frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 27 January 2025
DOI 10.3389/frai.2025.1504805

OPEN ACCESS

EDITED BY
Steffen Pauws,
Tilburg University, Netherlands

REVIEWED BY
Giacomo Rossettini,
University of Verona, Italy
Paolo Marcheschi,
Gabriele Monasterio Tuscany Foundation (CNR), Italy

*CORRESPONDENCE
Michel E. van Genderen
✉ m.vangenderen@erasmusmc.nl

RECEIVED 01 October 2024
ACCEPTED 06 January 2025
PUBLISHED 27 January 2025

CITATION
Workum JD, van de Sande D, Gommers D and van Genderen ME (2025) Bridging the gap: a practical step-by-step approach to warrant safe implementation of large language models in healthcare. *Front. Artif. Intell.* 8:1504805. doi: 10.3389/frai.2025.1504805

Bridging the gap: a practical step-by-step approach to warrant safe implementation of large language models in healthcare

Jessica D. Workum[1,2,3], Davy van de Sande[1,3], Diederik Gommers[1,3] and Michel E. van Genderen[1,3]*

[1]Department of Adult Intensive Care, Erasmus MC University Medical Center, Rotterdam, Netherlands, [2]Department of Intensive Care, Elisabeth-TweeSteden Hospital, Tilburg, Netherlands, [3]Erasmus MC Datahub, Erasmus MC University Medical Center, Rotterdam, Netherlands

Large Language Models (LLMs) offer considerable potential to enhance various aspects of healthcare, from aiding with administrative tasks to clinical decision support. However, despite the growing use of LLMs in healthcare, a critical gap persists in clear, actionable guidelines available to healthcare organizations and providers to ensure their responsible and safe implementation. In this paper, we propose a practical step-by-step approach to bridge this gap and support healthcare organizations and providers in warranting the responsible and safe implementation of LLMs into healthcare. The recommendations in this manuscript include protecting patient privacy, adapting models to healthcare-specific needs, adjusting hyperparameters appropriately, ensuring proper medical prompt engineering, distinguishing between clinical decision support (CDS) and non-CDS applications, systematically evaluating LLM outputs using a structured approach, and implementing a solid model governance structure. We furthermore propose the ACUTE mnemonic; a structured approach for assessing LLM responses based on Accuracy, Consistency, semantically Unaltered outputs, Traceability, and Ethical considerations. Together, these recommendations aim to provide healthcare organizations and providers with a clear pathway for the responsible and safe implementation of LLMs into clinical practice.

KEYWORDS

large language models, responsible AI, artificial intelligence, health care quality, access and evaluation, disruptive technology

Introduction

Large language models (LLMs) are artificial intelligence (AI) systems with the inherent capability of processing and interpreting natural language (Thirunavukarasu et al., 2023). LLMs show promise in transforming healthcare, offering a newfound flexibility in that, like a Swiss army knife, one single tool can be used for various applications, including administrative support and clinical decision-making (Schoonbeek et al., 2024; Levra et al., 2024). For example, LLMs can aid clinicians by efficiently summarizing medical records and crafting discharge documents. A recent study by Schoonbeek et al. demonstrated that the GPT-4 model proved to be as complete and correct as the clinician in summarizing clinical notes in preparation for outpatient visits, while being 28 times faster (Schoonbeek et al., 2024). Furthermore, LLMs have shown to offer a level of empathy in responding

to patient questions that could surpass human clinicians (Ayers et al., 2023; Luo et al., 2024). Beyond these administrative or documentation tasks, the application of LLMs in healthcare can be expanded to clinical decision support. For example, when comparing the performance of an LLM to medical-journal readers in diagnosing complex real-world cases, the LLM outperformed its human counterparts with 57% vs. 36% correct diagnoses (Eriksen et al., 2023). These examples represent a mere subset of potential applications of LLMs in healthcare, with the scope continuously expanding at rapid pace.

When used for clinical decision support (CDS), LLMs are likely to be considered a medical device and thus have to adhere to strict legislation, requiring thorough assessment to ensure quality standards (Keutzer and Simonsson, 2020; Jackups, 2023). However, for non-CDS applications, there is a lack of robust frameworks and regulatory oversight to ensure high quality output and responsible use of these models in clinical settings. Furthermore, existing legislations provide little guidance on responsible and safe implementation of LLMs from the healthcare organization or provider's perspective. This problem has also been identified recently the World Health Organization in their report on Ethics and Governance of AI for Health (World Health Organization, 2024). Current existing frameworks remain largely abstract and provide limited practical guidance (Raza et al., 2024). Thus, despite the growing use of LLMs in healthcare, a critical gap persists in clear, actionable guidelines for healthcare organizations and providers to ensure their responsible and safe implementation. In this paper, we propose a practical step-by-step approach, combined with an evaluation framework, to bridge this gap and support healthcare organizations and providers in warranting the responsible and safe implementation of LLMs into healthcare (Figure 1).

(1) Protect patient privacy

LLMs have the potential to inadvertently reveal sensitive information to third parties if this information has been previously used as an input in the LLM (Open et al., 2023). Currently, protective measures (i.e., safeguards) to prevent such data leaks are inconsistent, leaving gaps in privacy protection (Yao et al., 2024). Importantly, in the context of healthcare, adherence to legal frameworks designed to safeguard personal data, such as the General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the U.S., is crucial. These regulations mandate that patient information must not be disclosed to third parties, including the developers or hosts of LLMs. Consequently, publicly available LLMs, which typically log user interactions for the purpose of model improvement (retraining), are not viable for healthcare applications due to the risk of data exposure and misuse. To improve user acceptance, LLMs should ideally be integrated into healthcare Information Technology (IT) ecosystems that host these models locally or on secure hospital-owned cloud servers (Nazari-Shirkouhi et al., 2023). This approach guarantees that patient data are securely maintained within the digital infrastructure of the hospital, thereby reinforcing the confidentiality and privacy of patient data. However, this is often not feasible due to high costs and infrastructure demands. Furthermore, the best performing general-purpose LLMs cannot be deployed locally due to proprietary nature of these models, restricting deployment on local servers. Open-source or smaller language models might be considered, but their performance can be inferior to proprietary LLMs (Wu et al., 2024).

It is paramount that if third party hosts of LLMs are used in healthcare, patient privacy is protected by stablishing a secure way of data transmission and guaranteeing that the data is not retained and the model is not retrained with user data. As such, an application programming interfaces (APIs) can serve as a secure connection between the hospital and the third party LLM host by implementing robust encryption protocols. Importantly, healthcare providers must be aware that they should establish strict contractual agreements with third-party LLM hosts to prevent data retention and ensure that user or patient data is not utilized for model retraining.

(2) Consider healthcare-specific model adaptations

General-purpose LLMs still face performance limitations and may not suffice for complex and specialized healthcare tasks without modifications (Mao et al., 2023). Therefore, specific use cases might benefit from integrating medical domain knowledge in the language model. There are two main ways of doing so: by creating a healthcare specific language model or by adapting an existing LLM with medical domain knowledge, either through retraining or by giving it access to a database with specific medical knowledge.

Benefits of creating healthcare-specific models are that they could address challenges such as fairness, transparency, and data-inconsistency and might perform better for very specific medical domain knowledge (He et al., 2023). An additional benefit is that these models are typically smaller in size, leaving the possibility of running these models locally. However, it appears that the development of general-purpose LLMs is advancing more rapidly than that of healthcare-specific models, likely due to broader investment and scalability. By adapting an existing general-purpose LLM with medical domain knowledge, the performance of LLMs within the medical field increases dramatically (Ferber et al., 2024). This can be achieved either by periodically retraining the model with medical domain knowledge, or through Retrieval Augmented Generation (RAG), a technique that integrates an external knowledge database with an LLM through a pre-constructed index (Ng et al., 2025). Comparing both techniques to a human writer: with retraining, the memory of the writer has been expanded, and with RAG, the writer has continuous access to an up-to-date library of information. With RAG, the LLM is combined with a database of specific medical domain knowledge. The LLM draws information from this database when formulating a response, similar to a search engine. This ensures its responses are aligned with the latest medical knowledge while reducing the risk of hallucinations (Zakka et al., 2024). RAG significantly improves the performance of LLMs for healthcare-specific applications. For example, when connecting a RAG framework to international oncology guidelines, the LLM's response improved from 57% to

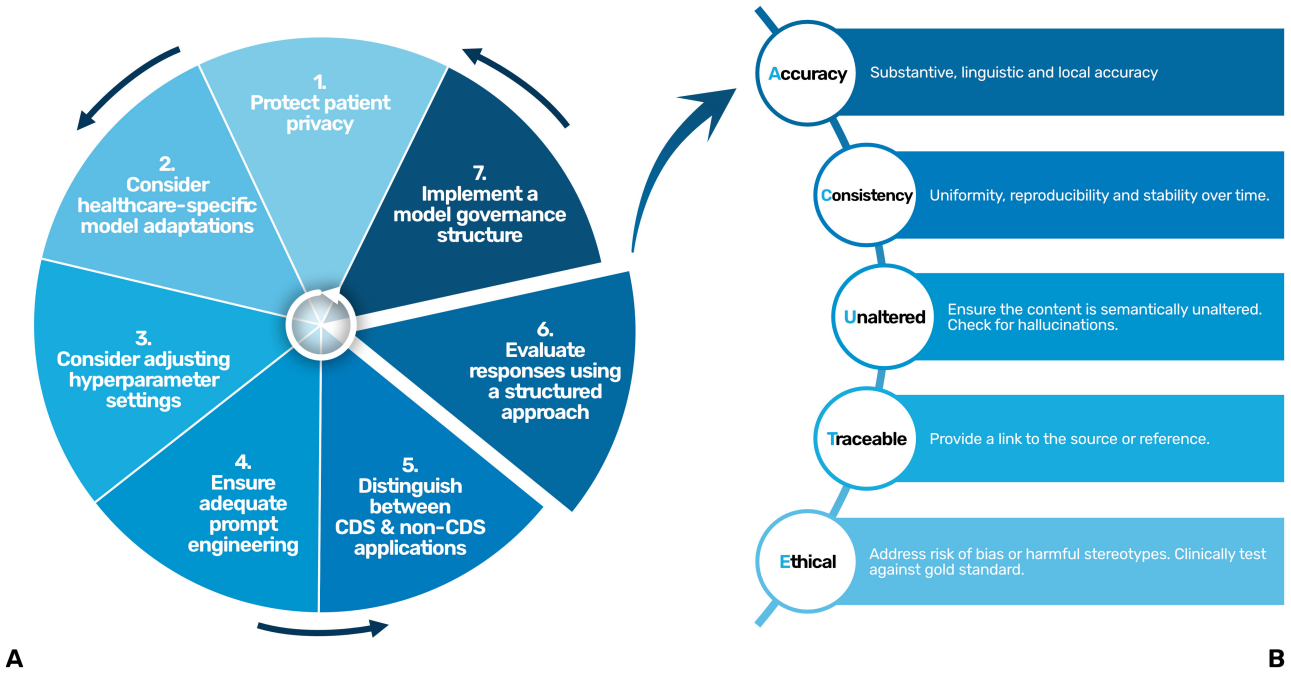

FIGURE 1
A step by step approach **(A)** and a structured evaluation approach (ACUTE) **(B)** for the responsible and safe use and development of LLMs in healthcare. CDS, Clinical Decision Support.

84% in answering questions correctly regarding the management of oncology patients (Ferber et al., 2024). Due to its flexibility, RAG is particularly beneficial in fields where knowledge evolves rapidly, such as medicine.

(3) Consider adjusting hyperparameter settings

Another way of improving an LLM's output is by adjusting its hyperparameters, particularly its temperature setting. The temperature controls between the randomness of the generated responses. Higher temperatures generate more variability, while lower temperatures result in more predictable and consistent responses, adhering more closely to the provided prompts (Pugh et al., 2024). Therefore, it is thought that lower temperatures are recommended when consistency is important, whereas higher temperatures might be useful in addressing ambiguity. However, despite the rationale for adjusting temperature settings based on the specific demands of a clinical use case, recent available evidence suggests that adjustment of temperature has no significant effect on the consistency of performance for various LLMs across different clinical tasks, possibly rendering this step obsolete in the future (Patel et al., 2024).

(4) Ensure adequate prompt engineering

An LLM's output is highly determined by the quality of the instructions or input to the model (prompt). Prompt engineering refers to the practice of designing and implementing prompts and is considered a new discipline within the field of AI. Advanced prompt engineering techniquesimprove the quality of the response of the model significantly (Zhang X. et al., 2024). Examples of advanced prompt engineering techniques are Few-Shot prompting and Chain-of-Thought (CoT) prompting. In Few-Shot prompting, the prompt includes a small number of examples to guide the model's understanding of the task. By providing these task-specific examples, the model is able to produce more accurate responses, even in scenarios where it has not been extensively trained. For example, in answering sample exam questions for the United States Medical Licensing Examination, 5-shot prompting improved the performance for the GPT-4 model from 84% to 87% correct (Nori et al., 2023). In CoT prompting, the model is instructed to engage in step-by-step reasoning by breaking down complex questions into smaller steps (Wei et al., 2022). This structured approach helps the model reason through tasks more effectively, improving coherence and accuracy of the outputs. CoT is especially useful for tasks requiring logical progression, making this technique of particular interest in CDS applications (Miao et al., 2024). However, various other prompt optimization approaches exist, reflecting the rapid evolution of this new discipline (Chang et al., 2024).

Currently, healthcare professionals rely heavily on extensive experimentation using LLMs, with a limited theoretical understanding of why a specific phrasing or formulation of a task is more sensible than others. Inadequate prompt engineering in medicine without strict constraints could lead to undesired outputs, such as (erroneous) medical advice. It is therefore vital that prompts in the medical field should be created by experts in medical prompt engineering (Chen et al., 2024).

(5) Distinguish between CDS and non-CDS applications

Due to regulatory oversight that warrant safe use of innovations in healthcare such as the Medical Device Regulation (MDR) in the European Union (EU) and the Food and Drug Administration (FDA) in the United States, it is important to differentiate between Clinical Decision Support (CDS) and non-CDS for the specific applications of LLMs. This differentiation strongly indicates whether the application is considered to be a medical device, and thus would fall under these specific regulations. CDS is generally understood to be any tool that assists clinicians in diagnostics or treatment decisions, and when it is used to inform clinical decisions that directly impact patient care, it is considered a medical device and would fall under these laws. In contrast, software that only provides supplementary information without driving clinical decisions, is not considered clinical decision support (non-CDS) and thus may not be classified as a medical device.

Consequently, an LLM that supports diagnostic or treatment processes would be classified as a medical device under, for example, the MDR. This prohibits the use of the tool until it has undergone a thorough assessment to ensure that it meets MDR-related quality standards, such as providing clinical evidence of their safety and effectiveness. This process may be time-consuming, possibly limiting the adoption of LLMs for CDS in healthcare.

Unlike traditional medical devices or AI-models, LLMs are inherently multi-purpose, capable of addressing diverse clinical and non-clinical queries. Subjecting LLMs to regulatory approval for each specific clinical purpose is impractical due to the immense effort and cost required. Their rapid evolution, with frequent updates in data, methods, and architectures, further complicates regulation. Regulatory sandboxes offer a supervised setting to explore regulatory requirements and evaluate LLM performance iteratively, providing a flexible pathway to address these challenges.

CDS applications

The use of LLMs for CDS seems very promising. When presented with United States Medical Licensing Examination (USMLE) sample exam questions, the GPT-4 model correctly answered 87% without any healthcare-specific adaptations (Nori et al., 2023). Additionally, on various publicly available benchmark datasets, such as the MedQA and the medical components of the Massive Multitask Language Understanding (MMLU), the GPT-4 model performed outstandingly well, answering over 80% correct for each benchmark (Nori et al., 2023). This indicates that general medical knowledge is inherently present in these models. Fewer studies have researched the capabilities of LLMs for specialized medical knowledge within clinical subdomains. For example, in a recent study, the GPT-4 model was able to correctly answer nephrology questions with a score of 73%, without healthcare-specific adaptations or advanced prompt engineering techniques, indicating its potential for highly specialized fields (Wu et al., 2024). When compared to human physicians, the performance of the GPT-4 model exhibited variation across medical specialties, although the model consistently met or exceeded the examination threshold in the majority of cases (Katz et al., 2024).

When implemented into healthcare, CDS will most likely require the use of healthcare-specific model adaptations, utilizing techniques such as RAG, to improve the accuracy of responses. Relevant references should be linked so that the source of the information can be checked.

However, the progression toward CDS necessitates more than the mere capability to answer clinical questions, as clinical decision-making encompasses a combination of medical knowledge, clinical reasoning, multidisciplinary collaboration, evidence-based practice and communication skills. Current advancements in LLMs, aimed at improving logical reasoning, bring the use of LLMs for CDS closer to fruition. Nevertheless, due to the potential significant impact on clinical decision-making, implementing LLMs for CDS demands tremendous diligence.

Non-CDS applications

The majority of non-CDS applications aims to reduce the administrative load for healthcare providers. Various examples are currently being implemented, such as composing draft responses to patient messages and creating summaries of the patient chart (Schoonbeek et al., 2024; van Veen et al., 2023; Garcia et al., 2024; Tai-Seale et al., 2024). If a use case is not considered CDS, there are currently no laws or guidelines in place to ensure responsible and safe use of LLMs. Given the swift development and adoption of LLMs in society, it is likely that additional non-CDS applications of LLMs are coming to healthcare rapidly.

While legal frameworks such as the EU AI Act, GDPR, and HIPAA establish important baseline requirements for data protection and accountability, they do not address the unique challenges posed by LLMs in clinical settings. For example, they lack requirements for clinical validation, i.e., objectively assessing whether outputs are sufficient for clinical use while accounting for risks like hallucinations, missing information and misinterpretations. These challenges underscore the need for healthcare-specific validation processes to complement existing legal frameworks.

(6) Evaluate using a structured approach

To ensure the responses of the LLM remain accurate, consistent, and aligned with clinical standards over time, a structured approach to evaluate their responses is essential. As LLMs are probabilistic by nature, their performance can vary, making continuous and systematic evaluation critical for maintaining quality and preventing errors, especially in high-stakes environments such as healthcare. Abbasian et al. proposed an extensive set of intrinsic and extrinsic evaluation metrics for assessing the performance of healthcare chatbots, including evaluating the quality of their response (Abbasian et al., 2024). However, their comprehensiveness limits their practicality in clinical settings. To balance comprehensiveness and simplicity, we've identified five key points that should be addressed when

evaluating the response of an LLM in clinical settings, being accuracy, consistency, semantically unaltered, traceable and ethical. The mnemonic "ACUTE" (Figure 1B) could be used as a helpful tool.

Accuracy encompasses three domains: first, substantive accuracy, meaning that responses are factually correct and contextually appropriate within the medical field, even for non-clinical decision support (non-CDS) applications. When determining if a response is substantively accurate, it is important to determine if the response is complete (i.e., determine if there is any information missing) and correct (i.e., determine if there are any factual errors). The second domain is linguistic accuracy, particularly for languages other than English. As foundational models are predominantly trained on English data, responses may exhibit reduced accuracy in other languages. Rigorously test for linguistic accuracy by adjusting the prompt. Frequently, writing the prompt in English and asking the LLM to translates yields better results. The third domain is local accuracy, which means, ensuring that the responses reflect each hospital's own policies and communication preferences.

When deployed in clinical practice, LLM responses need to be reproducible and stable over time, ensuring reliability in their outputs. As such, consistency is another key criterion. If the LLM provides inconsistent results, try adjusting the temperature settings or the prompt. If the inconsistency remains, try a different LLM for this specific clinical task.

The responses should also be semantically Unaltered. The response of LLM should accurately reflect the information presented in the patient chart without introducing extraneous content (hallucinations). Furthermore, the responses should be Traceable, making it clear where the LLM obtained its information, ideally by providing a reference to the source. For example, when utilizing RAG, the source of the information should be cited, and when summarizing notes in the patient chart, after each claim, the original note should be linked.

And lastly, the Ethical dimension mandates the responsible use of LLMs and aims to prevent that LLM responses do not perpetuate biases or harmful stereotypes, ensuring the responsible and fair use of these models in clinical practice. LLMs are typically trained on large datasets that include publicly available text, which often contains inherent biases reflective of societal inequalities. Studies have shown that these biases can perpetuate in LLM outputs, leading to disparities in diagnosis and treatment across different demographic groups. The Benchmark of Clinical Bias in Large Language Models (CLIMB benchmark) highlights how LLMs may exhibit these biases, resulting in unequal diagnostic accuracy across populations (Zhang Y. et al., 2024). Similarly, another study found that LLMs could reinforce harmful stereotypes, such as underdiagnosing conditions like smoking in young males and obesity in middle-aged females (Pal et al., 2023). This emphasizes the need for careful oversight to prevent biased decision-making in clinical practice. Ideally, each new use case should be clinically tested compared to its gold standard, which is generally the performance of the clinician.

TABLE 1 Critical questions to guide responsible LLM implementation in healthcare with actionable steps.

Recommendation	Critical questions
1. Protect patient privacy	How is patient data securely transmitted and stored?
	Are third-party agreements in place to prevent data retention or model retraining?
	Are the LLMs hosted on secure, hospital-controlled infrastructure?
2. Consider healthcare-specific model adaptations	Is medical domain knowledge paramount to the specific use case for which an LLM is deployed?
	Has the LLM been adapted or validated for the specific healthcare tasks it will perform? If so, how?
	Does the application utilize RAG (Retrieval-Augmented Generation) to integrate up-to-date medical knowledge?
3. Consider adjusting hyperparameter settings	Have hyperparameters, such as temperature, been adjusted to align with the specific clinical use case?
	Has the impact of hyperparameter adjustments been adequately evaluated?
4. Ensure adequate prompt engineering	Who is responsible for writing and maintaining the prompts?
	Have medical professionals been involved in designing and testing the prompts?
	Have the prompts been tested and refined in an iterative manner to minimize errors and undesired outputs?
5. Distinguish between CDS and non-CDS applications	Is the application clearly categorized as either Clinical Decision Support (CDS) or non-CDS?
	For CDS applications, does the LLM comply with potentially relevant medical device regulations (e.g., MDR, FDA)?
	For non-CDS applications, are barriers set in place to avoid unintended use as a medical device?
6. Evaluate using a structured approach	Are LLM outputs evaluated using a structured framework, such as the ACUTE criteria (Accuracy, Consistency, Unaltered meaning, Traceability, Ethical considerations)?
	Is there a process for documenting evaluation results and using them to guide improvements?
7. Implement a model governance structure	Is there a dedicated team in place to monitor and oversee LLM performance over time?
	Are evaluations performed regularly to ensure ongoing alignment with clinical standards?
	Are fallback mechanisms established to ensure continuity?

In contrast to existing frameworks that provide broad, cross-sectoral guidelines, the ACUTE framework offers a specialized and practical approach tailored to the unique requirements of healthcare, focusing specifically on evaluating LLM outputs for clinical relevance and patient safety. We believe that using the ACUTE mnemonic as a structured

TABLE 2 Checklist for the ACUTE framework, designed to evaluate LLM outputs in healthcare and ensure that each criterion is addressed effectively to minimize risks and enhance reliability.

Dimension	Criteria	Focus
Accuracy	Are responses factually correct and complete?	Substantive accuracy
	Are responses grammatically correct and clear, even in non-English languages?	Linguistic accuracy
	Do responses align with hospital policies and preferences?	Local accuracy
Consistency	Are responses consistent across repeated prompts?	Reproducibility
	Are responses stable across different sessions and model versions?	Stability over time
	Are inconsistencies addressed effectively through prompt refinements?	Mitigation of inconsistencies
Unaltered	Do responses avoid adding erroneous or fabricated information?	Hallucinations
	Do responses accurately reflect the input data, such as patient charts?	Reflection of source data
Traceability	If applicable, are claims and recommendations clearly linked to credible sources?	Source attribution
	If applicable, are external references provided when RAG or other systems are used?	Use of retrieval systems
Ethical	Do responses avoid perpetuating harmful biases or stereotypes?	Bias avoidance
	Are sensitive topics handled responsibly and respectfully?	Sensitive topics

approach balances simplicity and comprehensiveness for the evaluation of LLM responses and remains practical for real-world clinical use while still adequately addressing key challenges in LLM evaluation and deployment. Comparative analyses utilizing the ACUTE framework should be performed between LLM-generated outputs and clinician outputs for clinical validation.

(7) Implement a model governance structure

Eventually, it is crucial to ensure high quality performance and output of the LLM over time and therefore, a system for regular monitoring and continuous evaluation should be in place. This is of particular importance, as an LLM's performance can vary over time via retraining or is updated to a new version. Thus, establishing a governance framework to monitor the LLM's performance over time and implement adaptive maintenance strategies is crucial. In addition to model governance, robust data governance is essential, ensuring transparent data management and controlled access. Governance principles for both data and models should be traceable, securely stored, and readily accessible to notified bodies and competent authorities to support regulatory compliance. A dedicated team comprising medical and AI experts should be established to collect and evaluate user feedback, interpret model quality and outputs, and implement appropriate actions accordingly. Adaptive maintenance strategies could include periodic audits of LLM outputs and robust fallback mechanisms, such as maintaining access to legacy versions and options for model switching. By incorporating these measures, the governance structure will remain robust and futureproof, safeguarding both safety and reliability over time. The ACUTE framework mentioned in step 6 could offer such guidance.

Connect all the steps

To move toward the safe and responsible development and implementation of LLMs in both administrative tasks and clinical decision support in healthcare, connecting all the steps is essential. For example, by combining different prompt engineering techniques with healthcare-specific model adaptations like RAG the overall performance of an LLM on medical board examinations improves significantly, highlighting the importance of considering the steps outlined in this manuscript (Samaan et al., 2024). As a practical aid, we have transformed the recommendations into "critical questions" in Table 1, and the ACUTE framework into a checklist in Table 2. These critical questions are designed to assess the readiness for responsible LLM implementation in healthcare. If these questions cannot be answered adequately, there is a significant gap that must be addressed prior to utilizing LLMs in healthcare. The ACUTE checklist will help systematically evaluate the performance of an LLM application, while highlighting potential weaknesses.

Ultimately, we must bridge the gap between technological AI model development and trustworthy and responsible AI adoption in a clinical setting. Despite the growing use of LLMs, a critical gap persists in clear, actionable guidelines available to healthcare organizations and providers to ensure their responsible and safe implementation. The integration of a step-by-step approach, combined with a practical evaluation framework, could address this gap. By balancing simplicity with comprehensiveness, these recommendations could lower AI hesitancy, improve clinical implementation and unlock its full potential in improving healthcare. Future researchers are encouraged to validate the proposed framework across diverse clinical scenarios. Advancing the responsible implementation of LLMs in healthcare will require a collective effort from healthcare organizations, providers, researchers, and policymakers to ensure robust validation, responsible use and adequate monitoring of LLMs in clinical practice. The recommendations outlined in this manuscript provide a practical starting point for this collaborative

journey, offering guidance for the responsible and effective implementation of LLMs in healthcare.

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

JW: Conceptualization, Investigation, Visualization, Writing – original draft, Writing – review & editing. DS: Conceptualization, Writing – original draft. DG: Supervision, Writing – review & editing. MG: Conceptualization, 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.

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 Gen AI was used in the creation of this manuscript. Paperpal (version 3.209.2, source: paperpal.com) was utilized solely for the purpose to enhance language.

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

Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Shakeri Hossein Abad, Z., Thieme, A., et al. (2024). Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digit Med.* 7:82. doi: 10.1038/s41746-024-01074-z

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., et al. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* 183, 589–596. doi: 10.1001/jamainternmed.2023.1838

Chang, K., Xu, S., Wang, C., Luo, Y., Liu, X., Xiao, T., et al. (2024). Efficient prompting methods for large language models: a survey. *arXiv [Preprint].* arXiv:2404.01077. doi: 10.48550/arXiv.2404.01077

Chen, S., Guevara, M., Moningi, S., Hoebers, F., Elhalawani, H., Kann, B. H., et al. (2024). The effect of using a large language model to respond to patient messages. *Lancet Digital Health*. 6, e379–e381. doi: 10.1016/S2589-7500(24)00060-8

Eriksen, A. V., Möller, S., and Ryg, J. (2023). Use of GPT-4 to diagnose complex clinical cases. *NEJM AI.* 1, 2023–2025. doi: 10.1056/AIp2300031

Ferber, D., Wiest, I. C., Wölflein, G., Ebert, M. P., Beutel, G., Eckardt, J. N., et al. (2024). GPT-4 for information retrieval and comparison of medical oncology guidelines. *NEJM AI* 1:235. doi: 10.1056/AIcs2300235

Garcia, P., Ma, S. P., Shah, S., Smith, M., Jeong, Y., Devon-Sand, A., et al. (2024). Artificial intelligence–generated draft replies to patient inbox messages. *JAMA Netw. Open* 7:e243201. doi: 10.1001/jamanetworkopen.2024.3201

He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., et al. (2023). A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv [Preprint].* doi: 10.2139/ssrn.4809363

Jackups, R. (2023). FDA regulation of laboratory clinical decision support software: is it a medical device? *Clin. Chem.* 69, 327–329. doi: 10.1093/clinchem/hvad011

Katz, U., Cohen, E., Shachar, E., Somer, J., Fink, A., Morse, E., et al. (2024). GPT versus resident physicians — a benchmark based on official board scores. *NEJM AI* 1:192. doi: 10.1056/AIdbp2300192

Keutzer, L., and Simonsson, U. S. H. (2020). Medical device apps: An introduction to regulatory affairs for developers. *JMIR Mhealth Uhealth.* 8:e17567. doi: 10.2196/17567

Levra, A. G., Gatti, M., Mene, R., Shiffer, D., Costantino, G., Solbiati, M., et al. (2024). A large language model-based clinical decision support system for syncope recognition in the emergency department: a framework for clinical workflow integration. *Eur. J. Intern. Med.* 131, 113–120. doi: 10.1016/j.ejim.2024.09.017

Luo, M., Warren, C. J., Cheng, L., Abdul-Muhsin, H. M., and Banerjee, I. (2024). Assessing empathy in large language models with real-world physician-patient interactions. *arXiv [Preprint].* arXiv:2405.16402. doi: 10.48550/arXiv.2405.16402

Mao, R., Chen, G., Zhang, X., Guerin, F., and Cambria, E. (2023). *GPTEval: A Survey on Assessments of ChatGPT and GPT-4.* Available at: http://arxiv.org/abs/2308.12488 (accessed 23 August 2023).

Miao, J., Thongprayoon, C., Suppadungsuk, S., Krisanapan, P., Radhakrishnan, Y., and Cheungpasitporn, W. (2024). Chain of thought utilization in large language models and application in nephrology. *Medicina (Lithuania)* 60:148. doi: 10.3390/medicina60010148

Nazari-Shirkouhi, S., Badizadeh, A., Dashtpeyma, M., and Ghodsi, R. (2023). A model to improve user acceptance of e-services in healthcare systems based on technology acceptance model: an empirical study. *J. Ambient Intell. Humaniz. Comput.* 14, 7919–7935. doi: 10.1007/s12652-023-04601-0

Ng, K. K. Y., Matsuba, I., and Zhang, P. C. (2025). RAG in health care: a novel framework for improving communication and decision-making by addressing LLM limitations. *NEJM AI* 2:380. doi: 10.1056/AIra2400380

Nori, H., King, N., Mckinney, S. M., Carignan, D., Horvitz, E., and Openai, M. (2023). *Capabilities of GPT-4 on Medical Challenge Problems*. Available at: https://arxiv.org/abs/2303.13375v2 (accessed 28 September 2024).

Open, A. I., Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., et al. (2023). *GPT-4 Technical Report*. Available at: http://arxiv.org/abs/2303.08774 (accessed 15 March 2023).

Pal, R., Garg, H., Patel, S., and Sethi, T. (2023). Bias amplification in intersectional subpopulations for clinical phenotyping by large language models. *MedRxiv [Preprint].* doi: 10.1101/2023.03.22.23287585

Patel, D., Timsina, P., Raut, G., Freeman, R., Levin, M. A., Nadkarni, G. N., et al. (2024). Exploring temperature effects on large language models across various clinical tasks. *medRxiv [Preprint].* doi: 10.1101/2024.07.22.24310824

Pugh, S. L., Chandler, C., Cohen, A. S., Diaz-Asper, C., Elvevåg, B., and Foltz, P. W. (2024). Assessing dimensions of thought disorder with large language models: the tradeoff of accuracy and consistency. *Psychiatry Res.* 341:116119. doi: 10.1016/j.psychres.2024.116119

Raza, M. M., Venkatesh, K. P., and Kvedar, J. C. (2024). Generative AI and large language models in health care: pathways to implementation. *NPJ Digit. Med.* 7:62. doi: 10.1038/s41746-023-00988-4

Samaan, J. S., Margolis, S., Srinivasan, N., Srinivasan, A., Yeo, Y. H., Anand, R., et al. (2024). Multimodal large language model passes specialty board examination

and surpasses human test-taker scores: a comparative analysis examining the stepwise impact of model prompting strategies on performance. *medRxiv*. 2024:10809. doi: 10.1101/2024.07.27.24310809

Schoonbeek, R. C., Workum, J. D., Schuit, S. C. E., Doornberg, J. N., Van Der Laan, T. P., and Bootsma-Robroeks, C. M. H. H.T. (2024). *Completeness, Correctness and Conciseness of Physician-written versus Large Language Model Generated Patient Summaries Integrated in Electronic Health Records*. SSRN. Available at: https://ssrn.com/abstract=4835935

Tai-Seale, M., Baxter, S. L., Vaida, F., Walker, A., Sitapati, A. M., Osborne, C., et al. (2024). AI-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Netw. Open* 2024:E246565. doi: 10.1001/jamanetworkopen.2024.6565

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023). Large language models in medicine. *Nat. Med.* 29, 1930–1940. doi: 10.1038/s41591-023-02448-8

van Veen, D., van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., et al. (2023). Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* 30, 1134–1142. doi: 10.1038/s41591-024-02855-5

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models chain-of-thought prompting. *arXiv [Preprint]*. arXiv:2201.11903. doi: 10.48550/arXiv.2201.11903

World Health Organization (2024). *Ethics and Governance of Artificial Intelligence for Health. Guidance on Large Multi-modal Models*. Geneva: World Health Organization.

Wu, S., Koo, M., Blum, L., Black, A., Kao, L., Fei, Z., et al. (2024). Benchmarking open-source large language models, GPT-4 and Claude 2 on multiple-choice questions in nephrology. *NEJM AI* 1, 1–8. doi: 10.1056/AIdbp2300092

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., and Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. *High-Confid. Comp*. 2024:100211. doi: 10.1016/j.hcc.2024.100211

Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., et al. (2024). Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI* 1:68. doi: 10.1056/AIoa2300068

Zhang, X., Talukdar, N., Vemulapalli, S., Ahn, S., Wang, J., Meng, H., et al. (2024). Comparison of prompt engineering and fine-tuning strategies in large language models in the classification of clinical notes. *medRxiv [Preprint]*. doi: 10.1101/2024.02.07.24302444

Zhang, Y., Hou, S., Derek Ma, M., Wang, W., Chen, M., and Zhao, J. (2024). CLIMB: a benchmark of clinical bias in large language models. Available at: https://github.com/