Addition of simple classification tags (such as "symptom description," "treatment experience") to unstructured dialogues

Preprocessing was conducted by a linguist with medical background and an information technology specialist, and reviewed by the project's supervising physician. These preprocessing steps ensured that the content input into ChatGPT-4o was structured clearly and contained the necessary medical information while preserving the original dialogue content as much as possible. The same preprocessed text was also provided to the residents as the basis for their history recording to ensure fair comparison.

Data analysis

Data analysis was performed using SPSS 26.0 statistical software. First, paired *t*-tests was used to compare the time taken by residents and ChatGPT-4o to complete the records, assessing the statistical significance of any time differences. Wilcoxon signed-rank tests will be used to evaluate the quality differences between the two groups. Descriptive statistics will include means and standard deviations, and Kappa coefficients was used to analyze the consistency between the two attending physicians' scores to ensure the reliability and repeatability of the results. Additionally, in-depth analysis was conducted on items with significant differences, such as chief complaint and overall requirements, to identify specific aspects where ChatGPT-4o might need improvement.

Ethical considerations

The study received IRB approval from Ruijin hospital's ethics committee (approval number: 2024-443). Written informed consent was obtained from all participants prior to their participation in this study, ensuring that participation is voluntary and that participants are fully informed. All patient information collected during the study was kept confidential and anonymized, used solely for research purposes.

Results

Comparison of medical record quality scores

Statistical analysis of the 65 cases was conducted to compare the performance of the resident group and the ChatGPT-4o group in each scoring category. The results showed in Table 2.

Overall, the quality scores revealed that the resident and ChatGPT-4o groups performed similarly in several categories, with no significant differences between the groups. Specifically, no significant differences were found in the following categories: present illness, past medical history, personal history, family history, physical examination, auxiliary examination, and diagnosis (*p*-values: 0.42, 0.22, 0.26, 0.49, 0.27, 0.49, and 0.97, respectively) (Figure 1).

TABLE 2 Summary of comparative analysis across all evaluation metrics.

Scoring category	Resident group Mean ± SD	ChatGPT-4o group Mean ± SD	*p*-value
Chief complaint	5.70 ± 0.27	5.50 ± 0.38	0.009*
Overall requirements	4.48 ± 0.33	4.31 ± 0.41	0.041*
Present illness	28.64 ± 1.14	28.42 ± 1.55	0.42
Past medical history	9.52 ± 0.54	9.65 ± 0.48	0.22
Personal history	9.42 ± 0.63	9.53 ± 0.57	0.26
Family history	4.83 ± 0.23	4.87 ± 0.20	0.49
Physical examination	19.25 ± 0.84	19.08 ± 0.93	0.27
Auxiliary examination	9.78 ± 0.26	9.81 ± 0.24	0.49
Diagnosis	3.75 ± 0.27	3.75 ± 0.29	0.97
Total	**89.57 ± 2.66**	**88.94 ± 3.13**	**0.25**

*Indicates $p < 0.05$, statistically significant difference. The Kappa coefficient between the two evaluators was 0.82.

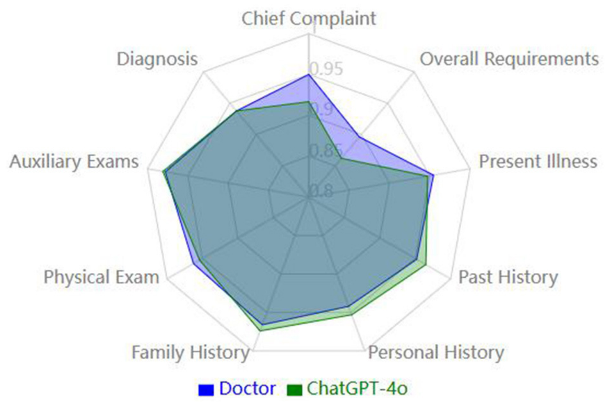


FIGURE 1
Comparison of medical history quality metrics (normalized).

However, in the "chief complaint" and "overall requirements" categories, the resident group scored significantly higher than the ChatGPT-4o group. In the "chief complaint" category, the resident group's mean score was 5.70 ± 0.27, while the ChatGPT-4o group's score was 5.50 ± 0.38, with a statistically significant difference ($p = 0.009$). In the "overall requirements" category, the resident group scored 4.48 ± 0.33 on average, while the ChatGPT-4o group scored 4.31 ± 0.41, which also showed a statistically significant difference ($p = 0.041$) (Figure 2).

In terms of total score across all categories, the resident group scored 89.57 ± 2.66, while the ChatGPT-4o group scored 88.94 ± 3.13. Paired *t*-test analysis showed no statistically significant difference between the two groups' total scores ($p = 0.25$), indicating that the overall quality of medical record documentation was comparable between the two groups.

10.3389/fmed.2025.1545730

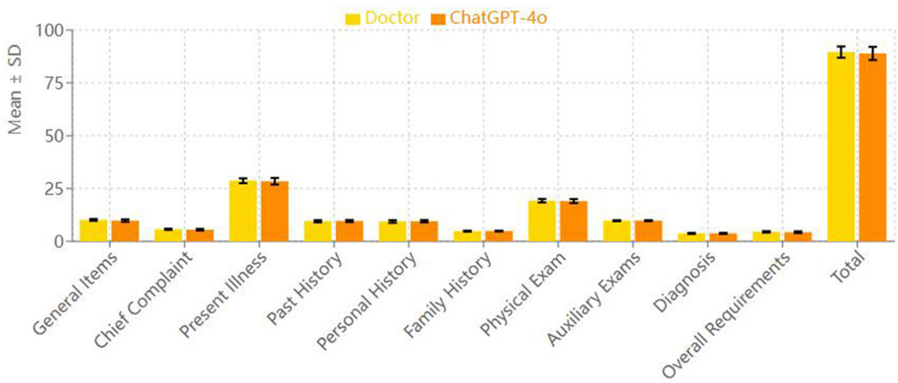


FIGURE 2
Comparison of mean scores (doctor vs. ChatGPT-4o).

Non-inferiority comparison of medical record quality

To assess whether ChatGPT-4o's performance in medical record quality was not inferior to that of the resident group, a non-inferiority analysis was conducted. The non-inferiority margin (Δ) was set at 5 points, meaning a difference of <5 points would indicate that ChatGPT-4o's performance was not inferior to the residents. The average total score for the resident group was 89.57, while the ChatGPT-4o group's average score was 88.94. The mean difference between the groups was 0.63 points, well below the non-inferiority margin ($\Delta = 5$). The non-inferiority test results indicated that the quality score for ChatGPT-4o fell within the pre-established non-inferiority margin ($p > 0.05$), confirming that ChatGPT-4o's performance in medical record quality was not inferior to that of the residents.

Comparison of medical record documentation time

The comparison of documentation time between the resident group and the ChatGPT-4o group showed that the resident group took an average of 893.2 seconds (~14.9 min) to complete the medical records, with a standard deviation of 28.0 s. In contrast, the ChatGPT-4o group completed the medical records in an average of 40.1 ± 4.4 s. Paired t-test analysis revealed that the time difference between the two groups was statistically significant ($p < 0.001$), indicating that ChatGPT-4o demonstrated significantly better efficiency in medical record documentation compared to the resident group.

Discussion

This study aims to assess the performance of ChatGPT-4o and resident physicians in terms of medical record efficiency and quality (16). The results indicate that while ChatGPT-4o maintains a comparable quality of medical records to the residents, it significantly reduces the time required for documentation. Specifically, ChatGPT-4o required only 40 s on average, whereas the resident physicians took ~15 min. This difference was statistically significant, highlighting ChatGPT-4o's clear advantage in time efficiency. However, it is important to note that the time required to process dialogue and correct transcription errors from speech recognition before generating the final record should also be considered. Improved speech recognition technology will be crucial for directly transcribing consultation processes into medical records through AI systems.

Although ChatGPT-4o demonstrated remarkable time efficiency, its quality scores were comparable to those of the residents. No significant differences were observed between the two groups in present illness, past medical history, personal history, family history, physical examination, auxiliary examinations, and diagnosis. However, in the "chief complaint" and "overall requirements" categories, the resident group scored significantly higher than the ChatGPT-4o group ($p = 0.009$ and $p = 0.041$, respectively). This suggests that, in these specific dimensions of medical record documentation, the residents performed better. These areas are more dependent on language proficiency and writing skills, and it is expected that AI models, including ChatGPT, may face some challenges in language generation, especially in non-native languages like Chinese.

From the perspective of non-inferiority analysis, although the residents scored slightly higher on certain items, ChatGPT-4o did not perform worse overall in terms of medical record quality. There was no statistically significant difference in total scores ($p = 0.25$), and the average difference between the groups was much smaller than the pre-set non-inferiority margin ($\Delta = 5$ points). This suggests that ChatGPT-4o can achieve a level of record quality similar to that of the resident physicians.

This finding holds significant clinical implications in the context of healthcare settings with heavy physician workloads (17). The high efficiency of ChatGPT-4o in record-keeping means it can alleviate physicians' burden while maintaining the quality of medical records, offering considerable potential to improve the overall efficiency of the healthcare system. ChatGPT-4o could be widely applied in various clinical settings, especially

10.3389/fmed.2025.1545730

in time-sensitive environments like emergency departments and intensive care units, where quick and efficient record support is critical. Additionally, in primary care settings, particularly in areas lacking experienced physicians, ChatGPT-4o could assist junior doctors in completing high-quality medical records, thus improving the quality of medical services.

However, despite the excellent performance of ChatGPT-4o, its clinical application faces several ethical challenges (18–20). Medical records involve sensitive patient information, and ensuring data security and privacy protection is a critical concern. Furthermore, over-reliance on AI could potentially diminish physicians' clinical reasoning abilities, thus impacting overall medical decision-making. Therefore, a balance must be struck between the use of technology and physician involvement to ensure clinical judgment is not compromised. Moreover, ethical review in medical record-keeping should ensure patient informed consent and clearly define the scope of data usage. Additionally, maintaining the model's focus and consistency remains a challenge in practical applications.

The limitations of this study include a small sample size, the focus on the hematology field, and the inability of the study design to cover all potential clinical complexities (21). In terms of sample selection, this study involved only five residents from a single specialty (hematology), which may limit the generalizability of the results. Future research should expand the sample size and explore the performance of ChatGPT-4o in other specialties. Each resident's background and experience level may influence their recording capabilities, and despite our attempt to minimize these differences through pre-assessment, selection bias may still exist. Additionally, there may be subjectivity in the standardization and scoring process, and while we attempted to reduce this through clear scoring criteria and independent scoring by two evaluators, the subjectivity of scoring remains inevitable. All clinicians in this study were from Ruijin Hospital, which may also limit the geographical representativeness of the results. Moreover, it is important to evaluate ChatGPT-4o's ability to handle complex cases and rare conditions, which would help comprehensively assess its applicability in clinical practice.

One promising research direction could involve integrating ChatGPT-4o with other AI systems, such as image recognition and retrieval-augmented generation (RAG) technologies, to create a multimodal clinical decision support system. This system could not only optimize medical record documentation but also provide real-time diagnostic suggestions and treatment plans. Such an integrated system would be particularly effective in assisting physicians with decision-making, especially in complex or rare cases.

Conclusion

This study provides strong evidence for the application of AI in medical history documentation, demonstrating the potential of ChatGPT-4o to improve clinical efficiency while maintaining medical history quality. As technology continues to develop, ChatGPT-4o or similar AI systems are expected to play a broader role in the healthcare field. However, how to maintain medical ethics and doctors' clinical abilities while applying these technologies will remain an ongoing and important issue.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving humans were approved by Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

XLu: Investigation, Writing – original draft. XG: Formal analysis, Investigation, Writing – original draft. XW: Data curation, Formal analysis, Writing – review & editing. ZG: Resources, Writing – review & editing. JC: Resources, Supervision, Writing – review & editing. WH: Resources, Supervision, Writing – review & editing. SW: Formal analysis, Software, Writing – review & editing. RW: Writing – review & editing. XLi: Data curation, Formal analysis, Funding acquisition, Resources, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Shanghai Jiao Tong University School of Medicine Postgraduate Medical Education Program (BYH20230315). Shanghai Science Technology and Innovation Action Plan Key Program on Medical Innovation Research (21Y31920400). Shanghai Shenkang Hospital Development Center's Project (SHDC12024141).

Acknowledgments

This manuscript details a research study that utilized AI chatbots as a key component of its investigations. Specifically, it features the application of ChatGPT version GPT-4o, a development by OpenAI. We acknowledge that all authors had access to the relevant data and contributed significantly to the composition of this manuscript. We acknowledge the support from the Shanghai Oriental Talent Plan Leading Program.

Conflict of interest

SW was employed by the WORK Medical Technology Group LTD.

The remaining authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. This manuscript details a research study that utilized AI chatbots as a key component of its investigations. Specifically, it features the application of ChatGPT version GPT-4o, a development by OpenAI.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2025.1545730/full#supplementary-material

References

1. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Sys*. (2020) 33:1877–901.

2. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008:128–44. doi: 10.1055/s-0038-1638592

3. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. (2016) 3:160035. doi: 10.1038/sdata.2016.35

4. Rotenstein LS, Torre M, Ramos MA, Rosales RC, Guille C, Sen S, et al. Prevalence of burnout among physicians: a systematic review. *JAMA*. (2018) 320:1131–50. doi: 10.1001/jama.2018.12777

5. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. (2017) 2:230–43. doi: 10.1136/svn-2017-000101

6. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT Performs on the Chinese National Medical Licensing Examination. *J Med Syst*. (2023) 47:86. doi: 10.1007/s10916-023-01961-0

7. Jin H, Lin Q, Lu J, Hu C, Lu B, Jiang N, et al. Evaluating the effectiveness of a generative pretrained transformer-based dietary recommendation system in managing potassium intake for hemodialysis patients. *J Ren Nutr*. (2024) 12:S1051-2276(24)00059-1. doi: 10.1053/j.jrn.2024.04.001

8. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. (2022) 28:31–8. doi: 10.1038/s41591-021-01614-0

9. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. guide to deep learning in healthcare. *Nat Med*. (2019) 25:24–9. doi: 10.1038/s41591-018-0316-z

10. OpenAI, "Hello GPT-4o." (2024). Available online at: https://openai.com/index/hello-gpt-4o/ (accessed November 30, 2024).

11. Mondillo G, Frattolillo V, Colosimo S, Perrotta A, Di Sessa A, Guarino S, et al. Basal knowledge in the field of pediatric nephrology and its enhancement following specific training of ChatGPT-4 "omni" and Gemini 1.5 Flash. *Pediatr Nephrol*. (2024) 40:151–157. doi: 10.1007/s00467-024-06486-3

12. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. (2018) 2:719–31. doi: 10.1038/s41551-018-0305-z

13. Angeli F, Verdecchia P, Vaudo G, Masnaghetti S, Reboldi G. Optimal use of the non-inferiority trial design. *Pharmaceut Med*. (2020) 34:159–65. doi: 10.1007/s40290-020-00334-z

14. Mauri L, D'Agostino RB Sr. Challenges in the design and interpretation of noninferiority trials. *N Engl J Med*. (2017) 377:1357–67. doi: 10.1056/NEJMra1510063

15. Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. *Trials*. (2011) 12:106. doi: 10.1186/1745-6215-12-106

16. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension *Lancet Digit Health*. (2020) 2:e537–48. doi: 10.1136/bmj.m3164

17. Patel VL, Shortliffe EH, Stefanelli M, Szolovits P, Berthold MR, Bellazzi R, et al. The coming of age of artificial intelligence in medicine. *Artif Intell Med*. (2009) 46:5–17. doi: 10.1016/j.artmed.2008.07.017

18. Elendu C, Amaechi DC, Elendu TC, Jingwa KA, Okoye OK, John Okah M, et al. Ethical implications of AI and robotics in healthcare: a review. *Medicine*. (2023) 102:e36671. doi: 10.1097/MD.0000000000036671

19. Nguyen A, Ngo HN, Hong Y, Dang B, Nguyen BT. Ethical principles for artificial intelligence in education. *Educ Inf Technol*. (2023) 28:4221–41.

20. Zhu J. AI ethics with Chinese characteristics? Concerns and preferred solutions in Chinese academia. *AI Soc*. (2022) 17:1–14. doi: 10.1007/s00146-022-01578-w

21. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*. (2021) 11:e047709. doi: 10.1136/bmjopen-2020-047709

frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 30 May 2025
DOI 10.3389/fdgth.2025.1497130

Check for updates

OPEN ACCESS

EDITED BY
Ariel Soares Teles,
Federal Institute of Education, Science and Technology of Maranhão, Brazil

REVIEWED BY
Luciano Reis Coutinho,
Universidade Federal do Maranhão, Brazil
Thomas F. Heston,
University of Washington, United States

*CORRESPONDENCE
Lifeng Han
l.han@lumc.nl;
lifeng.han@manchester.ac.uk

RECEIVED 16 September 2024
ACCEPTED 31 March 2025
PUBLISHED 30 May 2025

CITATION
Ren L, Belkadi S, Han L, Del-Pinto W and Nenadic G (2025) Synthetic4Health: generating annotated synthetic clinical letters.
Front. Digit. Health 7:1497130.
doi: 10.3389/fdgth.2025.1497130

Synthetic4Health: generating annotated synthetic clinical letters

Libo Ren[1], Samuel Belkadi[2], Lifeng Han[1,3]*, Warren Del-Pinto[1] and Goran Nenadic[1]

[1]Department of Computer Science, University of Manchester, Greater Manchester, Manchester, United Kingdom, [2]Department of Engineering, University of Cambridge, Cambridge, United Kingdom, [3]Leiden Institute of Advanced Computer Science (LIACS) and Leiden University Medical Center (LUMC), Leiden University, Leiden, Netherlands

Clinical letters contain sensitive information, limiting their use in model training, medical research, and education. This study aims to generate reliable, diverse, and de-identified synthetic clinical letters to support these tasks. We investigated multiple pre-trained language models for text masking and generation, focusing on Bio_ClinicalBERT, and applied different masking strategies. Evaluation included qualitative and quantitative assessments, downstream named entity recognition (NER) tasks, and clinically focused evaluations using BioGPT and GPT-3.5-turbo. The experiments show: (1) encoder-only models perform better than encoder–decoder models; (2) models trained on general corpora perform comparably to clinical-domain models if clinical entities are preserved; (3) preserving clinical entities and document structure aligns with the task objectives; (4) Masking strategies have a noticeable impact on the quality of synthetic clinical letters: masking stopwords has a positive impact, while masking nouns or verbs has a negative effect; (5) The BERTScore should be the primary quantitative evaluation metric, with other metrics serving as supplementary references; (6) Contextual information has only a limited effect on the models' understanding, suggesting that synthetic letters can effectively substitute real ones in downstream NER tasks; (7) Although the model occasionally generates hallucinated content, it appears to have little effect on overall clinical performance. Unlike previous research, which primarily focuses on reconstructing original letters by training language models, this paper provides a foundational framework for generating diverse, de-identified clinical letters. It offers a direction for utilizing the model to process real-world clinical letters, thereby helping to expand datasets in the clinical domain. Our codes and trained models are available at https://github.com/HECTA-UoM/Synthetic4Health.

KEYWORDS

pre-trained language models (PLMs), encoder-only models, encoder–decoder models, named entity recognition, masking and generating, synthetic data creation, clinical NLP (natural language processing)

1 Introduction

With the development of medical information systems, electronic clinical letters are increasingly used in communication between healthcare departments. These clinical letters typically contain detailed information about patients' visits, including their symptoms, medical history, medications, etc. (1). They also often include sensitive personal information, such as patients' names, phone numbers, and addresses (2, 3). As

 10.3389/fdgth.2025.1497130

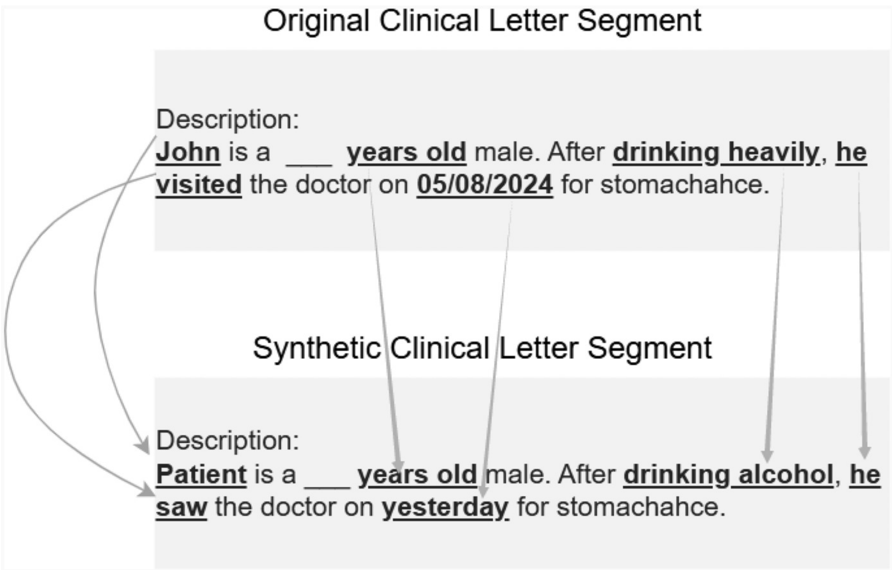


FIGURE 1
An example of the objective: sentence/segment-level generations.

a result, these letters are difficult to share and nearly impossible to use widely in clinical education and research.

In 2018, 325 severe breaches of protected health information were reported by CynergisTek (4) placing nearly 3,620,000 patients' records at risk (4). This data reflects just 1 year, and similar privacy breaches are unfortunately common. The most severe hacking incident affected up to 16,612,985 patients (4). Therefore, generating synthetic letters and applying de-identification techniques seem indispensable.

Additionally, due to privacy concerns and access controls, insufficient data remains a major challenge in clinical education, medical research, and healthcare system development (5). Some shared datasets offer de-identified annotated data, with the MIMIC series being a typical example. These datasets are accessible through PhysioNet. MIMIC-IV (6–8), the latest version, contains clinical data from 364,627 patients, collected from 2008 to 2019 at a medical center in Boston. It contains details about hospitalizations, demographics, and transfers. Numerous research studies have been conducted using this shared dataset. Another public dataset series in the clinical domain is i2b2/n2c2 (9), which is accessible through the DBMI Data Portal. This series includes unstructured clinical notes, such as process notes, radiology reports, and discharge summaries and is published for clinical informatics sharing and natural language processing (NLP) task challenges.

However, these shared datasets are often limited to specific regions and institutions, making them not comprehensive. Consequently, models and medical research outcomes derived from these datasets cannot be widely applied (10). Therefore, to address the lack of clinical datasets and reduce the workload for clinicians, it is essential to explore available technologies that can automatically generate de-identified clinical letters.

Existing systems generate clinical letters primarily by integrating structured data; however, there are not many studies that explore the use of natural language generation (NLG) models for this purpose (11–13). NLG attempts to combine clinical knowledge with general linguistic expressions to generate clinical letters that are both readable and medically accurate. However, NLG technology is not yet mature enough for widespread use in healthcare systems. Additionally, it faces numerous challenges, including medical accuracy, format normalization, and de-identification (12). Therefore, this investigation focuses on how NLG technology can be used to generate reliable and anonymous clinical letters, which can benefit medical research, clinical education, and clinical decision-making.

The main aim of our work is to *generate de-identified clinical letters* that can *preserve clinical information* while *differing from the original* letters. A brief example of our objective is shown in Figure 1. Based on this objective, different generation models are explored as a preliminary attempt. Then, the best models are selected and various techniques are tested to improve the quality of the synthetic letters. The synthetic letters are evaluated not only with quantitative and qualitative methods but also in downstream tasks, i.e., NER. We hope this work contributes to addressing the challenge of insufficient data in the clinical domain.

In summary, this work is centered on the research question (RQ): "How can we generate reliable and diverse clinical letters without including sensitive information?" Specifically, it answers the following related sub-questions (RQs)[1]:

[1]We report our extended solid investigation and outcomes based on our preliminary workshop paper findings (96).

1. How do different models perform in masking and generating clinical letters?
2. How should the text be segmented in clinical letter generation?
3. How do different masking strategies affect the quality of synthetic letters?
4. How can we evaluate the quality of synthetic letters?

To answer these questions, we explored various large language models (LLMs) for masking and generating clinical letters, ultimately focusing on one that performed well. The overall highlights of this work are summarized as follows:

1. Mask and generate clinical letters using different LLMs at the sentence level.
2. Explore methods to improve synthetic clinical letters' readability and clinical soundness.
3. Initially evaluate synthetic letters using both qualitative and quantitative methods.
4. Apply synthetic letters in downstream tasks and further evaluate them using clinically focused methods.
5. Explore post-processing methods to further enhance the quality of de-identified letters.

2 Background and literature review

We first introduce general language models, followed by their applications, especially within the clinical domain. We then present the generative language models based on the transformer architecture. These models serve as the technical foundation for most modern text generation tasks. Afterward, we review related works, discussing their relevance and connections to our work. Finally, all quantitative evaluation metrics used in this paper are introduced.

2.1 Development of language models (LMs)

The development of language models can be divided into three stages: rule-based approach, supervised modeling, and unsupervised modeling (14).

2.1.1 Rule-based approach

The rule-based approach, first used in the 1950s, marks the beginning of NLP (15). This approach relies on a set of predefined rules, which were written and maintained manually by specialists (16, 17). Although it can generate standardized text without being fed with extensive input data (17), it has numerous limitations. Initially, manually crafted rules are often ambiguous, and the dependencies between different rules increase the cost of maintenance (15). Second, these stylized models cannot perform well in understanding realistic oral English and ungrammatical text, such as clinical discharge records, although these texts are still readable to humans (15). Third, they are not objective enough, as they are affected by the editors of the rule library. Additionally, they are not flexible enough to deal with special cases. Therefore, the rule-based method is only suitable for analyzing and generating highly standardized texts like prescriptions (17).

2.1.2 Supervised language models

To address the limitations of the rule-based approach, supervised learning has been applied to NLP. The invention of statistical machine translation (SMT) in 1990 marked the rise of supervised NLP (14). It learns the correspondence rules between different languages by analyzing the input of bilingual texts (parallel corpora) (18). Supervised NLP models are trained on annotated labels to learn rules automatically. The learned rules will be used in word prediction or text classification. Hidden Markov model (HMM) and conditional random field (CRF) are two typical applications of this stage (19). Both of them work by tagging features of the input texts. HMM generates data by statistically analyzing word frequencies (20, 21). CRF, however, searches globally and calculates joint probabilities to get an optimal solution (22, 23). Long short-term memory (LSTM) is another typical example of supervised language modeling (24). In text generation tasks, the input consists of a set of labeled data or word vector sequences. By minimizing the loss between the predicted word vector and the actual word vector, LSTM can capture the dependencies between words in long texts (25, 26).

Although supervised language models perform better than the rule-based approach, domain experts still need to annotate the training dataset (14). In addition, collecting data in some domains is difficult due to privacy issues (such as *medical* and *legal* domains). This became an ongoing challenge in applying the supervised language models to specific tasks.

2.1.3 Unsupervised language models

To address the high cost and difficulty of obtaining labeled data, unsupervised neural networks are applied to the language modeling (27). The popularity of corpora such as Wikipedia and social media provides enough data for training unsupervised models (14). Word embedding is a significant technique in this stage (28). For example, Word2Vec represents words using vectors with hundreds of dimensions. The context can be captured by training word vectors in a sliding window. By adjusting hyperparameters to maximize the conditional probability of the target word, the model can learn semantic information accurately (29, 30) [e.g., "Beijing"-"China"+"America" => "Washington" (31)]. After training, each word usually has a fixed word vector regardless of the context in which it appears (known as static word embedding) (26).

Unlike Word2Vec, BERT and GPT use contextual word embeddings, meaning that their word vectors reflect the semantic information and are affected by the context (32). BERT focuses on contextual understanding (33) (e.g., in the sentence "The bank is full of lush willows," the word "bank" refers to a riverside rather than a financial institution). In contrast, GPT models focus on text generation within a specific context (34, 35) (e.g., Prompt: "Do you know Big Ben?" Answer: "Yes, I know Big Ben. It is the nickname for the Great Bell of the Clock located in London."). Although unsupervised language models have been able to train and understand text proficiently, they still face challenges in practical applications, such as difficulty handling ambiguity and high computing resource consumption. Therefore, language modeling still has a long way to go.

2.2 Language models applications in clinical domain

Based on the modeling methods mentioned above, a variety of language models have been developed. They play an important role in scientific research and daily life, especially in the field of healthcare. In this section, we discuss the *clinical language model* applications in detail from two aspects: NER and NLG.

2.2.1 Named entity recognition

NER was originally designed for text analysis and recognition of named entities, such as dates, organizations, and proper nouns (36). In the clinical domain, NER is used to identify *clinical events* (e.g., symptoms, drugs, treatment plans, etc.) from unstructured documents, along with their *qualifiers* (e.g., chronic, acute, mild), classify them, and extract the relationship between entities (37, 38). Earlier, NER systems relied on rule-based and machine learning methods that required extensive manual feature engineering. In 2011, Collobert et al. (39) used word embeddings and neural networks in NER. Since then, research in NER has shifted to automatic feature extraction.

spaCy[2] is an open-source NLP library used for tasks like POS tagging and text classification. Additionally, it offers a range of pre-trained NER models. ScispaCy,[3] a fine-tuned extension of spaCy on medical science datasets, can recognize entities such as "DISEASE," "CHEMICAL," and "CELL," which are essential for medical research. Although NER is useful in rapidly extracting clinical terms, several challenges remain, such as non-standardization (extensive use of abbreviated words in clinical texts), misspellings (due to manual input by medical staff), and ambiguity (often influenced by context, e.g., whether the word "back" refers to an adverb or an anatomical entity) (37). Existing research mitigates these problems using *entity linking* (mapping extracted clinical entities to medical repositories such as UMLS and SNOMED). More deep learning models and text analysis tools are being developed to solve these issues.

2.2.2 De-identification

The unprocessed clinical text poses a risk of personal information leakage. Additionally, manual de-identification is not only error-prone but also costly. Therefore, research on de-identification is indispensable for the secondary use of clinical data. Typically, de-identification is based on NER models to identify protected health information (PHI). Then, PHI is processed by different strategies (such as synonym replacement, removal, or masking) (40, 41).

Similar to NER, early de-identification approaches relied heavily on rule-based systems, machine learning, or hybrid models. PhysioNet DeID, the VHA best-of-breed (BoB), and MITRE's MIST are three typical examples (42). However, these algorithms require extensive handcrafted feature engineering. With the development of unsupervised learning, recurrent neural networks (RNNs) and transformers are widely used in de-identification tasks (43, 44).

Philter, a protected health information filter (45), is a pioneering system that combines rule-based approaches with state-of-the-art NLP models to identify and remove PHI. Although Philter outperforms many existing tools like PhysioNet and Scrubber, particularly in terms of recall and F2 score, it still requires large amounts of annotated data for training (45). Additionally, research has shown that while the impact of de-identification on downstream tasks is minimal, it cannot be completely ignored (46). Therefore, performing de-identification without mistakenly removing semantic information is still a challenge in this field.

2.2.3 Natural language generation

Both label-to-text and text-to-text generation are components of NLG (47). NLG consists of six primary sub-tasks, covering most of the NLG process. NLG architectures can generally be divided into three categories (47):

- **Modular architectures:** This architecture consists of three modules: the text planner (responsible for determining the content for generation), the sentence planner (which aggregates the synthetic text), and the realizer (which generates grammatically correct sentences). These modules are closely related to the six sub-tasks, and each module operates independently.
- **Planning perspectives:** This architecture considers NLG as a planning problem. It generates tokens dynamically based on the objectives, with potential dependencies between different steps.
- **Integrated or global approaches:** Currently the dominant architecture for NLG, this approach relies on statistical learning and deep learning. Common generative models, such as transformers and conditional language models, are included in this architecture.

In the field of healthcare, NLG applications include document generation and question-answering. Document generation involves discharge letters, diagnostic reports for patients, decision-making suggestions for experts, and personalized patient profiles for administrators (48). Some systems have already been implemented in practice. For instance, PIGLIT generates explanations of clinical terminology for diabetes patients (49), while MAGIC can generate reports for intensive care unit (ICU) patients (50). Question answering is another application of NLG. Tools like chatbots can provide patients with answers to basic healthcare questions (51).

[2]https://spacy.io/

[3]https://allenai.github.io/scispacy/

10.3389/fdgth.2025.1497130

Nowadays, NLG in the clinical field focuses on the development and training of transformer-based LLMs; examples of this work can be seen in (11, 52). These models perform well in specific domains such as semantic query (53) and electronic health record (EHR) generation (54). However, very few systems can reliably produce concise, readable, and clinically sound reports across multiple sub-domains (48).

2.3 Generative language models

2.3.1 Transformer and attention mechanism

Although RNNs and LSTM networks are effective at capturing semantic understanding, their recursive structure not only prevents parallel computation but also makes them prone to gradient vanishing (55). The introduction of the transformer architecture in 2017 addressed this issue by replacing the recurrent structure with a multi-head attention mechanism (56). Since then, most deep learning models have been based on the transformer framework. Transformer architecture is based on an encoder–decoder model (56). To understand this, we first need to overview auto-regressive models and the multi-head attention mechanism.

Auto-regressive models' predictions for each auto-regressive model token depend on the previous output. Therefore, it can only access the preceding tokens and operate iteratively. When the input sequence is X, the auto-regressive model aims to train parameters θ to maximize the log-likelihood of the conditional probability P (Equation 1) (56)

$$L(X) = \sum_{i} \log P(x_i \mid x_{i-k}, \ldots, x_{i-1}; \Theta) \tag{1}$$

Multi-head attention mechanism: The attention mechanism was initially proposed by Cho et al. (57). It can not only focus on the element being processed but also capture the context dependence (56). The scaled dot-product attention is computed as shown in Equation 2. Multi-head attention consists of several single-head attention (scaled dot-product attention) layers (56). Each word in the input sequence is converted into a high-dimensional vector representing semantic information by word embedding. These vectors are then passed through linear transformation layers to generate vectors for queries (Q), keys (K), and values (V). For each word, Q, K, and V are inputs to this single-head attention layer. The importance score of this word is calculated, and V corresponding to this word is multiplied to get the output of this head (called attention). Finally, outputs from all layers are concatenated to form a larger vector, which is the input to a feed-forward neural network (also the output of the multi-head attention layer) (56)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Transformer and pre-training language models (PLMs): Transformer consists of an encoder and a decoder. The auto-regressive model is the basis of the decoder. When the input sequence is $X = (x_1, \ldots, x_N)$ and the output sequence is $Y_M = (y_1, \ldots, y_M)$, the model can learn a latent feature representation $Z = (z_1, \ldots, z_N)$ from X to Y. The generation of each new element Y_M relies on the generated sequence $Y_{M-1} = (y_1, \ldots, y_{M-1})$ and feature representation Z. Both the encoder and the decoder use the multi-head attention mechanism (55, 56).

Many modern models are based entirely or partially on the transformer. They compute general feature representations for the training set by unsupervised learning. This is the concept of PLMs. They can be fine-tuned to adapt to the specific tasks on particular datasets (34, 55).

2.3.2 Encoder-only models

Since the transformer's encoder architecture can effectively capture the semantic features, some models only use this part for training. They are applied in text understanding tasks, such as text classification and NER. Bidirectional encoder representations from transformers (BERT) (58) is a representative model among them.

Unlike the transformer decoder, which uses an auto-regressive model, BERT is trained based on the masked language model (MLM) (34). It masks the word in the input sequence and uses the bidirectional encoder to understand the context semantically, which will be used in predicting the masked word (58). It has already been pre-trained on a 16 GB corpus. To deploy it, we only need to replace the original fully connected layer with a new output layer and then fine-tune the parameters on the dataset for specific tasks (58). This approach consumes fewer computing resources and less time than training a model from scratch. In the clinical domain, Bio_ClinicalBERT (59) and medicalai/ClinicalBERT (60) are fine-tuned in the clinical dataset based on the BERT architecture. Initially, due to BERT's focus on semantic understanding, it was rarely used for text generation (61).

Robustly optimized BERT pretraining approach (RoBERTa) (62) improved some key hyperparameters based on BERT. Instead of BERT's static mask, it uses a dynamic mask strategy, which helps it better adapt to multitasking. Additionally, it gained a stronger semantic understanding after training on five English datasets of 160 GB. However, it was trained with more epochs and larger batch sizes compared to BERT, indicating higher computational resource requirements and longer training time (63).

To better handle long sequences, the Longformer introduces a sparse attention mechanism to reduce computation (64). This allows each token to focus only on nearby tokens rather than the entire sequence. Unlike traditional models like BERT and RoBERTa, which can only process no more than 512 tokens, the Longformer can handle up to 4,096 tokens. It consistently performs better than RoBERTa in downstream tasks involving long documents (64). The Clinical-Longformer model (65) was fine-tuned for the clinical domain.

Supplementary Table S1 summarizes the encoder-only models used in our work and their corresponding fine-tuning datasets.

2.3.3 Decoder-only models

In 2020, the performance of ChatGPT-3 (66) in question answering task caught researchers' attention to decoder-only architectures. As mentioned earlier, the transformer decoder is an auto-regressive model. It can only refer to the synthesized words on the left side to generate the new word, without considering the context (which is called masked self-attention). This method made it more flexible in generating coherent text. Compared with BERT, the GPT series performed well in zero-sample and small-sample learning tasks by enlarging the size of the model. Even without fine-tuning, a simple prompt can help GPT generate a reasonable answer (67).

Unlike GPT, which improves models' performance by increasing dataset size and the number of parameters without limitations, Meta AI published a series of Llama models. These models aim to maximize the use of limited resources - in other words, by extending training, they reduce the overall demand on computing resources. The latest Llama3 model requires only 8–70 billion parameters (68), significantly less than GPT-3's 175 billion (67). Additionally, it outperforms GPT-3.5 Turbo in five-shot learning (69).

2.3.4 Encoder–decoder models

T5 family (70) is a classic example of the encoder–decoder model. This architecture is particularly suitable for text generation tasks that require deep semantic understanding (71). T5 transforms all kinds of NLP tasks into a text-to-text format (72). Unlike BERT, which uses word-based masking and prediction, T5 processes text at the *fragment* level using "span corruption" to understand semantics (72). For the fill-in-the-blank task, instead of replacing the specific words with <mask> like BERT, T5 replaces the text fragments with an ordered set of <extra_id_n> to reassemble the long sequence text. T5 needs to pre-process the input text according to the task requirements. A directive prefix should be added as a prompt.

Some language models fine-tuned with T5 on specific datasets, such as SciFive (fine-tuned in some science literature) (73) and ClinicalT5 (fine-tuned in clinical dataset MIMIC-III notes) (74), have shown excellent performance in their respective fields. The T5 family models used in this paper and their corresponding fine-tuned datasets are summarized in Supplementary Table S2.

2.3.5 Comparison and limitations

According to Cai et al. (71), the encoder–decoder architecture performs best with sufficient training data. However, challenges in data collection can negatively affect its performance. Despite these challenges, different architectures are well-suited to different tasks. For example, for tasks requiring semantic understanding, such as text summarization, the encoder–decoder architecture is the most effective. In contrast, for tasks that involve minor word modifications, the encoder-only structure works better. However, the decoder-only structure is not suitable for tasks with insufficient training data and long text processing, but performs well in few-shot question answering tasks (71, 75).

Following these discussions, transformer-based PLMs have demonstrated strong performance in NLP tasks, but many challenges still remain.

2.4 Related works on clinical text generation

2.4.1 LT3: label to text generation

LT3 (76) adopts an encoder–decoder architecture to generate synthetic text from labels. As shown in Supplementary Figure S1, labels such as medications are the input of the encoder, which can generate corresponding feature representations. The decoder generates prescription sequences based on these features. The pre-trained BERT tokeniser is used to split the input sequence into sub-words. LT3 is trained from scratch. Instead of using traditional greedy decoding, which may miss the global optimum, the authors proposed beam search decoding with backtracking (B2SD). This approach broadens the search range through a backtracking mechanism, preserving possible candidates for the optimal solution. To reduce time complexity, they used a probability difference function to avoid searching for low-probability words. Additionally, the algorithm penalizes repeated sub-sequences and employs a logarithmic heuristic to guide the exploration of generation paths. The authors test LT3 on the 2018-n2c2 dataset and evaluate the results using both quantitative metrics and downstream tasks. It was demonstrated that this model outperforms T5 in label-to-text generation.[4]

2.4.2 Seq2Seq generation for medical dataset augmentation

Amin-Nejad et al. (75) compared the performance of the Vanilla transformer and GPT-2 using the MIMIC-III dataset in seq2seq tasks. Specifically, they fed as input a series of structured patient information as conditions, as shown in Supplementary Figure S2, to generate discharge summaries. They demonstrated that the augmented data outperforms the original data in downstream tasks (e.g., readmission prediction). Furthermore, they proved that the Vanilla transformer performs better with large samples, while GPT-2 excels in few-shot scenarios. However, GPT-2 is not suitable for augmenting long texts. Additionally, they used Bio_ClinicalBERT for the downstream

[4]LT3 achieved significant improvements over the best-performing T5 model (T5 base) in label-to-text generation, achieving improvements of up to 6.5 BLEU points and 0.02 in the BERTScore. Unfortunately, when we tried applying B2SD to generate clinical letters, the results were somehow disappointing. This may be due to the length of clinical letters. B2SD consumes a lot of time on long text generation. Despite this, it still shows great potential in generating clinical data.

tasks and discovered that Bio_ClinicalBERT outperformed the baseline model (BERT) in almost all experiments. This suggests that Bio_ClinicalBERT can potentially replace BERT in the biomedical field. Interestingly, although the synthetic data have a low score on internal metrics (such as ROUGE and BLEU), the performance on downstream tasks is notably enhanced. This may be because augmenting text can effectively introduce noise into the original text, improving the model's generalization to unseen data.

According to their findings, decoder-only models like GPT-2 are not suitable for processing long texts. Bio_ClinicalBERT is particularly effective for tasks in the clinical area, and the Clinical transformer is promising in augmenting medical data. This provides more possibilities for our task of generating synthetic clinical letters.

2.4.3 Discharge summary generation using clinical guidelines and human evaluation framework

Unlike the traditional supervised learning of fine-tuning language models (which requires a large amount of annotated data), Ellershaw et al. (77) generated 53 discharge summaries using only a one-shot example and a clinical guideline. Their research consists of two aspects: generating discharge summaries and a manual evaluation framework.

As shown in Supplementary Figure S3, the authors used clinical notes from MIMIC-III as input and incorporated a one-shot summary along with clinical guidance as prompts to generate discharge summaries by GPT-4-turbo. Initially, five sample synthetic summaries were evaluated by a clinician. Based on the feedback, the clinical guidance was revised to adapt to the generation task. Through iterative optimization, the revised guidance, combined with the original one-shot sample, became the new prompt. Then, the authors generated 53 discharge summaries using this method and invited 11 clinicians to do a final manual quantitative evaluation. Clinicians were invited to evaluate the error rate at the section level (e.g., diagnoses, social context, etc.). It includes four dimensions:

- Minor omissions,
- Severe omissions,
- Unnecessary text, and
- Incorrect additional text.

Each discharge summary was evaluated by at least two clinicians, and the authors calculated agreement scores to evaluate the subjectivity during the human evaluation stage. Unfortunately, the inter-rater agreement was only 59.72%, raising concerns that the revised prompts based on such feedback might result in subjective synthetic summaries. Although this study partially addresses the issue of insufficient training data and provides reliable human quantitative evaluation methods, it is still not well-suited for our investigation. Specifically, it required 11 clinicians to evaluate 53 synthetic samples, demonstrating the considerable time and manpower required. Therefore, there is still a long way to go before this technique can be used for large-scale text generation tasks.

2.4.4 Comparison of masked and causal language modeling for text generation

Micheletti et al. (78) compared masked language modeling (MLM, including BERT, RoBERTa, BiomedNLP-PubMedBERT) and causal language modeling (CLM, including T5, BART, SciFive-large-Pubmed_PMC) across various datasets for masking and text generation tasks. They used qualitative and quantitative evaluations, as well as downstream tasks, to assess the quality of the synthetic texts. Their workflow is shown in Supplementary Figure S4. Based on these evaluations, the study yielded the following results:

- MLM models are better suited for text masking and generation tasks than CLM.
- Introducing domain-specific knowledge does not consistently improve model performance.
- Downstream tasks can adapt to the introduced noise. Although some synthetic texts might not achieve highly quantitative evaluation scores, they can still perform well in downstream tasks. This matches the findings from Amin-Nejad et al. (75).
- A lower random masking ratio (i.e., masked tokens/total tokens) can generate higher-quality synthetic texts.

These very recent findings provide insightful inspiration to our investigation. Our work builds on their research, expanding on masking strategies and focusing on the clinical domain.

3 Methodologies and experimental design

Due to the sensitivity of clinical information, many clinical datasets are not accessible. As mentioned in Section 2, numerous studies use NLG techniques to generate clinical letters and evaluate the feasibility of replacing the original raw clinical letters with synthetic letters. Most existing research involves fine-tuning PLMs or training transformer-based models from scratch on their datasets through supervised learning. These studies explore different ways to learn mapping from the original raw text to synthetic text and work on generating synthetic data that are similar (or even identical) to the original ones. Our work, however, aims to find a method that can generate clinical letters that can *keep the original clinical story, while not exactly being the same as the original letters*. To achieve this objective, we employed various models and masking strategies to generate clinical letters. The experiment follows these steps:

1. **Data collection and pre-processing**: We first accessed clinical letter examples (6–8) for an overview. The texts were segmented at the sentence level, and clinical entities and structural templates were extracted to capture the clinical narratives while maintaining clinical soundness.
2. **Randomly masking**: We randomly masked the context and generated clinical letters by predicting masked tokens using different LLMs.

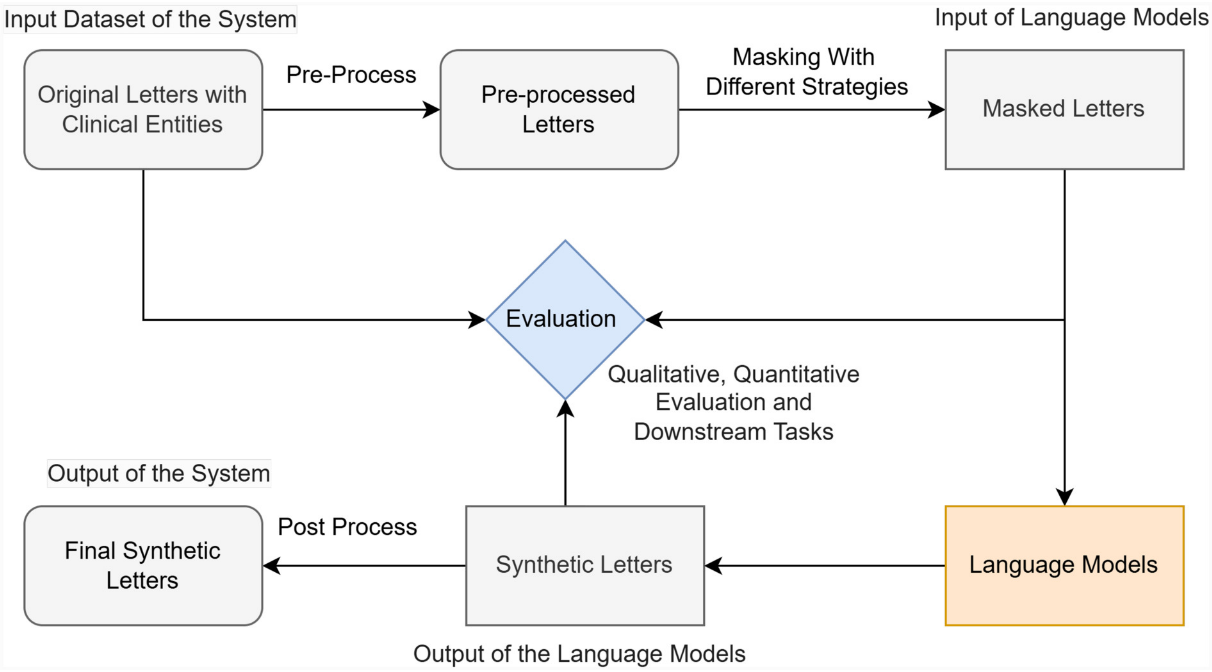


FIGURE 2
Overall investigation workflow for SYNTHETIC4HEALTH.

3. **Model evaluation**: We evaluated synthetic letters generated by different language models. Based on their performance, we selected Bio_ClinicalBERT and worked on it.
4. **Masking strategy exploration**: We explored multiple masking strategies to retain clinical stories and diversity while removing private information. After generating clinical letters using these strategies, we evaluated their quality.
5. **Post-processing**: We applied post-processing techniques to further enhance the readability of synthetic letters.
6. **Downstream task evaluation**: We compared the performance of synthetic and original letters in a downstream NER task to evaluate the usability of these synthetic letters.

An overall investigation workflow is shown in Figure 2.

3.1 Dataset

Based on the objective of this project, we need a dataset that includes both clinical notes and some clinical entities. The dataset we used was from the SNOMED CT Entity Linking Challenge (6–8). It includes 204 clinical letters and 51,574 manually annotated clinical entities.

Clinical letters: The clinical letters were from a subset of discharge summaries in MIMIC-IV-Note (6, 79). It uses clinical notes obtained from a healthcare system in the United States. These notes were de-identified by a hybrid method involving the rule-based approach and neural networks. To avoid releasing sensitive data, the organization also did a manual review of PHI. In these letters, all PHI was replaced with three underscores "___." The letters record the patient's hospitalisation information (including the reason for visiting, consultation process, allergy history, discharge instructions, etc.). They are saved in a comma-separated value (CSV) format file "mimic-iv_notes_training_set.csv." Each row of data represents an individual clinical letter. It consists of two columns, where the "note_id" column is a unique identifier for each patient's clinical letter, and the "text" column contains the contents of the clinical letter. Since most language models have a limitation on the number of tokens to process (80), we tokenized the clinical letters into words using the "NLTK" library and found that all clinical letters contained thousands of tokens. Therefore, it is necessary to split each clinical letter into multiple chunks to process them. These separated chunks must be merged in the end to generate the whole letter.

Annotated clinical entities: The entities were manually annotated based on SNOMED CT. A total of 51,574 annotations cover 5,336 clinical concepts. They were saved in another CSV document which includes four columns: "note_id," "start," "end," and "concept_id." The "note_id" column corresponds to the "note_id" in the "mimic-iv_notes_training_set.csv" file. The "start" and "end" columns indicate the position of annotated entities. The "concept_id" can be used for entity linking with SNOMED CT. For example, for the "note_id" "10807423-DS-19," the annotated entity "No Known Allergies" has a corresponding "concept_id": "609328004." This can be linked to SNOMED CT under the concept of "Allergic disposition" (81).

An example of text excerpted from the original letter is shown in Supplementary Figure S5. It contains the document structure and some free text. According to the dataset, document structure often corresponds to capital letters and colons ":." Our primary goal is to mask the context that is neither part of the document structure nor annotated entities, and then generate a new letter, as both structure and clinical entities are essential for understanding clinical information (46).

3.2 Software and environment

All codes and experiments in this paper were carried out in the integrated development environment (IDE) "Google Colab Pro+" using a 52 GB system RAM and 225 GB disk space. The built-in T4 GPU (16 GB VRAM) accelerates the inference process. The primary tools used in the paper include:

- **Programming language and environment:** Python `3.10` serves as the main programming language.
- **Deep learning framework:** PyTorch `2.3.1` is the core framework used for loading and applying pre-trained language models (PLMs).
- **Natural language processing libraries:** This includes Hugging Face Transformers `4.42.4`, NLTK (version $\geq$ `3.1`), and BERTScore `0.3.13`, among others. These are popular tools for text processing and evaluation in the NLP domain.
- **Auxiliary tools:** Libraries such as pandas (version $\geq$ `1.0.1`) and mpmath (`1.1.0` $\leq$ version < `1.4`) can support data management, mathematical operations, and other routine tasks.

3.3 Pre-processing

The collected dataset involves different files and comprises entirely raw data. It is necessary to pre-process these files before using them in generation tasks. The pre-processing of this system contains five steps: "Merge dataset based on 'note_id,'" "Annotated Entity Recognition," "Split Letters in Chunks," "Word Tokenization," and "Feature Extraction." The pre-processing pipeline is shown in Figure 3.

3.3.1 Merging dataset and annotated entity recognition

Initially, we merged the clinical letters file and annotations file into a new DataFrame. After this, we extracted manually annotated entities based on their index. An excerpt from an original letter is shown in Supplementary Figure S6, and the manually annotated entities are listed in Supplementary Table S3.

3.3.2 Splitting letters into variable-length chunks

Typically, PLMs such as BERT, RoBERTa, and T5 have a limit on the number of input tokens, usually capped at 512 (82). When dealing with text that exceeds this limit, common approaches include discarding the excess tokens or splitting the text into fixed-length chunks of 512 tokens. In addition, some studies evaluate the tokens' importance to decide which parts should be discarded (83).

In this work, each clinical letter ("note_id") contains thousands of tokens, as mentioned in Section 3.1, to preserve as much critical clinical information as possible; therefore, we avoided simply discarding tokens. Instead, we adopted a splitting strategy based on semantics. Each block is not a fixed length. Rather, they are complete paragraphs that are as close as possible to the token limit. This approach aims to help the model better capture the

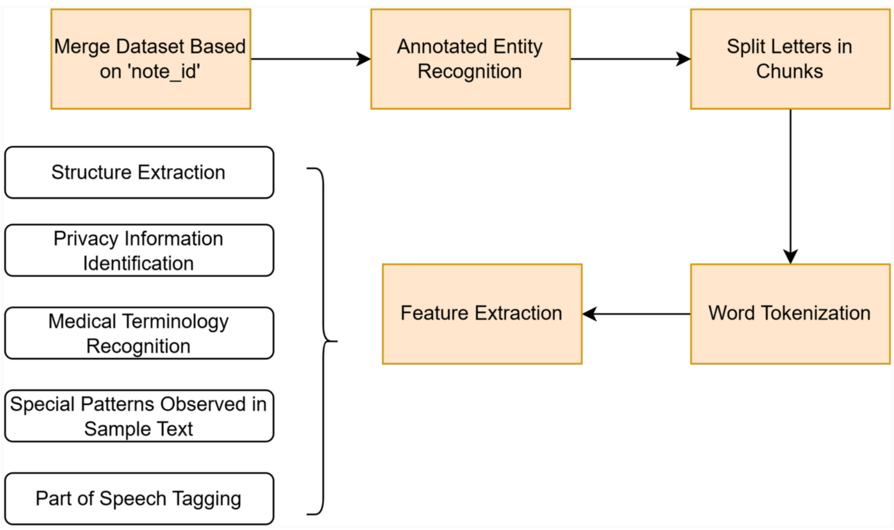


FIGURE 3
Pre-processing pipeline.

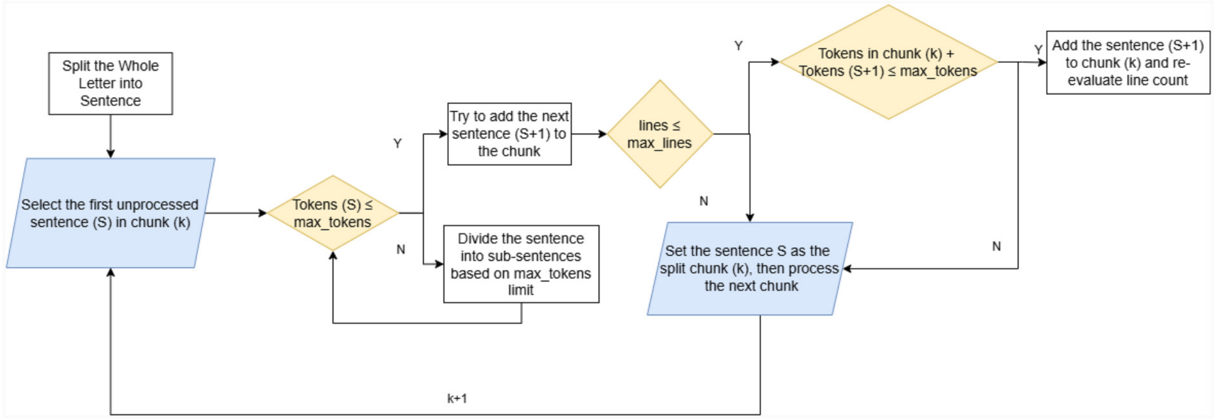


FIGURE 4
Text chunking workflow.

meaning and structure of clinical letters, thereby improving its ability to retain essential clinical information while efficiently processing the text. In fact, we initially generated letters at the sentence level. However, it was found that processing at the sentence level is not only time-consuming but also fails to provide the model with enough information for inference and prediction. This is why the letters were processed in chunks rather than in sentences.

As shown in Figure 4, each raw letter is split into sentences first. We used the pre-trained models provided by the "NLTK" library, which combines statistical and machine-learning approaches to identify sentence boundaries. Each clinical letter is treated as a separate processing unit, with the first sentence automatically assigned to the first text block (chunk). To control the length of each chunk, we set a maximum line count parameter (max_lines). If the first sentence already meets the value of "max_lines," the chunk will contain that single sentence only. Otherwise, subsequent sentences will be added to the chunk until the line count reaches the max_lines.

Extra care is needed when handling text with specific formats, such as medication dosage descriptions, as shown in Supplementary Figure S7. Because there is no clear sentence boundary, these sentences may exceed the tokens limitation. To address this, we first checked whether the sentence being processed exceeds the token limit (max_tokens). If it does not, the sentence will be added to the current chunk. Otherwise, the sentence should be split into smaller chunks, each no longer than "max_tokens." This operation helps balance processing efficiency while maintaining semantic integrity. In the example shown in Supplementary Figure S7, although using line breaks to split the text seems to be more flexible, considering time complexity and the requirement to index the annotated entities, this method was not chosen.

3.3.3 Word tokenization

To prepare the text for model processing, we split each chunk of text into smaller units: tokens. The tokenization methods can be categorised into two types: one for feature extraction and the other for masking and generation.

For the tokenization aimed at feature extraction, we used the "word_tokenise" method from the "NLTK" library. It is helpful to preserve the original features of the words, which is especially important for retaining clinical entities. For instance, in the sentence "Patient is a ___ yo male previously healthy presenting w/ fall from 6 ft, from ladder." Word boundaries such as spaces can be automatically detected for tokenization. The results of different tokenization methods are shown in the Supplementary Table S4.

As for the tokenization used for masking and generating, we retained the original models' tokenization methods. The specific tokenization approach varies by model, as shown in Supplementary Table S4. For example, BERT family models use word-piece tokenization, which initially splits text by spaces and then further divides the words into sub-words (62). This approach is particularly effective for handling words that are not in the pre-training vocabulary and is especially useful for predicting masked words. For complex clinical terms, however, these models rely heavily on a predefined dictionary, which can result in unsatisfactory tokenization and hinder the model's understanding. For instance, the word "COVID-19" is tokenized by BERT into ["co," "##vid," "–," "19"]. In contrast, the T5 family models use sentence-piece tokenization. It does not rely on space to split the text. Instead, this method tokenises directly from the raw text, making it better suited for handling abbreviations and non-standard characters (e.g., "COVID-19"), which are common in clinical letters.

It is important to note that although all BERT family models use word-piece tokenization, the results can still differ. This is because different models use different vocabularies during pre-training, leading to variations in tokenization granularity. The tokenization methods for each model are detailed in Supplementary Table S4. Each tokenization approach has its own advantages and disadvantages for processing clinical letters. Therefore, exploring how these models impact the clinical letter generation is also a requirement of our project.

3.3.4 Feature extraction

Since we aimed to generate de-identified clinical letters that can preserve clinical narratives during masking and generation, it is necessary to extract certain features beforehand. We extracted the following features, with an example provided in Supplementary Figure S8 and Supplementary Table S5.

- **Document structure:** This feature is identified by a rule-based approach. As mentioned in Section 3.1, structural elements (or templates) often correspond to the use of colons ":." They should not be masked to preserve the clinical context.
- **Privacy information identification:** In this part, we used a hybrid approach. To identify sensitive information such as "Name," "Date," and "Location (LOC)," we employed a NER toolkit from Stanza (84). To handle privacy information like phone numbers, postal codes, and e-mail addresses, we implemented a rule-based approach. Specifically, we devised several regular expressions to match the common formats of these data types. These pieces of private information should be masked.
- **Medical terminology recognition:** A NER toolkit pre-trained on the i2b2 dataset is used here (85). It can identify terms like "Test," "Treatment," and "Problem" in free text. Although our dataset has already been manually annotated, these identified terms can serve as a supplement to the pre-annotated terms.
- **Special patterns observed in sample text:** Some specific patterns, like medication dosages (e.g., enoxaparin 40 mg/0.4 ml) or special notations (e.g., "b.i.d."), may carry significant meaning. We retained these terms unless they were identified as private information to preserve the clinical background of the raw letters.
- **Part of speech (POS) tagging:** Different parts of speech (POS) play distinct roles in interpreting clinical texts. We aimed to explore how these POS influence the model's understanding of clinical text. To achieve this, we used a toolkit (85) trained on the MIMIC-III (86) dataset for POS tagging. It performs better than SpaCy[5] and NLTK in handling clinical letters.

3.4 Clinical letter generation

We discuss the models and masking strategies that are used in generating synthetic clinical letters. It is important to clarify that our key objective is to generate letters that differ from the original ones, rather than being exact copies, as the same statement may indirectly reveal the patients' privacy. Although fine-tuning the model can always improve precision and enhance the model's semantic comprehension ability, it tends to produce letters that are too closely aligned with the originals. This also causes the fine-tuned model to rely too heavily on the original dataset, compromising its ability to generalize. Therefore, simply fine-tuning the model is not ideal if the PLMs can already generate the readable text. Instead, we should concentrate on how to *protect clinical terms and patient narratives as well as avoid privacy breaches.*

As discussed in Sections 2.3 and 2.4, decoder-only models struggle with processing long texts that require contextual understanding (75). Additionally, deploying them requires substantial computing resources and time. Therefore, we explored various PLMs, including both encoder-only and encoder–decoder models, in this paper. After evaluating their ability to generate synthetic letters from our dataset, we focused on Bio_ClinicalBERT, a well-performed model in our task, to experiment with different masking strategies. Additionally, from the discussion in Section 3.3, we need to split the text into various-length-chunks. So, the appropriate *length of these chunks* is also experimented with Bio_ClinicalBERT.

3.4.1 Encoder-only models with random masking

As mentioned earlier, the primary method for this paper involves masking and generation. We focused extensively on encoder-only models because of their advantage in bi-directional semantic comprehension. These encoder-only models, including BERT, RoBERTa, and Longformer (detailed in Section 2.3) were compared for their performance. Given the clinical focus of this task, we particularly explored model variants that were fine-tuned on clinical or biological datasets. However, as no clinically fine-tuned RoBERTa (62) variant was available, the RoBERTa-base was used for comparisons. Specifically, the encoder-only models we explored include Bio_ClinicalBERT (59), medicalai/ClinicalBERT (60), RoBERTa-base (62), and Clinical-Longformer (65).

We used the standard procedure for masked language modeling (MLM). First, the tokens that need to be masked were selected. They were then corrupted, resulting in masked text that includes both masked and unmasked tokens. Next, the model predicts the masked tokens and replaces them with the ones having the highest probabilities.

3.4.2 Encoder–decoder models with random masking

Although encoder–decoder models are not typically used for masked language modeling, they are well-suited for text generation. The architecture of T5, in particular, is designed to maintain the coherence of the text (70). Therefore, we included the T5 family models for comparisons.

The process of generating synthetic letters with encoder–decoder models is very similar to that with encoder-only models. The difference is that, unlike the BERT family, which automatically masks tokens and replaces them with "`<mask>`," the T5 family models *do not have any built-in masking function.* As a result, we identified the words that needed to be masked by index and removed them, which are represented as "extra_id_x" in the T5 family models. The text, with these words removed, was then used for generation, which we refer to as "text with blanks." To maintain consistency in the format, we later replaced "extra_id_x" with "`<mask>`" when displaying the masked text. Additionally, the T5 family models require a prompt as part of the input. For this task, the complete input was structured as "Fill in the blanks in the

[5]https://spacy.io/

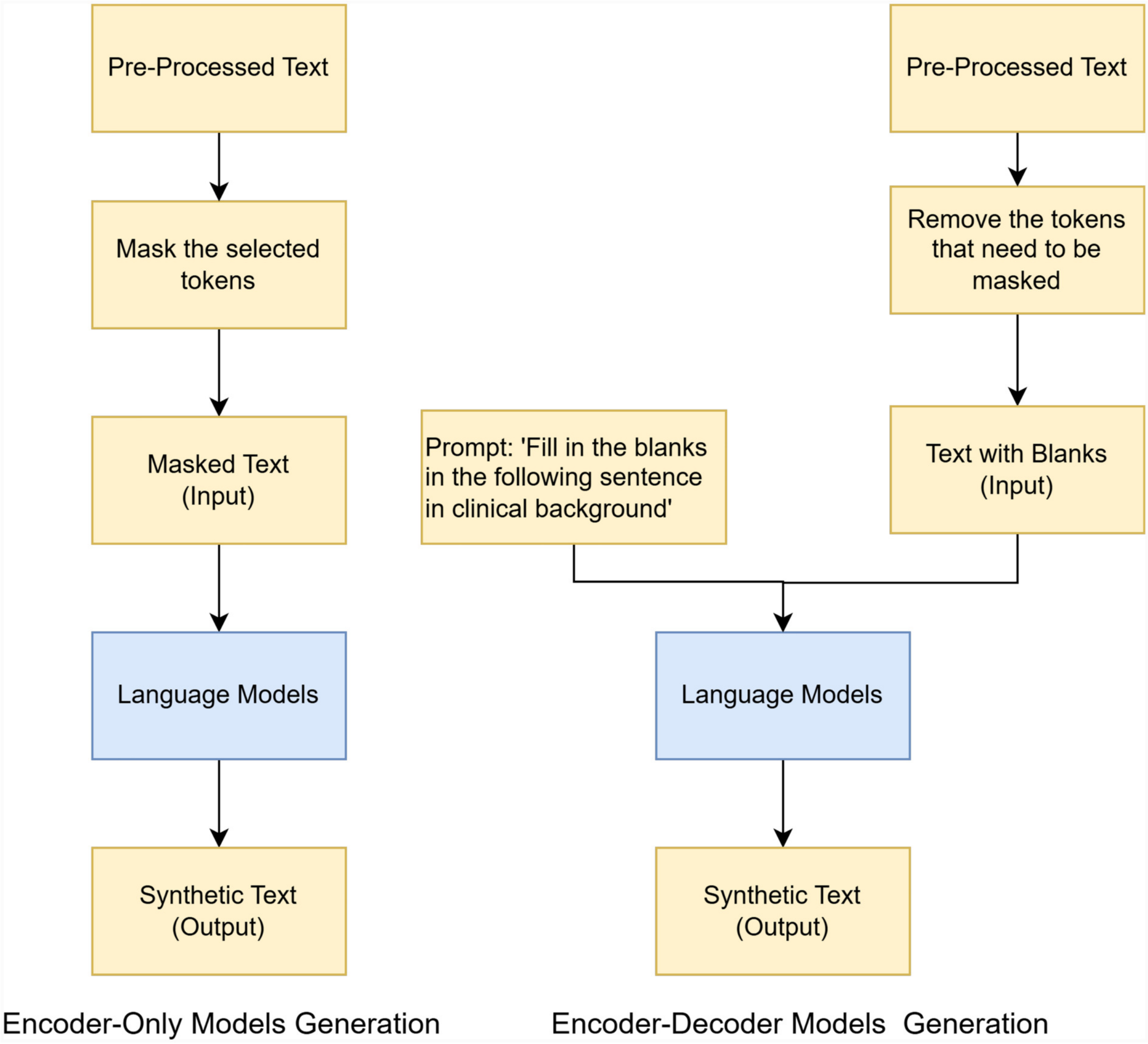


FIGURE 5
Comparison of encoder-only and encoder–decoder model architectures.

following sentence in the clinical background" + "text with blanks." In this paper, we used T5-base (70), Clinical-T5-Base (87, 88), Clinical-T5-Sci (87, 88), and Clinical-T5-Scratch (87, 88) for comparison. The comparison of encoder-only and encoder–decoder model architectures is shown in Figure 5.

3.4.3 Different masking strategies with Bio_ClinicalBERT

To make the synthetic letters more readable, clinically sound, and privacy-protective, different masking strategies were tested based on the following principles.

1. **Preserve annotated entities:** The manually annotated entities should not be masked to retain the clinical knowledge and context.
2. **Preserve extracted structures:** Tokens that are part of the document structure should be preserved as templates for clinical letters.
3. **Mask detected private information:** This is helpful in de-identification. Although the dataset we use is de-identified, this approach may be useful when this system is deployed with real-world data.
4. **Preserve medical terminology:** It still aims to retain clinical knowledge, as some diseases and treatments were not manually annotated.
5. **Preserve non-private numbers:** Certain numbers, such as drug dosage or heart rates, are indispensable for clinical diagnosis and treatment. However, only non-private numbers should be retained, while private information (such as phone numbers, ages, postal codes, dates, and email addresses) should be masked.
6. **Preserve punctuation:** Punctuation marks such as periods (".") and underscores ("___") should not be masked, as they clarify the sentence boundaries and make the synthetic letters more coherent (89).

7. **Retain special patterns in samples:** Tokens that match specific patterns (e.g., "Vitamin C ˆ1,000 mg," "Ibuprofen > 200 mg," etc.) should be retained, as they may contain important clinical details. These patterns are summarized by analyzing raw sample letters.

Based on the principles above, different masking strategies were experimented with:

1. **Mask randomly:** Tokens that can be masked are selected randomly from the text. We experimented with *masking ratios* ranging from 0% to 100% in 10% increments. This approach helps to understand how the number of masked tokens influences the quality of synthetic letters and provides a baseline for other masking strategies.
2. **Mask based on POS tagging:** We experimented with different configurations in this section, such as masking only nouns, only verbs, etc. It is helpful to understand how POS influences the models' context understanding. Similar to the random masking approach, we selected the tokens based on their POS configuration and masked them in 10% increments from 0% to 100%.
3. **Mask stopwords:** Stopwords generally contribute little to the text's main idea. Masking stopwords serves two purposes: reducing the *noise* for model understanding and increasing the *variety* of synthetic text by predicting these words. Moreover, they do not influence crucial clinical information. This approach is highly similar to the one used in "Mask based on POS tagging." The only difference is the criteria for selecting tokens. Specifically, tokens are selected based on whether they are stopwords rather than on their POS. The "NLTK" library was used for detecting stopwords in the text.
4. **Hybrid masking using different ratio settings:** After employing the aforementioned masking strategies, we observed the influence of these elements. Additionally, we experimented with their *combinations* at different masking ratios based on the outcomes, such as masking 50% nouns and 50% stopwords simultaneously.

3.4.4 Determining variable-length chunk size with Bio_ClinicalBERT

As mentioned in Section 3.3, we utilize two parameters in our chunk segment procedure: "max_lines" and "max_tokens." "max_lines" represents the desired length of each chunk, while "max_tokens" is related to the computing resources and model limitations. These two parameters determine the final length of each chunk together. Although most models we used have a limit of 512 tokens (except for the Longformer, which can process up to 4,096 tokens), we set 256 as the value for "max_tokens" due to computing resource constraints.

As for "max_lines," we experimented with values starting from 10 lines, increasing by 10 lines each time, and calculated the average tokens for each chunk. Once the token growth began to slow, we refined the search by using finer increments. Finally, we selected the number of lines at which the average tokens per chunk stopped growing. This is because more lines in each chunk provide more information for the model to predict masked tokens. However, if the chunk length reaches a critical threshold, it indicates that the primary limitation is "max_tokens" not "max_lines." Continuing to increase "max_lines" would lead to additional computational overhead, as the system would have to repeatedly check whether adding the next sentence meets the required line count.

3.5 Evaluation methods

Both quantitative and qualitative methods will be used to evaluate the performance. Additionally, a downstream task (NER) is employed to assess whether the synthetic clinical letters can replace the original raw data. The evaluation methods pipeline is illustrated in Figure 6.

3.5.1 Quantitative evaluation

To comprehensively evaluate the quality of the synthetic letters, we used quantitative evaluation from multiple dimensions, including the model's inference performance, the readability of the synthetic letters, and their similarity to the raw data. The specific metrics are listed in the following.

Standard NLG metrics: It covers standard NLG evaluation methods such as ROUGE, BERTScore, and METEOR. ROUGE measures literal similarity, the BERTScore evaluates semantic similarity, and METEOR builds on ROUGE by taking synonyms and word order into account. It provides a more comprehensive evaluation of the synthetic text (90).

These evaluations are performed by comparing the synthetic text with the original text. Moreover, a baseline is calculated by comparing the masked text to the original text. The evaluation score should exceed the baseline but remain below "1," ensuring that it does not exactly replicate the original text.

Readability metrics: To evaluate the readability, we calculated SMOG, Flesch Reading Ease, and Flesch–Kincaid Grade Level. Given our clinical focus, we prioritized SMOG as the primary readability metric, with Flesch Reading Ease and Flesch–Kincaid Grade Level as reference standards. In this analysis, we compared the readability metrics of the synthetic text with those of the original and masked texts. The evaluation results should closely approximate the original text's metrics. Significant differences (91)[6] may suggest that the model cannot preserve semantic coherence and readability adequately.

Advanced text quality metrics: In this part, we calculated the perplexity, subjectivity, and information entropy. We want the synthetic letters to be useful in training clinical models. Therefore, perplexity should not be far away from the value of the original letters. As for subjectivity and information entropy, we expect the synthetic letters to be both subjective and informative.

[6]We define a significant difference as a change of 1 SMOG grade, 1 Flesch–Kincaid Grade Level, or 10 points in Flesch Reading Ease, as these thresholds approximately correspond to a shift of one grade level or readability tier.

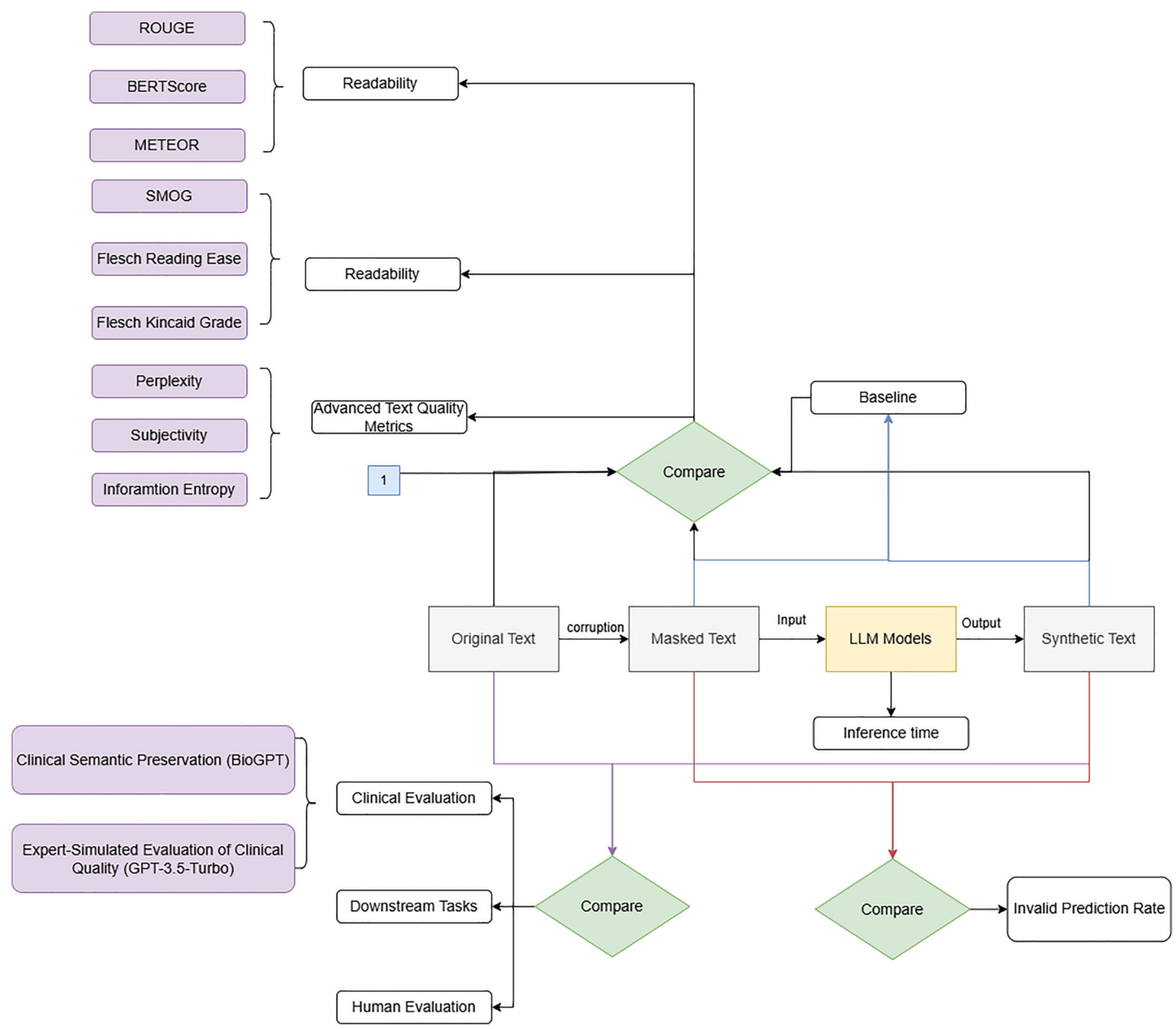


FIGURE 6
Evaluation pipeline.

Invalid prediction rate: We calculated the invalid prediction rate for each generation configuration. This ratio is determined by dividing the number of invalid predictions (such as punctuation marks or subwords) by the total number of masked words that need to be predicted. We expect the model to generate more meaningful words. Since punctuation marks are not masked, the model should avoid generating too many non-words. This metric can provide insights into the model's inference capability.

Inference time: The inference time for each generation configuration across the whole dataset (204 clinical letters) was recorded. Shorter inference times indicate lower computational resource consumption. When this system is deployed on large datasets, it is expected to save both time and computing resources.

3.5.2 Qualitative evaluation

In the quantitative evaluation, we not only calculated the evaluation metrics for the entire dataset but also recorded the results for each individual synthetic clinical letter. Interestingly, while some synthetic texts exhibited strong performance according to most metrics, they did not always appear satisfactory upon "visual" inspection. Conversely, some synthetic letters with average metrics may appear more visually appealing.

Although human evaluation is the most reliable approach for evaluating clinical letters, it is limited by availability and cost. Therefore, combining qualitative and quantitative evaluations helps in identifying suitable quantitative metrics for assessing the performance of our model. Once identified, one of these metrics can be used as the primary standard, while the others serve as supporting indicators. As a workaround, we selected a small sample of representative clinical letters based on the evaluation results. Subsequently, we reviewed the outcomes to better understand how different generation methods impacted these results, while also evaluating their correspondence with the quantitative metrics.

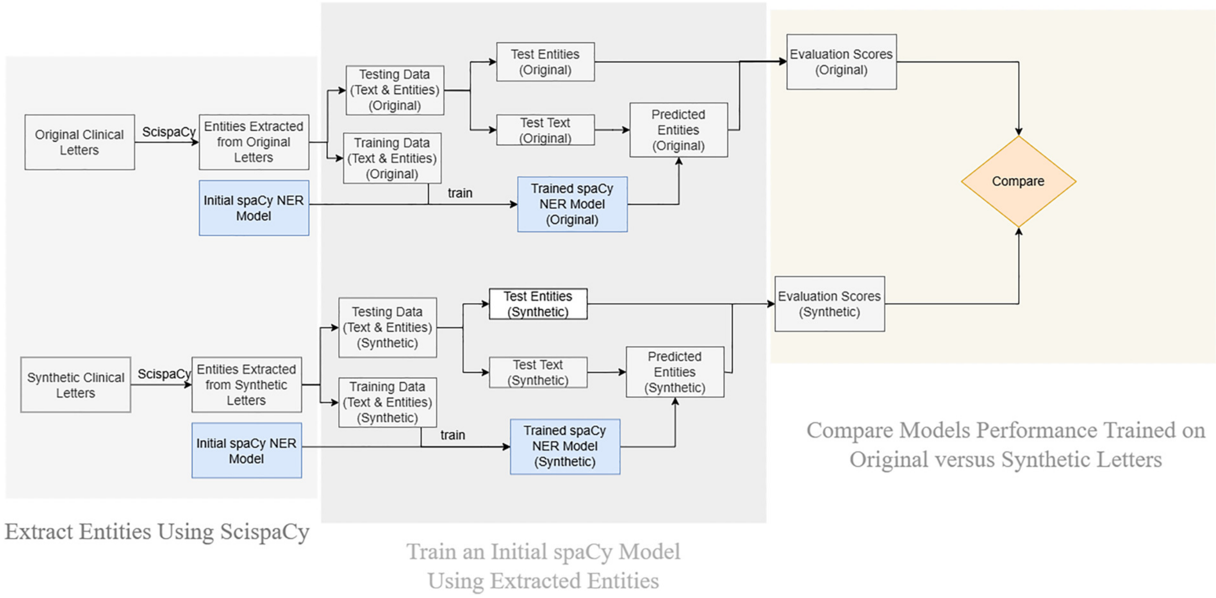


FIGURE 7
Workflow of the downstream NER task.

3.5.3 Downstream NER task

Beyond qualitative and quantitative evaluation, we can also apply synthetic clinical letters in a downstream NER task. This is helpful to further evaluate their quality and their potential to replace original ones in clinical research and model training.

ScispaCy[7] and spaCy[8] are used in this part. As shown in Supplementary Figure S9, they extract features from the text and learn the weights of each feature through neural networks. These weights are updated by comparing the loss between the predicted probabilities and actual labels. If a word does not belong to any label, it is classified as "O" (outside any entity). spaCy initializes these weights randomly. However, the version of ScispaCy we used, "en_ner_bc5cdr_md," is specifically fine-tuned on the BC5CDR corpus. It focuses more on "chemical" and "disease" entities while retaining the original general features.

In this downstream NER task, as shown in Figure 7, we initially extracted entities from letters using ScispaCy. Subsequently, these entities were used to train a base spaCy model. The trained model was then employed to extract entities from the testing set. Finally, we compared these newly extracted entities with those originally extracted by ScispaCy, and the evaluation scores were calculated. These steps were performed on both original clinical letters and synthetic letters, to assess whether the synthetic letters can potentially replace the original ones.

3.5.4 Clinical evaluation

Clinical semantic preservation: To evaluate how much clinical information is preserved, we used BioBERT (52) for a rough estimate. Specifically, we tokenized both the original and synthetic letters, obtained their embeddings using BioBERT, and computed the cosine similarity between them. Since BioBERT is trained on biomedical corpora, its embeddings are expected to capture clinical semantic features. A high similarity score indicates that clinical information is largely preserved. However, it is important to note that this method only evaluates the effectiveness of preserving clinical narratives at the semantic level and does not guarantee medical factuality.

Expert-simulated evaluation of clinical quality: To further evaluate the clinical usefulness of our synthetic letters, we employed GPT-3.5-Turbo (92) through prompt-based evaluation. Specifically, we evaluated the results from two perspectives: clinical soundness and narrative coherence. Clinical soundness measures whether the content aligns with medical factuality, while narrative coherence evaluates whether the letter is contextually consistent and resembles a real-world clinical letter. The prompt we used is shown in Figure 8.

3.6 Post-processing

3.6.1 Filling in the blanks

As described in Section 3, the dataset we used has been de-identified with all private information replaced by three

[7]https://allenai.github.io/scispacy/

[8]https://spacy.io/

```
Suppose you are a senior clinical doctor, you need to evaluate the quality of
clinical letters.

Please rate the following clinical letter from 0 to 1 on these two dimensions:

1. Clinical Soundness - Is the diagnosis and treatment medically reasonable?
2. Narrative Coherence - Does it flow logically like a real clinical note?

Use this format:
Clinical Soundness: <score> - <brief explanation>
Narrative Coherence: <score> - <brief explanation>

---

Letter:
\"\"\"{letter}\"\"\"
```

FIGURE 8
GPT-3.5-turbo prompt for clinical evaluation.

underscores "___." We hope that the synthetic clinical letters can maintain a certain degree of clinical integrity without disclosing any private patient information. To address this, a post-processing step was added to the synthetic results. This step involves masking the three underscores ("___") detected and using PLMs to predict the masked part again. For example, if the original text is "___ caught a cold," the post-processing result should ideally be "John caught a cold" or "patient caught a cold." Such synthetic clinical letters can better support clinical model training and teaching.

In this part, we used Bio_ClinicalBERT and BERT-base models. Although Bio_ClinicalBERT is better at clinical information understanding, this issue is not directly related to clinical practice, so we used BERT-base for comparison.

3.6.2 Spelling correction

Since our data come from real-world sources, it is inevitable that some words may be misspelled by doctors. These spelling errors can negatively impact the model's training process or hinder clinical practitioners' understanding of the synthetic clinical letters. Although some errors are masked and re-generated, our masking ratio is not always 100%, so some incorrect words may still exist. Toolkit "TextBlob" (93) was added to correct these errors. Specifically, it uses a rule-based approach that relies on a built-in vocabulary library to detect and correct misspellings.

3.7 Summary

In this section, we present the experimental design and subsequent implementation steps: these include defining project requirements, data collection and environmental setup, pre-processing, masking and generating the text, post-processing, the downstream NER task, clinical evaluation, and both qualitative and quantitative assessments. An example of the entire process flow is shown in Supplementary Figure S10.

4 Experimental results and analysis

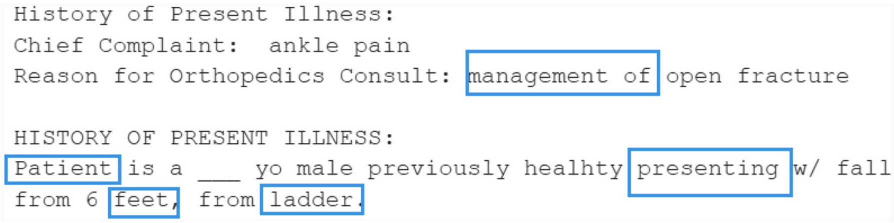
4.1 Chunk segmentation effects on inference time

As mentioned in Section 3.4.4, we set "max_lines" as a variable and "max_tokens" equal to 256. A series of increasing "max_lines" were tested until the average tokens per chunk peaked. We initially did this on a small sample (seven letters). The results are shown in Supplementary Table S6 for the Bio_ClinicalBERT model.

We can see that the average tokens per chunk reaches a peak as the "max_lines" parameter increases to 41. Similarly, inference time decreases as "max_lines" increases to 41, but it increases again once it exceeds this value. This experiment was also conducted on slightly larger samples of 10 and 30 letters. All of them showed the same trend. However, the inference time here may only reflect an overall trend, not exact results, as it is influenced by many factors, not only the chunk size but also the internet speed.

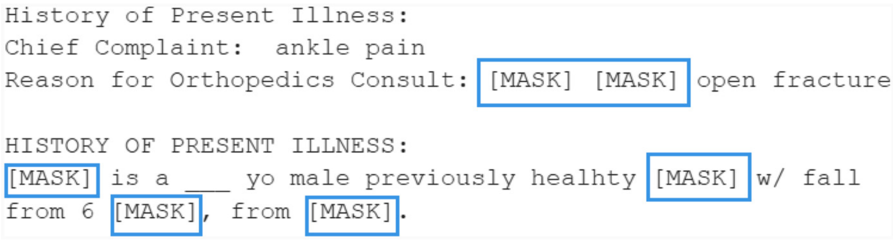
4.2 Random masking: qualitative results

We employed both encoder-only and encoder–decoder models to mask and generate the data, yielding numerous interesting results for human evaluation. Given space constraints, only a simple example is provided here. Following the masking principles in Section 3.4, the eligible tokens were randomly selected for masking. Although the initial intention was to mask 50% of tokens, the actual masking ratio was lower due to the requirement to preserve certain entities and structures.


History of Present Illness:
Chief Complaint: ankle pain
Reason for Orthopedics Consult: management of open fracture

HISTORY OF PRESENT ILLNESS:
Patient is a ___ yo male previously healhty presenting w/ fall
from 6 feet, from ladder.

FIGURE 9
Original unprocessed example sentence (6–8) ("note_id": "10807423-DS-19") (the circled tokens will be masked).


History of Present Illness:
Chief Complaint: ankle pain
Reason for Orthopedics Consult: [MASK] [MASK] open fracture

HISTORY OF PRESENT ILLNESS:
[MASK] is a ___ yo male previously healhty [MASK] w/ fall
from 6 [MASK], from [MASK].

FIGURE 10
An example of the masked sentence.

4.2.1 Encoder-only models

The original sentence is displayed in Figure 9. After feature extraction, the resulting structure is shown in Supplementary Figure S11. As detailed in Supplementary Table S7, certain manually annotated entities are excluded from masking. The output of this masking process is shown in Figure 10.

The generated text using Bio_ClinicalBERT is displayed in Figure 11. For "management of open fracture," the model produced "r," which is commonly used to denote "right" in clinical contexts, showing a relevant and logical prediction. Furthermore, the model's input "R ankle," despite not being in the figure due to space constraints, provided context for predicting "r" instead of "left." Interestingly, the term "admitted" was generated, even though it was not in the input, indicating the model's understanding of clinical context. Although the phrase "from 6 stairs, from home" is entirely different from the original ("from 6 feet, from ladder"), it remains contextually appropriate.

Overall, Bio_ClinicalBERT produced a clinically sound sentence, even though no tokens matched the original. In other examples, the predicted words may partially overlap with the original text. Nonetheless, this model effectively retains clinical information and introduces diversity without altering the text's meaning.

The results from medicalai/ClinicalBERT and Clinical-Longformer are shown in Supplementary Figures S12 and S13. All three clinical-related models correctly predicted "r" from the input context. The medicalai/ClinicalBERT model performs *comparably* to Bio_ClinicalBERT, despite adding an extra comma, which did not affect the text's clarity. However, Clinical-Longformer's predictions, while understandable, were *repetitive* and less satisfactory. Importantly, none of these three models altered the original meaning.

The result generated by RoBERTa-base is shown in Supplementary Figure S14. While the generated text initially seems reasonable, the predicted word "years" shifts the focus to a temporal context, which was not intended. This is likely because RoBERTa is pre-trained on a general corpus and lacks sufficient clinical knowledge for accurate text generation, or it could simply be a coincidence based on this specific sentence, where RoBERTa-base inferred "years" from its training data.

4.2.2 Decoder-only GPT-4o

Additionally, GPT-4o was used for comparison, with the prompt "Replace '<mask>' with words in the following sentence:." The results, shown in Supplementary Figure S15, are satisfactory. As discussed in Section 2.3, decoder-only models excel in few-shot learning (67), which is confirmed by this experiment. However, its performance may decline with long clinical letters (75).

4.2.3 Encoder–decoder models

To further evaluate different PLMs in generating synthetic letters, we tested the T5 family models. The generated results for

```
History of Present Illness:
Chief Complaint:  ankle pain
Reason for Orthopedics Consult: r for open fracture

HISTORY OF PRESENT ILLNESS:
This is a ___ yo male previously healhty admitted w/ fall
from 6 stairs, from home.
```

FIGURE 11
Example sentence generated by Bio_ClinicalBERT.

```
History of Present Illness:
Chief Complaint: ankle pain
Reason for Orthopedics Consult: fracture dislocation, open fracture

HISTORY OF PRESENT ILLNESS:
open is a ___ yo male previously healhty fracture w/ fall
from 6 Reason from for
```

FIGURE 12
Example sentence generated by T5-base.

the same sentence are shown in Figure 12 and Supplementary Figures S16–S18.

T5-base performs the best among these tested models. However, the results are still not fully rational, as it generated "open is a ___ yo male.'. The other three models tend to use de-identification (DEID) tags to replace the masked words, as these tags are part of their corpora. Furthermore, the T5 family models may predict multiple words for each token, aligning with findings in Section 2.3.

All these four T5 family models perform worse than the encoder-only models. This is consistent with the findings from Micheletti et al. (78) that MLM models outperform CLM models in medical datasets.

4.3 Random masking: quantitative results

4.3.1 Sentence-level quantitative results: encoder-only models

We first calculated representative quantitative metrics at the sentence level, matching the sample sentence used in Section 4.2. This approach allows for a better integration of quantitative and qualitative evaluations. Although SMOG is typically suited for medical datasets, it is less appropriate for sentence-level analysis, so the Flesch Reading Ease was used here. The results are presented in Table 1.

Our objective is to generate letters that differ from the original while maintaining clinical semantics and structure. Thus, high ROUGE scores are not desired, as they indicate substantial word/string overlap. The BERTScore is particularly useful for assessing semantic similarity, while METEOR offers a comprehensive evaluation considering word forms and synonyms theoretically. Flesch Reading Ease, on the other hand, provides a direct measure of textual readability.

We observed that clinical-related encoder-only models generally outperform RoBERTa-base in qualitative evaluation (see Section 4.2). However, from the quantitative perspective, RoBERTa-base shows mediocre performance across most metrics except for the BERTScore. In contrast, Bio_ClinicalBERT, despite no word overlap in this sample sentence, achieves a reasonable clinical context and the highest BERTScore among the models. Both medicalai/Clinical BERT and Bio_ClinicalBERT excel in Flesch Reading Ease, likely because they tend to predict tokens with fewer syllables that preserve the original meaning.

Surprisingly, while METEOR is designed to closely reflect human evaluation, the BERTScore appears to be more consistent with our evaluation criteria. This trend was observed in other sample texts as well. *Synthetic texts with higher BERTScore and lower ROUGE scores are more aligned with our objectives*. It is likely because the BERTScore is calculated using word embeddings, which can capture deep semantic similarity more effectively. All evaluation results *meet or exceed the baseline, affirming the effectiveness of these four encoder-only models* in generating clinical letters.

TABLE 1 Encoder-only models comparison at the sentence level (the "Baseline" without annotations was calculated by comparing the masked text to the original text).

Evaluation metric	Model evaluation			
	RoBERTa-base	medicalai/ClinicalBERT	Clinical-Longformer	Bio _ ClinicalBERT
ROUGE-1				
Generation performance	86.54	88.46	89.52	84.91
baseline	84.91	84.91	84.91	84.91
ROUGE-2				
Generation performance	74.51	78.43	79.61	73.08
baseline	73.08	73.08	73.08	73.08
ROUGE-L				
Generation performance	86.54	88.46	89.52	84.91
baseline	84.91	84.91	84.91	84.91
BERTScore F1				
Generation performance	0.81	0.83	0.84	0.85
baseline	0.79	0.65	0.79	0.65
METEOR				
Generation performance	0.87	0.88	0.90	0.86
baseline	0.85	0.85	0.85	0.85
Flesch Reading Ease				
Generation performance	10.24	18.70	9.22	16.67
baseline (original)	8.21	8.21	8.21	8.21
Baseline (mask)	16.67	16.67	16.67	16.67

4.3.2 Sentence-level quantitative results: encoder–decoder models

The evaluations for the encoder–decoder models, as presented in Table 2, generally underperform on most metrics compared to encoder-only models, except for METEOR. Interestingly, while the Flesch Reading Ease scores suggest a minimal impact on readability, the BERTScores are at least 0.05 lower than the baseline, indicating major deviations from the original meaning. This is consistent with our qualitative observations that the outputs from encoder–decoder models are largely unintelligible.

Collectively, the quantitative and qualitative results demonstrate that ***encoder–decoder models are not well-suited for generating clinical letters***, as they fail to preserve the original narratives. These results also support the validity of using BERTScore as the primary evaluation metric, with other metrics serving as supplementary references. We also tested this on the entire dataset, which produced *consistent* results.

4.3.3 Quantitative results on the full dataset: encoder-only models

Based on the findings above, we expect a higher BERTScore and a lower ROUGE score. We used the 0.4 masking ratio to illustrate the model comparison on the full dataset in Table 3. The other masking ratios show similar trends. Surprisingly, all encoder-only models this time showed comparable results, which contradicts our hypothesis that "Clinical-related" models would outperform base models. This suggests that *training on the clinical dataset has limited impact on the quality of synthetic letters*. This may be because most clinical-related tokens are preserved, with only the

TABLE 2 Encoder–decoder models comparison at the sentence level (the baseline without annotations was calculated by comparing the masked text to the original text).

Evaluation metric	Model evaluation			
	T5-base	Clinical-T5-base	Clinical-T5-scratch	Clinical-T5-Sci
ROUGE-1				
Generation performance	86.79	85.19	87.38	80.36
baseline	73.77	73.77	73.77	73.77
ROUGE-2				
Generation performance	75.00	71.70	75.25	69.09
baseline	63.33	63.33	63.33	63.33
ROUGE-L				
Generation performance	84.91	83.33	87.38	80.36
baseline	73.77	73.77	73.77	73.77
BERTScore F1				
Generation performance	0.44	0.40	0.45	0.40
baseline	0.50	0.50	0.50	0.50
METEOR				
Generation performance	0.85	0.83	0.83	0.82
baseline	0.85	0.85	0.85	0.85
Flesch Reading Ease				
Generation performance	8.21	8.21	19.71	8.21
baseline (original)	8.21	8.21	8.21	8.21
Baseline (mask)	8.21	8.21	8.21	8.21

TABLE 3 Encoder-only models comparison on the full dataset with Masking Ratio 0.4 (the baseline was calculated by comparing the masked text to the original text).

Evaluation metric	Model evaluation			
	RoBERTa-base	medicalai/ClinicalBERT	Clinical-Longformer	Bio_ ClinicalBERT
ROUGE-1				
Generation performance	92.98	93.63	94.66	93.18
baseline	85.64	85.44	85.64	85.61
ROUGE-2				
Generation performance	86.10	87.42	89.50	86.50
baseline	74.96	74.64	74.96	74.92
ROUGE-L				
Generation performance	92.54	93.22	94.38	92.71
baseline	85.64	85.44	85.64	85.61
BERTScore F1				
Generation performance	0.91	0.90	0.92	0.90
baseline	0.82	0.63	0.82	0.63

TABLE 4 Standard NLG metrics across different masking ratios using Bio_ClinicalBERT (the baseline was calculated by comparing the masked text to the original text).

Bio_ClinicalBERT	Masking ratio					
	1.0	0.8	0.6	0.4	0.2	0.0
ROUGE-1						
Generation performance	76.28	83.75	88.91	93.18	96.76	99.51
baseline	64.05	71.56	78.56	85.61	92.63	99.22
ROUGE-2						
Generation performance	62.60	70.77	78.81	86.50	93.42	99.02
baseline	51.72	57.88	65.38	74.92	86.27	98.61
ROUGE-L						
Generation performance	74.33	81.69	87.71	92.71	96.65	99.50
baseline	64.05	71.56	78.56	85.61	92.63	99.22
BERTScore						
Generation performance	0.63	0.75	0.83	0.90	0.95	0.99
baseline	0.29	0.39	0.50	0.63	0.79	0.98
METEOR						
Generation performance	0.70	0.80	0.87	0.93	0.97	1.00
baseline	0.66	0.72	0.78	0.85	0.92	0.99

remaining tokens being eligible for masking. Consequently, the normal encoder-only models can effectively understand the context and predict appropriate words while preserving clinical information. This differs slightly from the sentence-level comparisons, likely because the evaluation of a single sentence cannot fully represent the overall results. Despite this, the BERTScore as a primary evaluation metric remains useful, as the correspondence between qualitative and quantitative evaluation is consistent, whether at the sentence or dataset level.

We now explore how different *masking ratios* affect the quality of synthetic clinical letters. For each model, we generated data with masking ratios from 0.0 to 1.0, in increments of 0.1 (the masking ratios here refer only to the eligible tokens, as described in Section 3.4.3, and do not represent the actual overall masking ratio). Due to space limitations, we will present only the results for Bio_ClinicalBERT with a 0.2 increment here.

Table 4 presents that the higher masking ratio, the lower the similarity (metrics' scores). As we expected, all evaluation values are higher than the baseline, but still below "1." This means the model can understand the clinical context and generate understandable text. It is surprising that with a masking ratio of 1.0, the BERTScore increased from the baseline (0.29) to 0.63. Although this score is not very high, it still reflects that Bio_ClinicalBERT can generate clinical text effectively.

In Supplementary Table S8, we calculated three readability metrics, which are mentioned in Section 3.5. None of these metrics showed significant differences from the original ones. However, it is strange that the SMOG and Flesh–Kincaid Grade are not always between the original baseline and masking baseline. When the masking ratio is high, the evaluation values even fall below both the masking and the original baseline. This may be because a *higher masking ratio* *leads to a lower valid prediction rate*. If the predicted words include many spaces or punctuation marks, the readability will decrease obviously.

In Supplementary Table S9, considering the perplexity, the masking baseline is very high, while the values for synthetic letters are close to the original ones. This indicates that the synthetic letters are useful for training clinical models. For information entropy, regardless of the masking ratio, it can *effectively preserve the amount of information*. As for subjectivity, since all the values are similar, we do not need to worry that the synthetic letters will be biased.

As shown in Table 5, inference time for the entire dataset consistently ranges between 3 and 4 h. However, it decreases with either very high or very low masking ratios. A mid-range masking ratio of approximately 0.6 results in longer inference times, likely because lower ratios reduce the number of masked tokens to process, while higher ratios provide less context, reducing the computational load. This lack of effective context also increases the invalid prediction rate. Conversely, with a masking ratio of “0,” even a small number of prediction errors can substantially affect the overall accuracy, as only a few tokens are masked.

4.4 Other masking strategies using Bio_ClinicalBERT

There is a random selection when masking tokens at certain ratios. Masking different types of tokens will lead to different results, as shown in Figure 13 and Supplementary Figure S19. This variability is understandable since the encoder-only models use bidirectional attention, as mentioned in Section 2.3. *These models need to predict the masked tokens based on the context*. Therefore, it is necessary to experiment with different masking strategies based on the types of tokens. We used POS tagging and stopwords to observe how these strategies influence the quality of synthetic letters.

As discussed in Section 4.3, the BERTScore should be the primary evaluation metric for our objective. Additionally, the invalid prediction rate is useful for assessing the model’s ability to generate informative predictions, and ROUGE scores help evaluate literal diversity. Therefore, these quantitative metrics, calculated using different masking strategies, are shown in this section. Similar to Section 4.3, we experimented with different masking ratios calculated from the eligible tokens (masked tokens divided by eligible tokens). The ratios are increased in increments of 0.1, ranging from 0.0 to 1.0. Due to space constraints, only metrics with increments of 0.2 are shown here. A comparison with the same actual masking ratio (masked tokens divided by total tokens in the text) are also presented in this subsection.

TABLE 5 Inference time and invalid prediction rate across different masking ratios using Bio_ClinicalBERT.

Masking ratio	1.0	0.8	0.6	0.4	0.2	0.0
Inference time	3:12:05	3:28:56	3:33:26	3:25:16	3:13:26	3:01:11
Invalid prediction rate	0.72	0.47	0.34	0.28	0.25	0.37

4.4.1 Masking only nouns

Nouns often correspond to personally identifiable information (PII), so masking nouns can serve as a verification step for de-identification.

As shown in Supplementary Table S10, the fewer nouns we mask, the better all these metrics perform. This trend is consistent with random masking. When the noun masking ratio is 1.0, meaning that all nouns are masked, the BERTScore increases from a baseline of 0.70 to 0.89. This means that the *model predicted meaningful nouns*. A similar trend is observed for the ROUGE scores. All evaluations are higher than the baseline but lower than “1.” However, ROUGE scores show a

Masked Sentence
```
History of Present Illness:
Chief Complaint:  ankle pain
Reason for Orthopedics Consult: [MASK] of open fracture

HISTORY OF PRESENT ILLNESS:
[MASK] [MASK] [MASK] ___ yo male previously healhty presenting w/ fall
from 6 [MASK], [MASK] ladder.
```

Synthetic Sentence
```
History of Present Illness:
Chief Complaint:  ankle pain
Reason for Orthopedics Consult: repair of open fracture

HISTORY OF PRESENT ILLNESS:
Chief ##t _ ___ yo male previously healhty presenting w/ fall
from 6 stairs, hit ladder.
```

FIGURE 13
Example sentence 1 with different masked tokens.

smaller improvement than BERTScore. This may be because the model generates synonyms or paraphrases that retain the original meaning. As the noun masking ratio increases from 0.0 to 1.0, the BERTScore decrease from 0.99 to 0.89, indicating a significant decrease.

Therefore, to generate synthetic clinical letters that are distinguishable but still retain the original clinical information, we can only partially mask nouns (around 0.8 masking ratio). It helps maintain balanced evaluation scores. When all nouns are masked, the quality of synthetic letters deteriorates, with the BERTScore falling below 0.9 and the invalid prediction rate increasing to 0.37.

4.4.2 Masking only verbs

Masking only verbs also helps identify which token types are appropriate for masking to achieve our objective. While verbs are essential to describing clinical events, some can still be inferred from context. Therefore, *masking verbs may have a slight effect on the quality of synthetic clinical letters, but it can also introduce some variation.*

Supplementary Table S11 shows a similar trend for masking verbs as observed with other masking strategies in standard NLG metrics. However, it is surprising that as the masking ratio increases, both the invalid prediction rate and NLG metrics decrease. This phenomenon can be attributed to two main reasons. First, the model seems to prioritize predicting meaningful tokens (rather than punctuation, spaces, etc.) to generate coherent sentences. Contextual relevance is only considered after the sentence structure is complete. This may be due to the important role of verbs in sentences. Second, the original raw data may contain fewer verbs than nouns. Therefore, the number of actual masked tokens changes slightly when verbs are masked, making the model less sensitive to them. This is also reflected in BERTScore. If all verbs are masked, the BERTScore remains high at 0.95, whereas if all nouns are masked, the BERTScore drops to 0.89.

4.4.3 Masking only stopwords

As mentioned in Section 3.4.3, masking stopwords aims to reduce noise for model understanding while introducing variation in synthetic clinical letters. Supplementary Table S12 shows that *masking only stopwords follows a similar trend to random masking*, where a higher masking ratio leads to lower ROUGE Score and BERTScore. Additionally, the invalid prediction rate is at its lowest with a medium masking ratio. This is because higher masking ratios always result in more information loss. On the other hand, lower masking ratios lead to fewer tokens being masked, which makes small prediction errors more influential. The results show an overall low Invalid Prediction Rate and high BERTScore, indicating that *stopwords have only a limited influence on the model's understanding of context*. This is not because the original raw letters contain very few stopwords. In fact, there are even more stopwords than nouns and verbs, as seen in sample texts.

TABLE 6 Quantitative comparisons of 0.1 actual masking ratio (the baseline was calculated by comparing the masked text to the original text).

Bio_ClinicalBERT	Nouns masking (1.0)	Stopwords masking (0.6)	Random masking (0.3)
ROUGE-1			
Generation performance	93.29	96.56	95.10
baseline	88.13	89.04	89.16
ROUGE-2			
Generation performance	86.71	92.53	90.17
baseline	78.32	79.99	80.44
ROUGE-L			
Generation performance	93.00	96.23	94.86
baseline	88.13	89.04	89.16
BERTScore			
Generation performance	0.89	0.95	0.93
baseline	0.70	0.71	0.71
Invalid prediction rate			
Generation performance	0.37	0.20	0.26

4.4.4 Comparison of identical actual masking ratios

To further observe how different masking strategies influence the generation of clinical letters, we compared the results using the same actual masking ratios but with different strategies. In other words, the number of masked tokens is fixed, so the only variable is *the type of tokens being masked*. Supplementary Table S13 shows the results with a 0.04 actual masking ratio, and Table 6 shows the results with a 0.1 actual masking ratio.

As we can see, masking only stopwords achieved the highest BERTScore and lowest invalid prediction rate. Therefore, stopwords have little influence on the overall meaning of the text, which is consistent with our earlier findings. Additionally, masking nouns and verbs performed worse than random masking. Therefore, if we want to preserve the original meaning, we cannot mask too many nouns and verbs.

4.4.5 Hybrid masking

After comparing different strategies with the same actual masking ratio, we explored hybrid masking strategies and compared them with other strategies at the same actual ratio. The results are presented in Supplementary Table S14. The first three columns have the same actual masking ratio. Masking only stopwords achieved the strongest performance among these strategies. However, when nouns were also masked along with stopwords, the performance decreased, as masking nouns negatively affect the results. Despite this, it still performed better than random masking, indicating that stopwords have a greater influence than nouns. Next, we compared the last two columns. If 0.5 of nouns and 0.5 of stopwords were masked, adding an additional 0.5 of masked verbs led to worse performance, showing that verbs also negatively influence the model's performance.

TABLE 7 Comparison with and without entity preservation using Bio_ClinicalBERT.

Bio_ClinicalBERT	With entity preservation (0.4 nouns masking)	With entity preservation (0.3 random masking)	Without entity preservation (0.4 nouns masking)
ROUGE-1	97.62	95.10	97.31
ROUGE-2	95.12	90.17	94.46
ROUGE-L	97.56	94.86	93.71
BERTScore	0.96	0.93	0.91

4.4.6 Comparison with and without (w/o) entity preservation

To further explore whether keeping entities is useful for our task, we compared our results with a baseline that does not retain any entities. The baseline was trained with four epochs of fine-tuning on our dataset. Specifically, 0.4 of nouns from all tokens were randomly masked during baseline training. In contrast, in our experiments, only eligible tokens—excluding clinical information—were selected for masking. The comparisons are shown in Table 7.

As we can see, when 0.4 nouns were masked while preserving entities, the models performed much better than those without any entity preservation. Interestingly, when we randomly masked 0.3 while preserving entities, the model achieved lower ROUGE-1 and ROUGE-2 scores but higher ROUGE-L and BERTScores compared to models without entity preservation. This trend is consistent across different settings. This suggests that models preserving entities show less overlap with the original text, while they can retain the original narrative better. Additionally, the higher ROUGE-L score suggests that the step of preserving document structure is indeed effective.

These results also confirm our initial hypothesis that, for our objective—generating clinical letters that can keep the original meaning while adding some variety—retaining entities is much more effective than just fine-tuning the model. Moreover, this approach can effectively preserve useful information while avoiding overfitting.

4.5 Downstream NER task

To further evaluate whether synthetic letters have the potential to replace the original raw letters, particularly in the domains of clinical research and model training, a downstream NER task was implemented. Two spaCy NER models were trained separately on original raw letters and synthetic letters. Specifically, the synthetic letters were generated with 0.3 random masking while preserving entities.

As shown in Table 8, spaCy models trained on original and synthetic letters showed similar evaluation scores. They even achieved F1 scores comparable to ScispaCy's score of 0.843. Therefore, the unmasked context appears to have minimal influence on model understanding. Consequently, *our synthetic letters can be used in NER tasks to replace real-world clinical letters, thereby further protecting sensitive information.*

TABLE 8 Comparisons on downstream NER task.

Metric	spaCy trained on original letters	spaCy trained on synthetic letters	Performance Delta (Δ)
F1 Score	0.855	0.853	−0.002
Precision	0.865	0.863	−0.002
Recall	0.846	0.843	−0.003

4.6 Clinical evaluation

4.6.1 Clinical semantic preservation

As mentioned in Section 3.5.4, we used BioGPT with a random masking ratio of 0.3 to evaluate the integrity of clinical narrative preservation. As shown in Table 9, the mean similarity score reaches 0.98, which is slightly higher than the score obtained using the BERTScore metric. This may be because BioGPT evaluates semantic similarity from a clinical perspective. Additionally, such a high score suggests that the synthetic clinical letters can potentially serve as replacements for the original ones.

4.6.2 Expert-simulated evaluation of clinical quality

As mentioned in Section 3.5.4, we prompted GPT-3.5-Turbo to simulate a clinical expert and evaluate clinical soundness and narrative coherence. The masked letters (with text replaced by "<mask>") continued to serve as a baseline. The results are shown in Table 10

Clinical soundness: The average clinical soundness score of the generated letters (0.604) is slightly lower than that of the original letters (0.766). Surprisingly, it is even lower than the score of the masked letters (0.611). We further identified all cases where the generated letters scored lower than the masked ones in clinical soundness. These cases account for 14% (29 out of 204) of the processed letters. One possible explanation is that Bio_ClinicalBERT occasionally produces hallucinatory content, which may obscure or distort the original clinical semantics. However, in the majority of cases, the generated letters achieve clinical soundness scores comparable to the masked letters and close to the original ones—demonstrating the overall potential of our synthetic letters to replace real ones.

Narrative coherence: As expected, the narrative coherence score of the generated letters (0.460) is slightly lower than that of the original ones (0.664), but higher than that of the masked letters (0.418). These results further support the feasibility of using synthetic letters as substitutes for real clinical letters.

TABLE 9 Results of clinical semantic preservation evaluation using BioGPT.

Metric	Max score	Min score	Mean score	Std deviation	Evaluation set size
Value	0.9996	0.9037	0.9896	0.0147	204

TABLE 10 Expert-simulated evaluation results.

Metric	Avg.	Max.	Min.	Std.
Clinical soundness				
Baseline (original)	0.766	1.0	0.5	0.11
Baseline (masked)	0.611	1.0	0.5	0.147
Generation performance	0.604	1.0	0.5	0.145
Narrative coherence				
Baseline (original)	0.664	0.8	0.3	0.082
Baseline (masked)	0.418	0.7	0.2	0.168
Generation performance	0.460	0.7	0.2	0.177

4.7 Post-processing results

4.7.1 Filling in the blanks

One example text without post-processing is shown in Supplementary Figure S20. After filling in the blanks, the results with BERT-base and Bio_ClinicalBERT are shown in Figure 14 and Supplementary Figure S21, respectively. We can see that both models can partially achieve the goal of making the text more complete. However, neither of them created a coherent story to fill in these blanks. They just used general terms like "hospital" and "clinic." Perhaps other decoder-only models, more suitable for generating stories like GPT, could perform better and should be explored in the future.

4.7.2 Spelling correction

Supplementary Figure S22 shows that if the incorrect words are masked, the models may be able to correct the misspelled tokens by predicting them. However, the masking process is random. Additionally, sometimes the predicted words will be incorrect because some models tokenise the sentence into word-pieces. Therefore, a post-processing step is necessary for correcting spelling.

As shown in Supplementary Figure S23, tooltik "TextBlob" (93) can successfully correct misspelled words ("healhty") in our sample text. However, if clinical entities are not preserved during the pre-processing step, "TextBlob" (93) may misidentify some clinical terms as spelling errors. This may be because "TextBlob" (93) was developed on the general corpus rather than a clinical one. Additionally, its corrections are limited to the word level and do not consider any context. Therefore, if words are misspelled deliberately, they could be processed incorrectly. Thus, *developing a clinical misspelling correction toolkit is a promising* research direction in the future.

4.8 Discussion

We found that different masking strategies result in notable differences in model performance. To enhance the practical applicability of our research, we provide a guideline for selecting appropriate masking strategies for different scenarios, as presented in Table 11.

As mentioned earlier, we observe that when most clinical terms are preserved, fine-tuning the model may not be necessary. In terms of clinical evaluation, hallucinated content was found to negatively affect clinical soundness, suggesting that retrieval-augmented generation (RAG) or integration with a clinical knowledge graph may be beneficial for future improvements. Further exploration is also needed—such as dynamic vocabulary construction—to better handle clinical abbreviations and novel terms. Our synthetic framework for clinical letters did not show any notable negative effects on narrative coherence or semantic preservation, and the high performance in downstream NER tasks further supports the feasibility of using synthetic letters as substitutes for original ones. Although filling in blanks and correcting spelling errors are essential for improving text quality, mitigating errors in processing rare clinical terms remains a major challenge, as previously discussed.

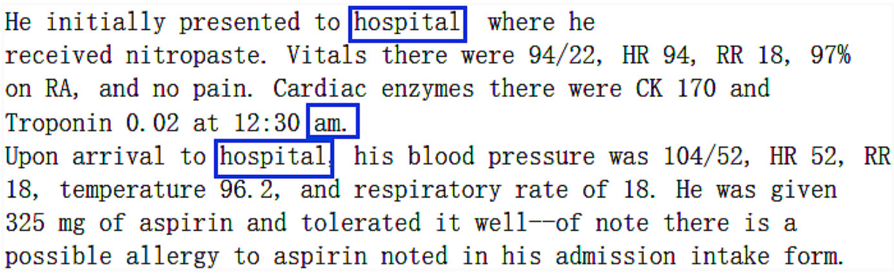
5 Conclusions and future work

5.1 Key findings

These results provided some useful findings in generating clinical letters, including

- **Encoder-only models generally perform much better** in clinical-letter masking and generation tasks, which is consistent with a very recent study by Micheletti et al. (78). When clinical information is preserved, **base encoder-only models perform comparably to clinical-related models**.
- To generate clinical letters that preserve clinical narrative while adding variety, **BERTScore should be the primary evaluation metric**, with other metrics serving as supporting references. This is because BERTScore focuses more on semantic rather than literal similarity, and it is consistent with qualitative assessment results.
- **Different types of masked tokens influence** the quality of synthetic clinical letters. Stopwords exert a positive impact, while nouns and verbs exert negative impacts.
- For our objective, preserving useful tokens is more effective than just fine-tuning the model without preserving any entities.
- The unmasked context has minimal influence on the models' understanding. As a result, the synthetic letters can be effectively used in the downstream NER task to replace original real-world letters.
- The synthetic letters largely preserve the consistency and coherence of clinical narratives from the original letters. However, Bio_ClinicalBERT occasionally generates hallucinated content, which may negatively impact clinical soundness and factuality.

10.3389/fdgth.2025.1497130



He initially presented to hospital where he received nitropaste. Vitals there were 94/22, HR 94, RR 18, 97% on RA, and no pain. Cardiac enzymes there were CK 170 and Troponin 0.02 at 12:30 am.
Upon arrival to hospital, his blood pressure was 104/52, HR 52, RR 18, temperature 96.2, and respiratory rate of 18. He was given 325 mg of aspirin and tolerated it well--of note there is a possible allergy to aspirin noted in his admission intake form.

FIGURE 14
Post-processing results with BERT-Base.

TABLE 11 Priority-based masking guidelines.

Priority	Note	Suggested masking strategy	Application scenarios
Diversity first	To improve the model's generalisation	Random masking (primarily), clinical terms masking (limited)	Basic clinical model pre-training; data augmentation
Clinical soundness first	The synthetic letters should satisfy clinical factuality	Keep clinical terms (complete); stopwords masking (extensive); verbs/ nouns masking	Clinical education; clinical QA model training; clinical model fine-tuning
Privacy first	To prevent PHI disclosure and mitigate privacy reconstruction through adversarial attacks	Private tokens masking (complete); nouns masking (extensive); verbs/ stopwords masking (medium)	Building open-source datasets; commercial deployment

5.2 Limitations

Although the strategies mentioned above help generate diverse, de-identified synthetic clinical letters, there are still some limitations in applying this method generally.

- **Challenges in the dataset:** Since these clinical letters are derived from the real world, certain issues are inevitable. For example, there may be spelling errors in the dataset. In note_id "10807423-DS-19," the word "healthy" is misspelled as "healhty." Such errors can negatively impact the usability of the synthetic text. Additionally, some polysemous words may cause contextual ambiguity. For instance, the word "back" can refer to an anatomical entity (e.g., the back of the body), or be used as an adverb.
- **Data volume:** Due to the difficulty in collecting annotated data, only 204 clinical letters were included in our research. This limited sample size may not be sufficiently representative, which could restrict the generalizability of our findings to a broader scenario. Moreover, the data we used were already de-identified. Although we considered de-identification and took steps to mask all private information, the effectiveness of these approaches cannot be thoroughly evaluated, as we do not have access to sensitive datasets.
- **Evaluation metrics:** In this paper, we primarily used BERTScore as our main evaluation metric, while also incorporating other metrics such as ROUGE and readability metrics. However, there is currently no comprehensive evaluation framework that can assess all aspects simultaneously, including maintaining the original meaning, diversity, readability, clinical soundness, and even privacy protection effectiveness.
- **Clinical knowledge understanding:** While the model can often preserve clinical entities and generate contextually reasonable tokens, it sometimes makes comprehension errors. For example, in a context where "LLE" ("left lower extremity") is used, Bio_ClinicalBERT incorrectly predicts the nearby masked token as "R ankle" ("right ankle"). In this case, the model fails to accurately capture the side clinical knowledge. Other challenges lie in handling long-tail phenomena and understanding abbreviated expressions, which are common in clinical language. Although spell correction techniques are explored in our project, distinguishing between a genuinely novel term and a simple misspelling remains difficult.
- **Computing resources:** Due to resource limitations, we explored a limited range of language generation models. Alternative architectures—such as enhanced decoder-only models—may be more suitable for our task.

5.3 Future work

Based on the limitations mentioned above, we outlined some potential directions to further explore:

- **Test on more clinical datasets:** To further evaluate the effectiveness of these masking strategies, more annotated clinical letters should be tested to assess system generalization.
- **Assess de-identification performance:** A quantitative metric for de-identification evaluation should be included in the future. Non-anonymous synthetic datasets can be used to evaluate the de-identification process, so that this system can be applied directly to real-world clinical letters in the future.
- **GRPO-based reinforcement learning:** The group relative policy optimization (GRPO) algorithm, as proposed in DeepSeek (94), has the potential to effectively balance multiple objectives, including clinical soundness, semantic integrity, textual diversity, and de-identification quality.

- **Evaluation benchmark:** A new metric suitable for our task should be developed. Specifically, this metric should consider both similarity and diversity. Weighting parameters for each dimension could be useful and can be obtained through neural networks. For evaluating clinical soundness, it is necessary to invite more clinicians to assess the synthetic letters based on multiple dimensions (77). Furthermore, mapping from clinical letters to their quality scores can be learned using deep learning.
- **Balancing knowledge from both clinical and general domains:** Although there are numerous clinical-related encoder-only models, only a few can effectively integrate clinical and general knowledge. Xie et al. (95) demonstrated that mixing the clinical dataset with the general dataset in a certain proportion can help the model better understand clinical knowledge. Therefore, a new BERT-based model should be trained from scratch using both clinical and general domain datasets.
- **Synonymous substitution:** We focused on exploring the range of eligible tokens for masking. Additionally, a masking strategy similar to BERT's can be integrated with our results (58). Specifically, we can select certain tokens to mask, some to retain, and replace others with synonyms. This approach can further enhance the variety of synthetic clinical letters. Moreover, the retained clinical entities can also be substituted using entity linking to SNOMED CT.
- **Spelling correction:** As mentioned in Section 4.7, very few toolkits are available for spelling correction in the clinical domain. Standard spelling correction tools may misidentify clinical terms as misspelled words. Therefore, it is necessary to develop a specialized spell-checking tool adapted to the clinical domain.

Data availability statement

The data used in this study are from the publicly available MIMIC database (Medical Information Mart for Intensive Care), accessible to qualified researchers who complete the PhysioNet credentialing process and agree to the Data Use Agreement (https://physionet.org/content/mimiciv/2.2/).

Author contributions

LR: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft; SB: Data curation, Formal analysis, Software, Validation, Writing – review & editing; LH: Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing; WD-P: Methodology, Supervision, Writing – review & editing; GN: Conceptualization, Funding acquisition, Investigation, Methodology, Resources, Supervision, Writing – review & editing.

Funding

The authors declare that financial support was received for the research and/or publication of this article. LH is grateful to the EU project 4D Picture funded by the European Union under Horizon Europe Work Programme 101057332. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The UK team are funded under the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number: 10041120.

Acknowledgments

We thank the insightful feedback from Nicolo Micheletti, which is very valuable to the start of this work. Generative AI technology (GPT) was used for model result comparison in the Results section, and the model name and version have been clearly indicated as required.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2025.1497130/full#supplementary-material

Supplementary Figure S1
An example of LT3 (76).

Supplementary Figure S2
An input example of conditional text generation (75).

Supplementary Figure S3
Workflow of discharge summary generation using clinical guidelines.

Supplementary Figure S4
Workflow of MLM and CLM comparison in text generation.

Supplementary Figure S5
Text excerpt from the original letter (6–8) ("note_id": "17656866-DS-6").

Supplementary Figure S6
Sample text from original letters (6–8) ("note_id": "10807423-DS-19").

Supplementary Figure S7
Sentence fragment exceeding token limit (6–8) ("note_id": "10807423-DS-19").

Supplementary Figure S8
Example sentence for feature extraction (see Supplementary Table S5 for Extracted Features).

Supplementary Figure S9
Training process for spaCy NER model.[9]

Supplementary Figure S10
An example of masking and generating.

Supplementary Figure S11
Extracted structure from the example sentence.

Supplementary Figure S12
Example sentence generated by medicalai/ClinicalBERT.

Supplementary Figure S13
Example sentence generated by Clinical-Longformer.

Supplementary Figure S14
Example sentence generated by RoBERTa-base.

Supplementary Figure S15
Example sentence generated by GPT-4o.

Supplementary Figure S16
Example sentence generated by Clinical-T5-Base.

Supplementary Figure S17
Example sentence generated by Clinical-T5-Scratch.

Supplementary Figure S18
Example sentence generated by Clinical-T5-Sci.

Supplementary Figure S19
Example sentence 2 with different masked tokens.

Supplementary Figure S20
Example sentence before post-processing (6–8) ("note_id": "16441224-DS-19").

Supplementary Figure S21
Post-processing results with Bio_ClinicalBERT.

Supplementary Figure S22
Spelling correction by masking and generating (6–8) ("note_id": "10807423-DS-19").

Supplementary Figure S23
Spelling correction (6–8) ("note_id": "10807423-DS-19").

Supplementary Table S1
Encoder-only models and their fine-tuned datasets.

Supplementary Table S2
The T5 family models used in our work.

Supplementary Table S3
Extracted entities and their details.

Supplementary Table S4
Comparison of tokenization methods for different LMs on sentence "Patient is a ___ yo male previously healthy presenting w/fall from 6 feet, from ladder."

Supplementary Table S5
Example: summary of feature extraction operations and extracted features.

Supplementary Table S6
Relation between chunk sizes and model inference time, average token number.

Supplementary Table S7
Annotated entities extracted from the example sentence (they should be preserved from masking).

Supplementary Table S8
Readability metrics across different masking ratios using Bio_ClinicalBERT (the Baseline without annotations was calculated by comparing masked text to the original text).

Supplementary Table S9
Advanced text quality metrics across different masking ratios using Bio_ClinicalBERT (the baseline without annotations was calculated by comparing masked text to the original text).

Supplementary Table S10
Quantitative comparisons of noun masking ratio (the "Baseline" was calculated by comparing masked text to the original text).

Supplementary Table S11
Quantitative comparisons of verb masking ratios (the "Baseline" was calculated by comparing masked text to the original text).

Supplementary Table S12
Quantitative comparisons of stopword masking ratios (the "Baseline" was calculated by comparing masked text to the original text).

Supplementary Table S13
Quantitative comparison of different masking strategies at a 0.04 actual masking ratio (the "Baseline" was calculated by comparing masked text to the original text).

Supplementary Table S14
Quantitative comparisons for hybrid masking (the baseline was calculated by comparing masked text to the original text).

References

1. Rayner H, Hickey M, Logan I, Mathers N, Rees P, Shah R. Writing outpatient letters to patients. *BMJ*. (2020) 368:1–4. doi: 10.1136/bmj.m24

2. Tarur SU, Prasanna S. Clinical case letter. *Indian Pediatr*. (2021) 58:188–9. doi: 10.1007/s13312-021-2144-3

3. Tucker K, Branson J, Dilleen M, Hollis S, Loughlin P, Nixon MJ, et al. Protecting patient privacy when sharing patient-level data from clinical trials. *BMC Med Res Methodol*. (2016) 16:5–14. doi: 10.1186/s12874-016-0169-4

4. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. *J Big Data*. (2018) 5:1–18. doi: 10.1186/s40537-017-0110-7

5. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. *JMIR Med Inform*. (2020) 8:e17984. doi: 10.2196/17984

6. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*. (2000) 101:e215–20. doi: 10.1161/01.cir.101.23.e215

7. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, et al. Mimic-iv (2024). doi: 10.13026/hxp0-hg59.

8. Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. (2023) 10:1. doi: 10.1038/s41597-022-01899-x

9. the President and of Harvard College, F. DBMI data portal (2023). Available at: https://portal.dbmi.hms.harvard.edu/ (Accessed September 4, 2024).

10. Humbert-Droz M, Mukherjee P, Gevaert O. Strategies to address the lack of labeled data for supervised machine learning training with electronic health records: case study for the extraction of symptoms from clinical notes. *JMIR Med Inform*. (2022) 10:e32903. doi: 10.2196/32903

11. Amin-Nejad A, Ive J, Velupillai S. Exploring transformer text generation for medical dataset augmentation. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference* (2020). p. 4699–708.

12. Hüske-Kraus D. Text generation in clinical medicine: a review. *Methods Inf Med*. (2003) 42:51–60. doi: 10.1055/s-0038-1634209

[9]Image Credit https://spacy.io/usage/training

13. Tang R, Han X, Jiang X, Hu X. Does synthetic data generation of LLMs help clinical text mining? *ArXiv abs/2303.04360* (2023).

14. Boonstra MJ, Weissenbacher D, Moore JH, Gonzalez-Hernandez G, Asselbergs FW. Artificial intelligence: revolutionizing cardiology with large language models. *Eur Heart J*. (2024) 45:332–45. doi: 10.1093/eurheartj/ehad838

15. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. (2011) 18:544–51. doi: 10.1136/amiajnl-2011-000464

16. Satapathy R, Cambria E, Hussain A. *Sentiment analysis in the bio-medical domain*. Cham: Springer (2017).

17. van der Lee C, Krahmer E, Wubben S. Automated learning of templates for data-to-text generation: comparing rule-based, statistical and neural methods. In: *Proceedings of the 11th International Conference on Natural Language Generation* (2018). p. 35–45.

18. Brown PF, Cocke J, Della Pietra SA, Della Pietra VJ, Jelinek F, Lafferty J, et al. A statistical approach to machine translation. *Comput Linguist*. (1990) 16:79–85. doi: 10.5555/92858.92860

19. Sharma S, Diwakar M, Singh P, Singh V, Kadry S, Kim J. Machine translation systems based on classical-statistical-deep-learning approaches. *Electronics*. (2023) 12:1716. doi: 10.3390/electronics12071716

20. Eddy SR. Hidden Markov models. *Curr Opin Struct Biol*. (1996) 6:361–5. doi: 10.1016/S0959-440X(96)80056-X

21. Masuko T, Kobayashi T, Tamura M, Masubuchi J, Tokuda K. Text-to-visual speech synthesis based on parameter generation from HMM. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*. Vol. 6. IEEE (1998). p. 3745–8.

22. Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel H-P. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinf*. (2008) 9:1–14. doi: 10.1186/1471-2105-9-207

23. Sutton C, McCallum A. An introduction to conditional random fields. *Found Trends® Mach Learn*. (2012) 4:267–373. doi: 10.1561/2200000013

24. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. (1997) 9:1735–80. doi: 10.1162/neco.1997.9.8.1735

25. Graves A, Graves A. Long short-term memory. *Superv Seq Label Recurr Neural Netw*. (2012) 385:37–45. doi: 10.1007/978-3-642-24797-2_4

26. Santhanam S. Context based text-generation using LSTM networks. *arXiv* [e-prints]. *arXiv–2005* (2020).

27. Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model. *Adv Neural Inf Process Syst*. (2000) 3:13. doi: 10.5555/944919.944966

28. Johnson SJ, Murty MR, Navakanth I. A detailed review on word embedding techniques with emphasis on word2vec. *Multimed Tools Appl*. (2024) 83:37979–8007. doi: 10.1007/s11042-023-17007-z

29. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *arXiv* [Preprint]. *arXiv:1301.3781* (2013).

30. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Process Syst*. (2013) 26:3111–9. doi: 10.5555/2999792.2999959

31. Ma L, Zhang Y. Using Word2Vec to process big text data. In: *2015 IEEE International Conference on Big Data (Big Data)*. IEEE (2015). p. 2895–7.

32. Camacho-Collados J, Pilehvar MT. From word to sense embeddings: a survey on vector representations of meaning. *J Artif Intell Res*. (2018) 63:743–88. doi: 10.1613/jair.1.11259

33. Zhang Z, Wu Y, Zhao H, Li Z, Zhang S, Zhou X, et al. Semantics-aware bert for language understanding. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34 (2020). p. 9628–35.

34. Li J, Tang T, Zhao WX, Nie J-Y, Wen J-R. Pre-trained language models for text generation: a survey. *ACM Comput Surv*. (2024) 56:1–39. doi: 10.1145/3649449

35. Liu J, Shen D, Zhang Y, Dolan WB, Carin L, Chen W. What makes good in-context examples for GPT-3? In: *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures* (2022). p. 100–14.

36. Grishman R, Sundheim BM. Message understanding conference-6: a brief history. In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics* (1996).

37. Bose P, Srinivasan S, Sleeman IV WC, Palta J, Kapoor R, Ghosh P. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Appl Sci*. (2021) 11:8319. doi: 10.3390/app11188319

38. Kundeti SR, Vijayananda J, Mujjiga S, Kalyan M. Clinical named entity recognition: challenges and opportunities. In: *2016 IEEE International Conference on Big Data (Big Data)*. IEEE (2016). p. 1937–45.

39. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. *J Mach Learn Res*. (2011) 12:2493–537. doi: 10.5555/1953048.2078186

40. Berg H, Henriksson A, Dalianis H. The impact of de-identification on downstream named entity recognition in clinical text. In: *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis* (2020). p. 1–11.

41. Berg H, Henriksson A, Fors U, Dalianis H. De-identification of clinical text for secondary use: research issues. In: *14th International Joint Conference on Biomedical Engineering Systems and Technologies-BIOSTEC 2021, 11-13 Februar, 2021, Vienna, Austria*. SciTePress (2021). p. 592–9.

42. Meystre SM. De-identification of unstructured clinical data for patient privacy protection. In: Gkoulalas-Divanis A, Loukides G, editors. *Medical Data Privacy Handbook*. Springer (2015). p. 697–716.

43. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. *J Am Med Inform Assoc*. (2017) 24:596–606. doi: 10.1093/jamia/ocw156

44. Kovačević A, Bašaragin B, Milošević N, Nenadić G. De-identification of clinical free text using natural language processing: a systematic review of current approaches. *Artif Intell Med*. (2024) 151:102845. doi: 10.1016/j.artmed.2024.102845

45. Norgeot B, Muenzen K, Peterson TA, Fan X, Glicksberg BS, Schenk G, et al. Protected health information filter (Philter): accurately and securely de-identifying free-text clinical notes. *NPJ Digit Med*. (2020) 3:57. doi: 10.1038/s41746-020-0258-y

46. Meystre SM, Ferrández O, Friedlin FJ, South BR, Shen S, Samore MH. Text de-identification for privacy protection: a study of its impact on clinical text information content. *J Biomed Inform*. (2014) 50:142–50. doi: 10.1016/j.jbi.2014.01.011

47. Gatt A, Krahmer E. Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *J Artif Intell Res*. (2018) 61:65–170. doi: 10.1613/jair.5477

48. Cawsey AJ, Webber BL, Jones RB. Natural language generation in health care. *J Am Med Inform Assoc*. (1997) 4:473–82. doi: 10.1136/jamia.1997.0040473

49. Hirst G, DiMarco C, Hovy E, Parsons K. Authoring and generating health-education documents that are tailored to the needs of the individual patient. In: *User Modeling: Proceedings of the Sixth International Conference UM97 Chia Laguna, Sardinia, Italy June 2–5 1997*. Springer (1997). p. 107–18.

50. McKeown K, Jordan DA, Pan S, Shaw J, Allen BA. Language generation for multimedia healthcare briefings. In: *Fifth Conference on Applied Natural Language Processing* (1997). p. 277–82.

51. Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. Natural language processing in medicine: a review. *Trends Anaesth Crit Care*. (2021) 38:4–9. doi: 10.1016/j.tacc.2021.02.007

52. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinformatics*. (2022) 23:bbac409. doi: 10.1093/bib/bbac409

53. Kong M, Huang Z, Kuang K, Zhu Q, Wu F. Transq: transformer-based semantic query for medical report generation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer (2022). p. 610–20.

54. Lee SH. Natural language generation for electronic health records. *NPJ Digit Med*. (2018) 1:63. doi: 10.1038/s41746-018-0070-0

55. Gillioz A, Casas J, Mugellini E, Abou Khaled O. Overview of the transformer-based models for NLP tasks. In: *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE (2020). p. 179–83.

56. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. (2017). p. 6000–10. NIPS'17.

57. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics (2014). p. 1724.

58. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics (2019). p. 4171–86. doi: 10.18653/v1/N19-1423.

59. Alsentzer E, Murphy J, Boag W, Weng W-H, Jindi D, Naumann T, et al. Publicly available clinical bert embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (2019). p. 72–8.

60. Wang G, Liu X, Ying Z, Yang G, Chen Z, Liu Z, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nat Med*. (2023) 29:2633–42. doi: 10.1038/s41591-023-02552-9

61. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: generalized autoregressive pretraining for language understanding. *Adv Neural Inf Process Syst*. (2019) 32:5753–63. doi: 10.5555/3454287.3454804

62. Zhuang L, Wayne L, Ya S, Jun Z. A robustly optimized BERT pre-training approach with post-training. In: Li S, Sun M, Liu Y, Wu H, Liu K, Che W, et al., editors. *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Huhhot, China: Chinese Information Processing Society of China (2021). p. 1218–27.

63. Acheampong FA, Nunoo-Mensah H, Chen W. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artif Intell Rev*. (2021) 54:5789–829. doi: 10.1007/s10462-021-09958-2

64. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. *arXiv* [Preprint]. *arXiv:2004.05150* (2020).

65. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. A comparative study of pretrained language models for long clinical text. *J Am Med Inform Assoc*. (2023) 30:340–7. doi: 10.1093/jamia/ocac225

66. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. (2020). NIPS '20.

67. Wu Y. Large language model and text generation. In: Xu H, Demner D, editors. *Natural Language Processing in Biomedicine: A Practical Guide*. Springer (2024). p. 265–97.

68. AI, M. Llama3 model card (2023). Available at: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md (Accessed May 4, 2024).

69. Context.ai. Compare Llama 3 70b instruct to GPT-3.5 Turbo (2024) (Online; accessed August 14, 2024).

70. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res*. (2020) 21:1–67. https://dl.acm.org/doi/abs/10.5555/3455716.3455856

71. Cai P-X, Fan Y-C, Leu F-Y. Compare encoder–decoder, encoder-only, and decoder-only architectures for text generation on low-resource datasets. In: *Advances on Broad-Band Wireless Computing, Communication and Applications: Proceedings of the 16th International Conference on Broad-Band Wireless Computing, Communication and Applications (BWCCA-2021)*. Springer (2022). p. 216–25.

72. Tsirmpas D, Gkionis I, Papadopoulos GT, Mademlis I. Neural natural language processing for long texts: a survey on classification and summarization. *Eng Appl Artif Intell*. (2024) 133:108231. doi: 10.1016/j.engappai.2024.108231

73. Phan L, Anibal JT, Tran HT, Chanana S, Bahadroglu E, Peltekian A, et al. SciFive: a text-to-text transformer model for biomedical literature. *arXiv abs/2106.03598* (2021).

74. Lu Q, Dou D, Nguyen T. ClinicalT5: a generative language model for clinical text. In: Goldberg Y, Kozareva Z, Zhang Y, editors. *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics (2022). p. 5436–43. doi: 10.18653/v1/2022.findings-emnlp.398.

75. Amin-Nejad A, Ive J, Velupillai S. Exploring transformer text generation for medical dataset augmentation. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, et al., editors. *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association (2020). p. 4699–708.

76. Belkadi S, Micheletti N, Han L, Del-Pinto W, Nenadic G. Generating medical instructions with conditional transformer. In: *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI* (2023).

77. Ellershaw S, Tomlinson C, Burton OE, Frost T, Hanrahan JG, Khan DZ, et al. Automated generation of hospital discharge summaries using clinical guidelines and large language models. In: *AAAI 2024 Spring Symposium on Clinical Foundation Models* (2024).

78. Micheletti N, Belkadi S, Han L, Nenadic G. Exploration of masked and causal language modelling for text generation. *CoRR abs/2405.12630* (2024).

79. Johnson A, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV-Note: deidentified free-text clinical notes. *PhysioNet*. (2023). doi: 10.13026/1n74-ne17.

80. Sun S, Iyyer M. Revisiting simple neural probabilistic language models. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2021). p. 5181–8.

81. International, S. Snomed ct browser (2024). Available at: https://browser.ihtsdotools.org/? (Accessed September 3, 2024).

82. Zeng W, Jin S, Liu W, Qian C, Luo P, Ouyang W, et al. Not all tokens are equal: human-centric visual analysis via token clustering transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022). p. 11101–11.

83. Hou L, Pang RY, Zhou T, Wu Y, Song X, Song X, et al. Token dropping for efficient bert pretraining. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2022). p. 3774–84.

84. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: a python natural language processing toolkit for many human languages. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (2020). p. 101–8.

85. Zhang Y, Zhang Y, Qi P, Manning CD, Langlotz CP. Biomedical and clinical English model packages for the Stanza Python NLP library. *J Am Med Inform Assoc*. (2021) 28:1892–9. doi: 10.1093/jamia/ocab090

86. Johnson AE, Pollard TJ, Shen L, Lehman L-wH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. (2016) 3:1–9. doi: 10.1038/sdata.2016.35

87. Eric L, Johnson A. Clinical-T5: large language models built using MIMIC clinical text. *PhysioNet*. (2023). doi: 10.13026/rj8x-v335.

88. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. Physiobank, physiotoolkit, and physionet. *Circulation*. (2000) 101:e215–20. doi: 10.1161/01.CIR.101.23.e215.

89. Lamprou Z, Pollick F, Moshfeghi Y. Role of punctuation in semantic mapping between brain and transformer models. In: *International Conference on Machine Learning, Optimization, and Data Science*. Springer (2022). p. 458–72.

90. Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein J, Lavie A, Lin C-Y, Voss C, editors. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics (2005). p. 65–72.

91. DuBay WH. The principles of readability. *Online submission* (2004).

92. OpenAI. GPT-3.5-Turbo (2023). Available at: https://platform.openai.com/docs/models/gpt-3.5-turbo (Accessed March 26, 2025).

93. Loria S. Textblob: simplified text processing (2024). Available at: https://textblob.readthedocs.io/en/dev/ (Accessed September 3, 2024).

94. Shao Z, Wang P, Zhu Q, Xu R, Song J, Bi X, et al. Deepseekmath: pushing the limits of mathematical reasoning in open language models. *arXiv* [Preprint]. *arXiv:2402.03300* (2024).

95. Xie Q, Chen Q, Chen A, Peng C, Hu Y, Lin F, et al. Me-llama: foundation large language models for medical applications. (version 1.0.0). *PhysioNet*. (2024). doi: 10.13026/wwfd-2t39

96. Ren L, Belkadi S, Han L, Del-Pinto W, Nenadic G. Beyond reconstruction: generating privacy-preserving clinical letters. In: Habernal I, Ghanavati S, Jain V, Igamberdiev T, Wilson S, editors. *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*. Albuquerque: Association for Computational Linguistics (2025). p. 60–74. https://aclanthology.org/2025.privatenlp-main.6/

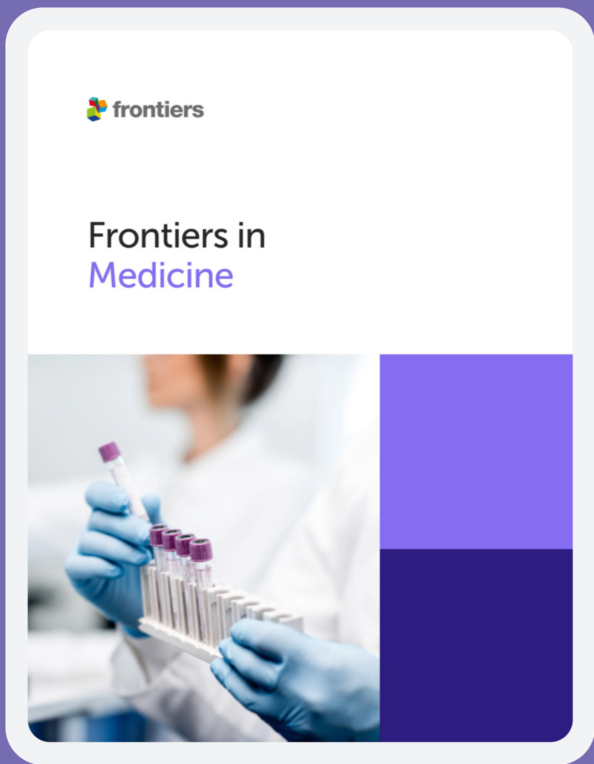
Frontiers in Medicine

Translating medical research and innovation into improved patient care

A multidisciplinary journal which advances our medical knowledge. It supports the translation of scientific advances into new therapies and diagnostic tools that will improve patient care.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact