frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 25 June 2025
DOI 10.3389/fmed.2025.1516442

OPEN ACCESS

EDITED BY
Weihua Yang,
Southern Medical University, China

REVIEWED BY
Xiaolai Zhou,
Sun Yat-sen University, China
Shulin Liu,
First Affiliated Hospital of Chongqing Medical University, China
Yuqing Chen,
Second Military Medical University, China

*CORRESPONDENCE
Ling-Ping Cen
✉ cenlp@hotmail.com

†These authors have contributed equally to this work

RECEIVED 24 October 2024
ACCEPTED 29 May 2025
PUBLISHED 25 June 2025

CITATION
He H-J, Zhao F-F, Liang J-J, Wang Y, He Q-Q, Lin H, Cen J, Chen F, Li T-P, Hu Z, Yang J-F, Chen L, Cheung CY, Tham Y-C and Cen L-P (2025) Evaluation and comparison of large language models' responses to questions related optic neuritis. *Front. Med.* 12:1516442. doi: 10.3389/fmed.2025.1516442

Evaluation and comparison of large language models' responses to questions related optic neuritis

Han-Jie He[1,2†], Fang-Fang Zhao[1†], Jia-Jian Liang[1], Yun Wang[1], Qian-Qian He[1,2], Hongjie Lin[1], Jingyun Cen[3], Feifei Chen[1], Tai-Ping Li[1], Zhanchi Hu[4], Jian-Feng Yang[1], Lan Chen[1], Carol Y. Cheung[5], Yih-Chung Tham[6,7,8] and Ling-Ping Cen[1,2,4,9*]

[1]Joint Shantou International Eye Center of Shantou University and The Chinese University of Hong Kong, Shantou, Guangdong, China, [2]Shantou University Medical College, Shantou, Guangdong, China, [3]Shaoguan University Medical College, Shaoguan, Guangdong, China, [4]Dongguan Guangming Eye Hospital, Dongguan, Guangdong, China, [5]Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China, [6]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore, [7]Centre of Innovation and Precision Eye Health, Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore, [8]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore, [9]Guangdong Provincial Key Laboratory of Medical Immunology and Molecular Diagnostics, The First Dongguan Affiliated Hospital, School of Medical Technology, Guangdong Medical University, Dongguan, China

Objectives: Large language models (LLMs) show promise as clinical consultation tools and may assist optic neuritis patients, though research on their performance in this area is limited. Our study aims to assess and compare the performance of four commonly used LLM-Chatbots—Claude-2, ChatGPT-3.5, ChatGPT-4.0, and Google Bard—in addressing questions related to optic neuritis.

Methods: We curated 24 optic neuritis-related questions and had three ophthalmologists rate the responses on two three-point scales for accuracy and comprehensiveness. We also assessed readability using four scales. The final results showed performance differences among the four LLM-Chatbots.

Results: The average total accuracy scores (out of 9): ChatGPT-4.0 (7.62 ± 0.86), Google Bard (7.42 ± 1.20), ChatGPT-3.5 (7.21 ± 0.70), Claude-2 (6.44 ± 1.07). ChatGPT-4.0 (p = 0.0006) and Google Bard (p = 0.0015) were significantly more accurate than Claude-2. Also, 62.5% of ChatGPT-4.0's responses were rated "Excellent," followed by 58.3% for Google Bard, both higher than Claude-2's 29.2% (all $p \leq 0.042$) and ChatGPT-3.5's 41.7%. Both Claude-2 and Google Bard had 8.3% "Deficient" responses. The comprehensiveness scores were similar among the four LLMs (p = 0.1531). Note that all responses require at least a university-level reading proficiency.

Conclusion: Large language models-Chatbots hold immense potential as clinical consultation tools for optic neuritis, but they require further refinement and proper evaluation strategies before deployment to ensure reliable and accurate performance.

KEYWORDS

eye diseases, optic nerve diseases, optic neuritis, artificial intelligence, natural language processing

1 Introduction

Recent advancements in artificial intelligence (AI) have unlocked limitless possibilities for transforming medicine. Thanks to machine learning and deep learning technologies, AI has shown tremendous potential in healthcare (1). Currently, key applications of AI in medicine include enhancing interaction and communication, improving image recognition, supporting diagnostics and nursing, optimizing healthcare management and administrative processes, and assisting in surgeries and drug development.

Large language models (LLMs) are AI systems based on neural network architectures and use deep learning models, trained on extensive databases for natural language processing tasks. These models possess human-like language capabilities, offering substantial benefits to healthcare professionals and patients (2). Among them, ChatGPT, a generative AI developed by OpenAI (San Francisco, CA, United States), stands out for its widespread use as an AI chatbot by the general public. It leverages vast amounts of internet text data to produce coherent responses tailored to specific inputs (3). Unlike traditional search engines, ChatGPT and similar chatbots excel in simplicity, specificity, and interactivity, sparking increasing interest in their potential for medical consultations (4).

The application of large language models in ophthalmology is increasingly prevalent. Chatbots are utilized to assess proficiency in ophthalmology (4–8), educate clinical medical students, (9) assist in diagnostic processes for clinicians (10, 11), perform image diagnostics (12), manage clinical electronic records (13), educate patients (14) and aid in personalized patient management (14, 15). Notably, one study showed that ChatGPT achieved an accuracy rate above 90% on the Taiwanese medical licensing examination, yet it exhibited the highest error rate (28.95%) in ophthalmology-related questions (8). Moreover, while ChatGPT performed adequately in general ophthalmology queries, it showed weaknesses in neuro-ophthalmology and ocular pathology (5). The efficacy of LLM-Chatbots in addressing optic neuritis-specific questions remains unexplored. Given their potential role as assistants to doctors and patients, particularly in neuro-ophthalmology, there is a pressing need for further exploration of chatbots in this field.

Although the incidence of optic neuritis varies across regions and ethnic groups, it is reported that there are still about 4–8 per 100,000 person years globally (16). Due to the specialized expertise required to diagnose and treat optic neuritis, many patients struggle to access timely medical consultations with well-trained doctors. In such cases, more accessible alternatives like online consultations or chatbots become increasingly appealing (17). However, the effectiveness of chatbots in managing optic neuritis-related inquiries needs thorough evaluation.

In this study, we will compare the accuracy, comprehensiveness, and readability of four widely used and openly accessible LLM-Chatbots—ChatGPT-3.5, ChatGPT-4.0, Google Bard (now updated to Google Gemini), and Claude-2 (now updated to Claude-3) in clinical consultations for optic neuritis. Our findings will provide valuable insights into the effectiveness of LLM-Chatbots in clinical consultations for optic neuritis.

2 Materials and methods

In this study, we compared the accuracy, comprehensiveness, and readability of responses to optic neuritis-related questions generated by four LLM-Chatbots. The study was conducted from 23 January 2024 to 4 March 2024. Given that the study did not involve any patients or animals, approval from an ethics committee was not required.

2.1 Study design

Questions on optic neuritis were collaboratively developed by clinical ophthalmologists (LPC, HJH, FFZ), based on common issues faced by patients in clinical settings, frequently asked questions on online platforms, and authoritative information from esteemed health websites, such as the National Eye Institute and the American Academy of Ophthalmology (18, 19). Upon aggregating all relevant information, the three doctors, leveraging their clinical experience, consolidated 24 questions related to optic neuritis, categorized into four groups: general, diagnosis, treatment, and follow-up and prevention. This categorization aimed to assess the varied performance of different LLM-Chatbots across question types. For the study, conducted from 23 January to 30 January 2024, we utilized four LLMs: Claude-2 (Anthropic, San Francisco, California), ChatGPT (versions GPT-3.5 and GPT-4.0, OpenAI, San Francisco, California), and Google Bard (Google, Mountain View, California). While Claude-2, ChatGPT-3.5, and Google Bard are freely available, ChatGPT-4.0 requires a paid subscription. Nevertheless, given its enhanced performance in neuro-ophthalmology and its relative affordability and ease of use for patients, ChatGPT-4.0 was included in our study (5, 7).

Each of the 24 optic neuritis-related questions was directly entered into four LLM-Chatbots using separate, newly opened windows to prevent interference and preserve response integrity,

with no specialized prompts employed in the process. A research member (HJH) uniformly collected all responses (Supplementary Tables 1–4), formatted them into plain text, and removed any identifiable features of each LLM-Chatbot without altering the main content. This ensured that evaluators could not determine which LLM-Chatbot produced the replies. To minimize potential biases across evaluations, three rounds of accuracy assessment were conducted, spaced 48 hours apart, with the sequence of responses rearranged before each round.

2.2 Evaluation of readability

Four validated readability scales (Supplementary Table 5) were used to assess responses from all LLM-Chatbots to optic neuritis-related questions (20–22), as well as authoritative online consultations accessible to patients, including content from the Mayo Clinic, Cleveland Clinic, and American Academy of Ophthalmology. These assessments utilized the Gunning Fog Index, Flesch-Kincaid Grade Level, Simple Measure of Gobbledygook (SMOG) score, and Coleman-Liau Index. Each of these scales measures word length, syntactic complexity, and sentence length, assigning a United States academic grade level necessary for comprehension. We calculated the readability scores for each response using a freely available online tool (23).

2.3 Evaluation of accuracy

The assessment team was made up of three neuro-ophthalmologists (FFZ, TPL, YW), each boasting at least 6 years of clinical experience. The assessment team was made up of three neuro-ophthalmologists (FFZ, TPL, YW), each boasting at least 6 years of clinical experience. Although the evaluators are non-native English speakers and may have certain shortcomings in understanding the linguistic nuances and cultural background of English content, the LLM-generated output on optic neuritis is predominantly medical (involving fewer complex cultural or idiomatic language issues), and importantly, all three hold medical master's degrees and possess strong proficiency in ophthalmic medical English. They routinely use professional English in their clinical practice, which qualifies them to serve as evaluators. To ensure impartiality, these evaluators were not informed beforehand which LLM-Chatbot provided the responses. They independently assessed the accuracy of the replies using a three-point scale:

1. "Deficient": Indicates responses that could significantly mislead and potentially harm patients due to inaccuracies.
2. "Marginal": Signifies responses that contain possible factual errors, but with a lower risk of misleading or harming patients.
3. "Excellent": Represents responses that are free from errors.

Ratings were determined based on a majority rule approach, where the responses from the LLM-Chatbots were assigned a rating after the majority rule was applied in each assessment round. If a discrepancy occurred among the three grading doctors' opinions, the response was classified as "Pending." After three evaluation rounds, each reply received three ratings. A final rating was established through the majority rule process (Supplementary Table 7). If this final rating remained "Pending," a senior doctor (LPC) would then provide a conclusive rating.

2.4 Evaluation of comprehensiveness

Responses from chatbots deemed "Excellent" in accuracy will undergo further evaluation for comprehensiveness by the assessment panel. The evaluation utilizes a three-tiered scale:

1. Incomplete: Responses lack crucial key information necessary for completeness.
2. Comprehensive: Responses include all essential key information required.
3. Highly Comprehensive: Responses not only provide all key information but also include additional useful details that were not anticipated.

2.5 Statistical analysis

Statistical analyses for this study were conducted using GraphPad Prism (version 8.3.0). Descriptive statistics are presented as mean values and standard deviations (SD). For the parametric data, readability scores were analyzed using one-way ANOVA followed by Tukey's multiple comparison post-hoc test across the four LLM-Chatbots and the overall web content. For non-parametric data, the Kruskal-Wallis Rank Sum test and Dunn's multiple comparison post-hoc test were employed to evaluate the total accuracy and comprehensiveness scores across the four models. Additionally, a two-tailed Pearson's χ^2 was used to assess the distribution of accuracy ratings among the chatbots. The Bonferroni correction method was applied to adjust p-values for multiple comparisons, with a p-value below 0.05 considered statistically significant.

3 Results

3.1 Summary of response lengths

Table 1 presents the responses of all LLMs to optic neuritis questions. The average word count ± SD was: Claude-2 (220.29 ± 20.88), ChatGPT-3.5 (238.75 ± 71.36), Google Bard (299.75 ± 99.41), and ChatGPT-4.0 (269.25 ± 62.96). The average character count ± SD was: Claude-2 (1181.46 ± 140.96), ChatGPT-3.5 (1316.46 ± 410.83), Google Bard (1711.96 ± 576.22), and ChatGPT-4.0 (1461.63 ± 348.74). The average sentence count ± SD was: Claude-2 (15.42 ± 3.12), ChatGPT-3.5 (13.25 ± 4.08), Google Bard (15.46 ± 5.41), and ChatGPT-4.0 (13.83 ± 4.10).

3.2 Readability

Figure 1 shows the average readability scores (Gunning Fog, Flesch-Kincaid, Coleman-Liau and SMOG; see

10.3389/fmed.2025.1516442

TABLE 1 Overview of response length from large language models (LLM)-Chatbots to optic neuritis-related questions.

LLM	Response (words)			Response (characters)			Response (sentences)		
	Mean (SD)	Minimum	Maximum	Mean (SD)	Minimum	Maximum	Mean (SD)	Minimum	Maximum
Claude-2	220.29 (20.88)	171	258	1181.46 (140.94)	904	1453	15.42 (3.12)	11	22
ChatGPT-3.5	238.75 (71.36)	127	413	1316.46 (410.83)	729	2227	13.25 (4.08)	7	20
Google Bard	299.75 (99.41)	41	470	1711.96 (576.22)	214	2569	15.46 (5.41)	3	23
ChatGPT-4.0	269.25 (62.96)	144	390	1461.63 (348.74)	722	2096	13.83 (4.10)	7	26

Supplementary Table 6) for LLMs' responses and professional web content on optic neuritis. The LLMs' averages were: Claude-2, 12.47 ± 1.93; Google Bard, 13.64 ± 2.33; ChatGPT-4.0, 14.75 ± 2.11; and ChatGPT-3.5, 15.37 ± 1.60—all at the college level. The web content averaged 11.41 ± 1.73 (college level) (one-way ANOVA, $p = 0.0188$). ChatGPT-3.5 (Tukey's *post hoc*, $p = 0.0172$) and ChatGPT-4.0 (Tukey's *post hoc*, $p = 0.0407$) had significantly higher readability scores than the web content.

3.3 Accuracy

Figure 2 depicts the average overall accuracy scores for optic neuritis responses from each LLM, as rated by three neuro-ophthalmologists over three rounds. ChatGPT-4.0 scored highest (7.62 ± 0.86), significantly outperforming Claude-2 (6.44 ± 1.07, Dunn's post-hoc test, p = 0.0006). Google Bard ranked second (7.42 ± 1.20; p = 0.0015 compared to Claude-2), followed by ChatGPT-3.5 (7.21 ± 0.70). Detailed scores for each question are in Supplementary Table 7.

Figure 3 presents the final ratings for optic neuritis responses from each LLM after three rounds. ChatGPT-4.0 had 62.5% "Excellent" responses and Google Bard had 58.3%, both significantly higher than Claude-2's 29.2% (Pearson's chi-squared test, all $p \leq 0.042$). ChatGPT-3.5 had 41.7% "Excellent." "Deficient" ratings: 8.3% for Claude-2 and Google Bard, compared to 0% for both ChatGPT models. Detailed ratings for each LLM are in Supplementary Table 7.

Table 2 illustrates the rating distributions for LLMs' responses to optic neuritis. ChatGPT-4.0 excelled in all categories (0% "Deficient"). ChatGPT-3.5 had no "Deficient" ratings but more "Marginal" ratings. Google Bard performed well in diagnosis and follow-up and prevention but had "Deficient" ratings in general and treatment. Claude-2 showed multiple Marginal ratings, two Deficient in treatment and no "Excellent" in follow-up and prevention.

3.4 Comprehensiveness

Supplementary Table 8 shows "Excellent" response comprehensiveness scores. All chatbots performed similarly: Claude-2 (2.67 ± 0.34), ChatGPT-3.5 (2.43 ± 0.39), Google Bard (2.71 ± 0.34), and ChatGPT-4.0 (2.74 ± 0.19). No significant differences were found (Kruskal-Wallis test, $p = 0.1531$).

4 Discussion

Our study conducted a rigorous evaluation of four widely-used LLM-Chatbots—Claude-2, ChatGPT-3.5, Google Bard, and ChatGPT-4.0—on their handling of optic neuritis-related questions. We sourced common questions from multiple venues and had them input systematically into the chatbots by professional neuro-ophthalmologists. Responses were anonymized and randomized before being assessed across three rounds by experienced doctors, with senior doctors resolving any inconsistencies. Responses rated as "Excellent" were

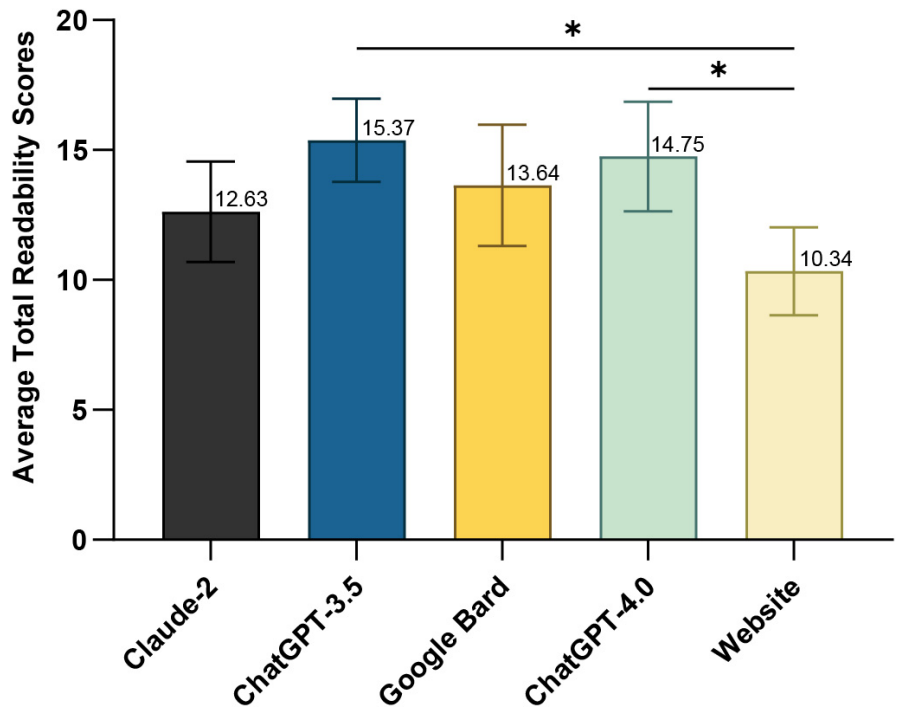

FIGURE 1
Average total readability scores of responses generated by large language models (LLM)-Chatbots and official website content. *$P \leq 0.05$.

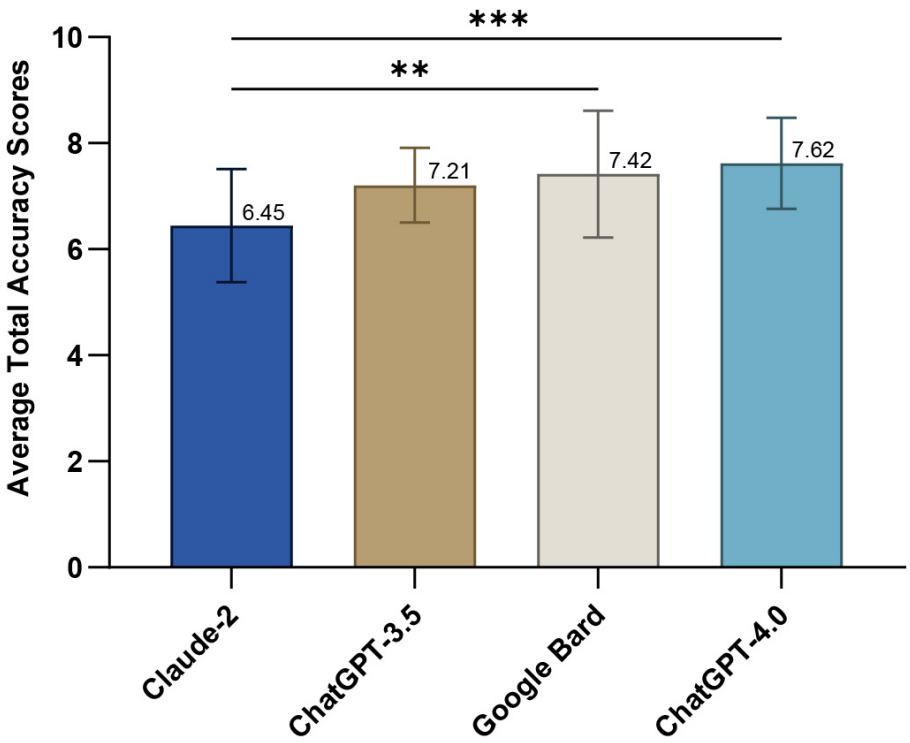

FIGURE 2
Average total accuracy scores of responses generated by large language models (LLM)-Chatbots. **$P \leq 0.01$; ***$P \leq 0.001$.

further examined for comprehensiveness. We also evaluated the readability of outputs from the LLM-Chatbots and established medical websites using an online tool. While previous research has explored LLM-Chatbots' role in neuro-ophthalmology, such as producing patient handouts (24), comparing their responses with human experts on neuro-ophthalmology issues

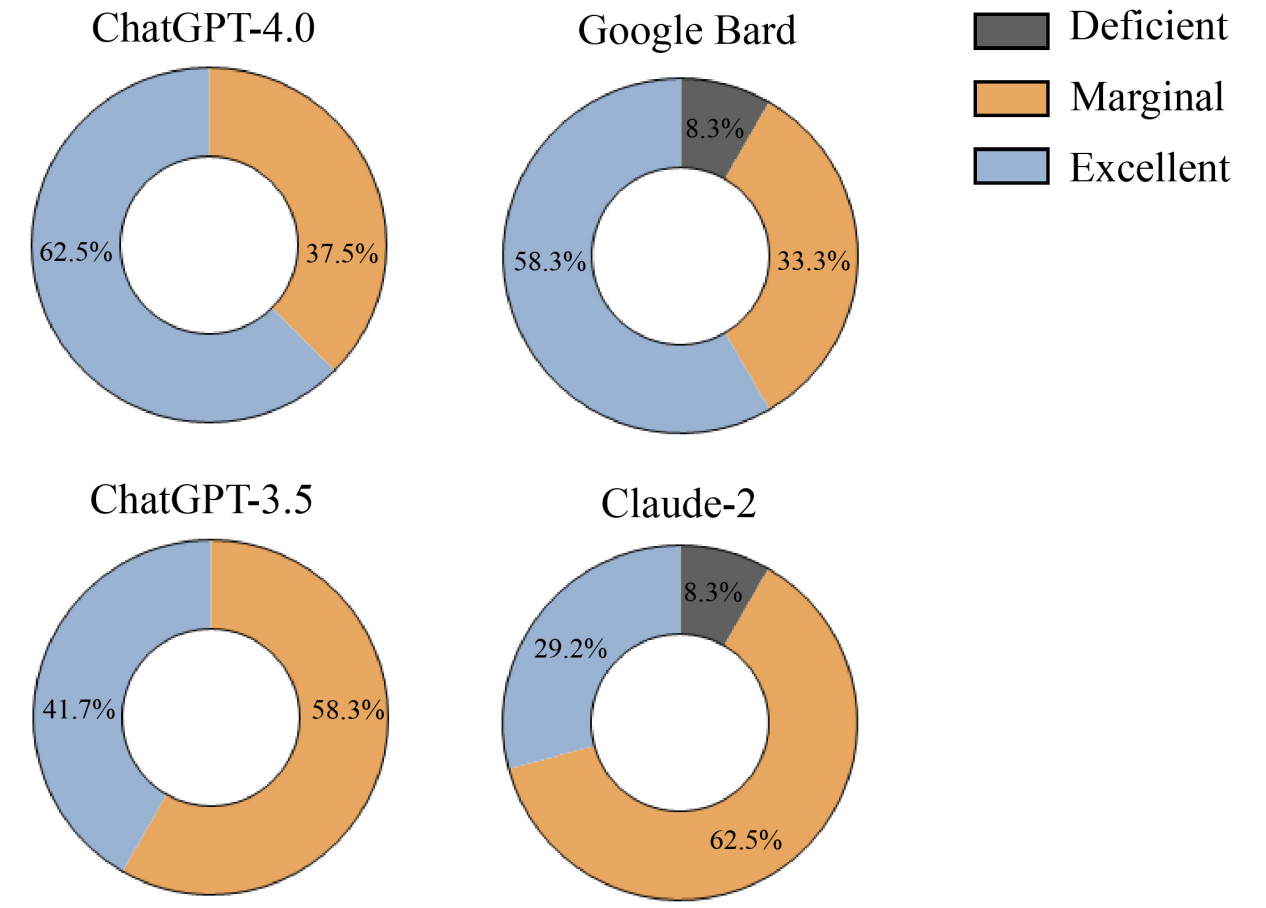

FIGURE 3
Final rating of responses generated by large language models (LLM)-Chatbots determined by the majority rule.

(25), and neuro-ophthalmic disease diagnosis (26), no prior studies have evaluated these chatbots on the three key aspects of accuracy, comprehensiveness, and readability for optic neuritis-specific questions—essentially what online patients need the most. Our findings could significantly enhance the use of LLM-Chatbots in neuro-ophthalmology, potentially establishing them as a new avenue for online consultations on optic neuritis, thereby underscoring our study's substantial practical importance.

Regarding readability, both LLM-Chatbot responses and the content from accessible authoritative websites require a college-level reading proficiency, as indicated in Supplementary Table 6. This is considerably higher than the sixth-grade or lower level recommended by The American Medical Association (AMA) (27). This discrepancy echoes previous findings where online patient education materials (PEMs) on major ophthalmology websites significantly exceeded recommended reading levels (28). Poor readability of LLM-generated responses will diminish their utility in optic nerve clinical consultations, as patients who cannot comprehend the information—even if highly accurate—cannot benefit from it. For patients with lower health literacy, low-readability responses may lead to misunderstandings of medical information and even delay treatment. The relatively poor readability of LLM-Chatbots compared to standard PEMs may be attributed to the LLMs being trained on vast databases, including texts from specialized ophthalmology websites (1). Moreover, the highly specialized and somewhat niche nature of optic neuritis-related content means that LLMs trained with such information undoubtedly necessitate a higher reading level. Our research highlights the challenges LLMs face in balancing accuracy and readability. For instance, Claude-2 has lower accuracy but better readability, while ChatGPT-4.0 is the opposite. Contrary to other studies suggesting ChatGPT-4.0's superior readability among LLMs, our findings suggest otherwise (20). Given that LLMs have the potential to simplify complex information, patients with lower educational levels could benefit by requesting simplified responses, thus maintaining content quality while making it more accessible (20, 22). This approach could be particularly useful for patients using LLM tools to address optic neuritis-related inquiries, guiding them in leveraging these technologies effectively.

In addressing questions related to optic neuritis, ChatGPT-4.0 demonstrates a significant advantage, achieving the highest average accuracy score and the most "Excellent" rated responses (Figures 2, 3). Google Bard closely follows, with performance nearly matching that of ChatGPT-4.0. ChatGPT-3.5 ranks in the middle, while Claude-2 shows the least favorable performance. Regarding comprehensiveness, the four LLM-Chatbots have successfully balanced accuracy and comprehensiveness, with their average scores all exceeding 2, thus achieving at least

TABLE 2 The percentage of ratings generated by each large language models (LLM)-Chatbots in different categories.

Category	Num-ber of quest-ions	Claude-2, *n* (%)			ChatGPT-3.5, *n* (%)			Google Bard, *n* (%)			ChatGPT-4.0, *n* (%)		
		Deficient	Marginal	Excellent	Deficient	Marginal	Excellent	Deficient	Marginal	Excellent	Deficie-t	Marginal	Excellent
General	6	0 (0)	4 (66.7)	2 (33.3)	0 (0)	3 (50)	3 (50)	1 (16.7)	2 (33.3)	3 (50)	0 (0)	2 (33.3)	4 (66.7)
Diagnose	5	0 (0)	3 (60)	2 (40)	0 (0)	3 (60)	2 (40)	0 (0)	1 (20)	4 (80)	0 (0)	2 (40)	3 (60)
Treatment	6	2 (33.3)	1 (16.7)	3 (50)	0 (0)	3 (50)	3 (50)	1 (16.7)	1 (16.7)	4 (66.7)	0 (0)	2 (33.3)	4 (66.7)
Follow-up and prevention	7	0 (0)	7 (100)	0 (0)	0 (0)	5 (71.4)	2 (28.6)	0 (0)	4 (57.1)	3 (42.9)	0 (0)	3 (42.9)	4 (57.1)

a "Comprehensive" rating (Supplementary Table 8). Our results corroborate earlier studies, indicating that ChatGPT-4.0 consistently outperforms other LLMs in the medical field, particularly in neuro-ophthalmology-related inquiries (5, 7, 29). The superior performance of ChatGPT-4.0 can be attributed to its enhanced model size and parameters, its expanding user base, and the incorporation of reinforcement learning from human feedback (RLHF), which helps in generating more relevant and contextually accurate responses (30, 31).

Our study reveals that while the majority of responses from the four models to optic neuritis-related questions were rated "Marginal" or better, Claude-2 and Google Bard each had responses categorized as "Deficient" in the general and treatment categories (Supplementary Table 7). This indicates that using LLMs to provide medical advice increases the risk of misleading information, especially for responses that are not rated as "Excellent." Moreover, because most patients currently have access only to general-purpose LLMs, which have not received formal medical certification, their use may also raise legal and ethical concerns. For example, the clinical use of non-certified LLMs raises ethical concerns, because although LLMs are ethically prohibited from providing harmful information, erroneous medical advice may still indirectly harm patients, leading to delayed treatment or inappropriate self-medication. Given the swift advancement and intricate deployment of LLMs, the ethical challenges highlighted above are unlikely to be adequately safeguarded by current laws and regulations. Notably, Google Bard sometimes includes source links in its responses, but these links are often fabricated and lack authenticity (Supplementary Table 3). Previous research indicates that Google Bard has a tendency to generate fictitious or incorrect information (32), an issue that remains unresolved. Similarly, when ChatGPT-4.0 is prompted to provide sources, it might face the same problem (33). Therefore, caution is advised when considering the source information provided by LLMs.

Unlike traditional search engines, LLMs benefit from deep learning capabilities, continuously enhancing their knowledge from diverse online databases and user feedback. This highlights the significant potential of LLM-Chatbots in clinical settings. Numerous studies have explored LLM applications in various medical fields. For instance, Lim et al. (34) identified potential in handling consultations related to myopia, particularly with ChatGPT-4.0. Meng et al. (35) found that ChatGPT can provide appropriate responses to fracture prevention and medical queries. However, as Cappellani et al. (14) noted, ChatGPT can still generate incomplete, incorrect, or potentially harmful information about common ophthalmic diseases, reflecting the variable performance of LLMs across different medical fields. This variability is largely influenced by the specificity and development of those fields—the richer and more frequent the user interactions, the more effectively LLMs can learn and improve their performance.

Our study has several limitations. First, the small number of optic neuritis-related questions and the limited variety within categories might not fully represent the issues patients usually face, indicating a need for more diverse questions in future research. Additionally, to mitigate evaluator subjectivity, we used multiple evaluation rounds and majority rule decisions.

While readability metrics help assess the educational level needed for understanding, they don't encompass all comprehension factors. Lastly, the rapid evolution of LLMs, driven by new training data and user feedback, means our findings are time-sensitive, when new models emerge, longitudinal re-evaluation studies will be extremely valuable.

Our study confirms the potential of LLM-Chatbots to provide online clinical consultations for optic neuritis, offering accurate and comprehensive information across distances. However, their readability issues might affect user experience. More critically, any misinformation from LLM-Chatbots could lead to unforeseen harmful consequences. Patients using LLM-Chatbots need to proceed with caution and maintain open communication with their doctors, who in turn should guide their use of these tools effectively. Enhancing the readability, accuracy, and comprehensiveness of LLM-Chatbots is essential. A table summarizing the main findings can be found in Supplementary Table 9.

Future research should concentrate on refining assessment strategies for LLMs by developing more comprehensive scoring criteria. Additionally, ongoing training and targeted improvements are crucial to enhance the accuracy and readability of LLMs. Such efforts will ensure that their performance in addressing questions related to optic neuritis becomes increasingly robust and reliable.

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

H-JH: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – original draft. F-FZ: Investigation, Methodology, Resources, Writing – review and editing. J-JL: Investigation, Resources, Writing – review and editing. YW: Investigation, Writing – review and editing. Q-QH: Resources, Writing – review and editing. HL: Writing – review and editing. JC: Writing – review and editing. FC: Writing – review and editing. T-PL: Writing – review and editing. ZH: Writing – review and editing. J-FY: Writing – review and editing. LC: Writing – review and editing. CC: Writing – review and editing. Y-CT: Writing – review and editing. L-PC: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This study was supported by the National Natural Science Foundation of China (81570849), and the Natural Science Foundation of Guangdong Province, China (2020A1515011413).

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) verify and take full responsibility for the use of generative AI in the preparation of this manuscript. Generative AI was used during the preparation of this work the authors used ChatGPT-4.0 in order to edit the entire article, correct grammatical errors, and enhance sentence coherence and academic style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

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.2025.1516442/full#supplementary-material

References

1. Li Z, Wang L, Wu X, Jiang J, Qiang W, Xie H, et al. Artificial intelligence in ophthalmology: The path to the real-world clinic. *Cell Rep Med.* (2023) 4:101095. doi: 10.1016/j.xcrm.2023.101095

2. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. *Nat Med.* (2019) 25:24–9. doi: 10.1038/s41591-018-0316-z

3. OpenAI. *Introducing ChatGPT*. (2024). Available online at: https://openai.com/blog/chatgpt (accessed April 16, 2024).

4. Tan S, Xin X, Wu D. ChatGPT in medicine: Prospects and challenges: A review article. *Int J Surg.* (2024) 110:3701–6. doi: 10.1097/JS9.0000000000001312

5. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmol Sci.* (2023) 3:100324. doi: 10.1016/j.xops.2023.100324

6. Antaki F, Milad D, Chia M, Giguère C, Touma S, El-Khoury J, et al. Capabilities of GPT-4 in ophthalmology: An analysis of model entropy and progress towards

human-level medical question answering. *Br J Ophthalmol.* (2023) 108:1371–8. doi: 10.1136/bjo-2023-324438

7. Teebagy S, Colwell L, Wood E, Yaghy A, Faustina M. Improved performance of ChatGPT-4 on the OKAP examination: A comparative study with ChatGPT-3.5. *J Acad Ophthalmol.* (2023) 15:e184–7. doi: 10.1055/s-0043-1774399

8. Lin S, Chan P, Hsu W, Kao C. Exploring the proficiency of ChatGPT-4: An evaluation of its performance in the Taiwan advanced medical licensing examination. *Digit Health.* (2024) 10:20552076241237678. doi: 10.1177/20552076241237678

9. Momenaei B, Mansour H, Kuriyan A, Xu D, Sridhar J, Ting D, et al. ChatGPT enters the room: What it means for patient counseling, physician education, academics, and disease management. *Curr Opin Ophthalmol.* (2024) 35:205–9. doi: 10.1097/ICU.0000000000001036

10. Zandi R, Fahey J, Drakopoulos M, Bryan J, Dong S, Bryar P, et al. Exploring diagnostic precision and triage proficiency: A comparative study of GPT-4 and bard in addressing common ophthalmic complaints. *Bioengineering.* (2024) 11:120. doi: 10.3390/bioengineering11020120

11. Delsoz M, Raja H, Madadi Y, Tang A, Wirostko B, Kahook M, et al. The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. *Ophthalmol Ther.* (2023) 12:3121–32. doi: 10.1007/s40123-023-00805-x

12. Mihalache A, Huang R, Popovic M, Patil N, Pandya B, Shor R, et al. Accuracy of an artificial intelligence Chatbot's interpretation of clinical ophthalmic images. *JAMA Ophthalmol.* (2024) 142:321–6. doi: 10.1001/jamaophthalmol.2024.0017

13. Ittarat M, Cheungpasitporn W, Chansangpetch S. Personalized care in eye health: Exploring opportunities, challenges, and the road ahead for Chatbots. *J Pers Med.* (2023) 13:1679. doi: 10.3390/jpm13121679

14. Cappellani F, Card K, Shields C, Pulido J, Haller J. Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients. *Eye.* (2024) 38:1368–73. doi: 10.1038/s41433-023-02906-0

15. Bernstein I, Zhang Y, Govil D, Majid I, Chang R, 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

16. Del Negro I, Pauletto G, Verriello L, Spadea L, Salati C, Ius T, et al. Uncovering the genetics and physiology behind optic neuritis. *Genes.* (2023) 14:2192. doi: 10.3390/genes14122192

17. Pan X, Zhou X, Yu L, Hou L. Switching from offline to online health consultation in the post-pandemic era: The role of perceived pandemic risk. *Front Public Health.* (2023) 11:1121290. doi: 10.3389/fpubh.2023.1121290

18. American Academy of Ophthalmology *American Academy of Ophthalmology [Internet]*. (2024). Available online at: https://www.aao.org/search/results?q=optic%20neuritis&realmName=_UREALM_&wt=json&rows=10&start=0 (accessed April 13, 2024).

19. National Eye Institute *Search Results for "Optic Neuritis"*. (2024). Available online at: https://www.nei.nih.gov/search?terms=optic%20neuritis (accessed April 13, 2024).

20. Srinivasan N, Samaan J, Rajeev N, Kanu M, Yeo Y, 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

21. Herbert A, Nemirovsky A, Hess D, Walter D, Abraham N, Loeb S, et al. An evaluation of the readability and content-quality of pelvic organ prolapse youtube transcripts. *Urology.* (2021) 154:120–6. doi: 10.1016/j.urology.2021.03.009

22. Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. *JAMA Netw Open.* (2023) 6:e2336997. doi: 10.1001/jamanetworkopen.2023.36997

23. Readability Formulas *Free Readability Assessment Tools [Internet]*. (2024) Available online at: https://readabilityformulas.com/ (accessed April 16, 2024).

24. Tao B, Handzic A, Hua N, Vosoughi A, Margolin E, Micieli J. 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

25. Tailor P, Dalvin L, Starr M, Tajfirouz D, Chodnicki K, Brodsky M, et al. A comparative study of large language models, human experts, and expert-edited large language models to neuro-ophthalmology questions. *J Neuroophthalmol.* (2024) 45:71–7. doi: 10.1097/WNO.0000000000002145

26. Madadi Y, Delsoz M, Lao P, Fong J, Hollingsworth T, Kahook M, et al. ChatGPT assisting diagnosis of neuro-ophthalmology diseases based on case reports. *J Neuroophthalmol.* (2023):doi: 10.1097/WNO.0000000000002274 [Epub ahead of print].

27. Weiss B. *Health Literacy and Patient Safety: Help Patients Understand*. Chicago, IL: American Medical Association Foundation and American Medical Association (2007).

28. Huang G, Fang C, Agarwal N, Bhagat N, Eloy J, Langer P. Assessment of online patient education materials from major ophthalmologic associations. *JAMA Ophthalmol.* (2015) 133:449–54. doi: 10.1001/jamaophthalmol.2014.6104

29. Raimondi R, Tzoumas N, Salisbury T, Di Simplicio S, Romano M. Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. *Eye.* (2023) 37:3530–3. doi: 10.1038/s41433-023-02563-3

30. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. *GPT-4 Technical Report [Internet]*. (2024). Available online at: http://arxiv.org/abs/2303.08774 (accessed April 13, 2024).

31. ChatGPT Statistics and User Numbers *OpenAI Chatbot [Internet]*. (2024). Available online at: https://www.tooltester.com/en/blog/chatgpt-statistics/#top (accessed April 13, 2024).

32. Kumar M, Mani U, Tripathi P, Saalim M, Roy S. Artificial hallucinations by Google bard: Think before you leap. *Cureus.* (2023) 15:e43313.

33. Huang K, Mehta N, Gupta S, See A, Arnaout O. Evaluation of the safety, accuracy, and helpfulness of the GPT-4.0 large language model in neurosurgery. *J Clin Neurosci.* (2024) 123:151–6. doi: 10.1016/j.jocn.2024.03.021

34. Lim Z, Pushpanathan K, Yew S, Lai Y, Sun C, Lam J, et al. Benchmarking large language models' performances for myopia care: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine.* (2023) 95:104770. doi: 10.1016/j.ebiom.2023.104770

35. Meng J, Zhang Z, Tang H, Xiao Y, Liu P, Gao S, et al. Evaluation of ChatGPT in providing appropriate fracture prevention recommendations and medical science question responses: A quantitative research. *Medicine.* (2024) 103:e37458. doi: 10.1097/MD.0000000000037458