scientific reports

Check for updates

OPEN

Large language models provide discordant information compared to ophthalmology guidelines

Andrea Taloni[1,2,3,4], Antonia Carmen Sangregorio[4], Giuseppe Alessio[4], Maria Angela Romeo[4], Giulia Coco[5], Linda Marie Louise Busin[2,6], Andrea Sollazzo[1,2,3], Vincenzo Scorcia[4] & Giuseppe Giannaccare[7]✉

To evaluate the agreement of LLMs with the Preferred Practice Patterns® (PPP) guidelines developed by the American Academy of Ophthalmology (AAO). Open questions based on the AAO PPP were submitted to five LLMs: GPT-o1 and GPT-4o by OpenAI, Claude 3.5 Sonnet by Anthropic, Gemini 1.5 Pro by Google, and DeepSeek-R1-Lite-Preview. Questions were classified as "open" or "confirmatory with positive/negative ground-truth answer". Three blinded investigators classified responses as "concordant", "undetermined", or "discordant" compared to the AAO PPP. Undetermined and discordant answers were analyzed to assess harming potential for patients. Responses referencing peer-reviewed articles were reported. In total, 147 questions were submitted to the LLMs. Concordant answers were 135 (91.8%) for GPT-o1, 133 (90.5%) for GPT-4o, 136 (92.5%) for Claude 3.5 Sonnet, 124 (84.4%) for Gemini 1.5 Pro, and 119 (81.0%) for DeepSeek-R1-Lite-Preview ($P = 0.006$). The highest number of harmful answers was reported for Gemini 1.5 Pro ($n = 6$, 4.1%), followed by DeepSeek-R1-Lite-Preview ($n = 5$, 3.4%). Gemini 1.5 Pro was the most transparent model (86 references, 58.5%). Other LLMs referenced papers in 9.5–15.6% of their responses. LLMs can provide discordant answers compared to ophthalmology guidelines, potentially harming patients by delaying diagnosis or recommending suboptimal treatments.

Keywords Large language model, Artificial intelligence, Guidelines, Preferred practice patterns, American Academy of Ophthalmology, AAO

Large language models (LLMs) are dramatically changing the current scenarios of education and productivity, and the application of generative artificial intelligence (AI) in the medical field aroused significant interest in the scientific community. Several authors have investigated the proficiency of LLMs in answering ophthalmology questions, both from patients and examination boards, reporting encouraging results[1–8]. However, the reliability of LLM use in clinical practice is still a controversial topic. The main offenders are safety, transparency, and accountability concerns. Despite rapid technological advancements, LLMs can still provide inaccurate, wrong or non-sensical answers, often referred to as "hallucinations"[9]. In addition, chatbots do not always provide references for their outputs, because answers are based on mixed information from training data sets, which can include unreliable sources[10]. Even if references are specifically requested by the user, they can be inaccurate or completely made up[11,12]. Finally, medical practitioners have no control on the way patients interact with generative AI. In case of discrepancies between physician and LLM, doctor-patients relationship may become challenging.

Given the growing influence of LLMs on patient care and clinical practice, it is essential that these tools undergo rigorous, systematic evaluations to determine their alignment with established medical guidelines and good practice standards. The purpose of this study is to evaluate the agreement of publicly available LLMs with the Preferred Practice Patterns® (PPP) guidelines developed by the American Academy of Ophthalmology (AAO) and to identify cases in which AI outputs may be harmful to patients[13].

[1]Department of Translational Medicine, University of Ferrara, Ferrara, Italy. [2]Department of Ophthalmology, Ospedali Privati Forlì "Villa Igea", Forlì, Italy. [3]Istituto Internazionale di Ricerca e Formazione in Oftalmologia, Forlì, Italy. [4]Department of Ophthalmology, University Magna Graecia of Catanzaro, Catanzaro, Italy. [5]Department of Clinical Sciences and Translational Medicine, University of Rome Tor Vergata, Rome, Italy. [6]Department of Ophthalmology, Fatebenefratelli Hospital, Milan, Italy. [7]Department of Surgical Sciences, Eye Clinic, University of Cagliari, Via Università 40, 09124 Cagliari, Italy. ✉email: giuseppe.giannaccare@unica.it

Methods

In this study, five publicly available LLMs were selected to investigate their agreement with the AAO PPP: GPT-o1 and GPT-4o by OpenAI (San Francisco, CA, USA), Claude 3.5 Sonnet by Anthropic (San Francisco, CA, USA), Gemini 1.5 Pro by Google (Mountain View, CA, USA), and DeepSeek-R1-Lite-Preview (DeepSeek, Hangzhou, Zhejiang, China). Open-source LLMs which require local deployment were excluded from the study because they necessitate substantial computational resources, coding expertise, and technical customization, which would have limited the generalizability of our findings.

Open questions based on the "Highlighted findings and recommendations for care" summaries in each AAO PPP document were prepared by an investigator (A.T.) and reviewed by a Fellow of the European Board of Ophthalmology (G.G.). For each finding reported in the guidelines, one question was crafted. Questions were classified as "open" or "confirmatory with positive ground truth answer", and "confirmatory with negative ground-truth answer". Prompts were submitted to the LLMs in separate chat sessions, to prevent any potential influence of previous queries on subsequent ones. All questions were submitted between December 27th, 2024 and January 2nd, 2025.

Three blinded investigators (Giulia Coco, MD, PhD, researcher, and corneal surgeon consultant; Andrea Sollazzo, MD, researcher, and Andrea Taloni, MD and PhD student) collaboratively reviewed responses to classify them as "concordant", "undetermined", or "discordant" with respect to the AAO PPP documents. Any disagreement between investigators was resolved through discussion until a consensus was reached. Simple randomization was used to ensure variation in the order of exposure of answers by the five LLM models. Agreement was considered to be achieved if the chatbot's response exactly matched the AAO recommendations or provided conclusions that aligned with them. Answers were defined as undetermined if they failed to mention specific aspects of the findings reported in the AAO PPP, if they were too vague, or presented minor numerical discrepancies for quantitative measurements (e.g. slightly different percentages for the incidence of a certain disease). Finally, discordant responses presented opposite statements compared to the AAO guidelines or were severely lacking in addressing key aspects of the topic. An example of questions with concordant, undetermined, and discordant answers is reported in Table 1.

Furthermore, investigators collaboratively analyzed undetermined and discordant answers to assess whether their content had the potential to harm patients. This evaluation focused on identifying unclear, misleading or incorrect information that could lead to inappropriate clinical decisions, delayed diagnosis, or suboptimal treatment protocols. Examples of potentially harmful answers have been included in the Supplementary Materials.

Finally, instances in which LLM responses referenced peer-reviewed articles were recorded, while citations to non-peer-reviewed sources, as well as any tangential or hallucinated references, were excluded from the analysis.

Statistical analysis

The Chi-square test of independence or Fisher's exact test were performed to compare categorical variables as appropriate. The word count for answers by LLMs was calculated. The Shapiro–Wilk test was performed to determine the normality of data. Wilcoxon-Mann-Whitney U test for non-parametric variables was used to compare word count between LLMs. Values for continuous variables were expressed as mean ± standard deviation (95% confidence intervals [CI]). Bonferroni adjustment was applied to control for multiple comparisons. A P value < 0.01 (0.05/5) was considered statistically significant when comparing the outcomes of the 5 different LLMs, while a P value < 0.016 (0.05/3) was considered statistically significant when comparing outcomes between concordant, undetermined, and discordant answers. All tests were two-sided. All answers were recorded in Microsoft Excel (Microsoft Corp., Redmond, WA, USA). Statistical analysis was performed using GraphPad Prism (version 10.3.1; GraphPad Software, Boston, MA, USA).

Results

In total, 147 questions across the 25 AAO PPP documents were submitted to the selected LLMs. Table 2 includes results about answers agreement, harming potential, and inclusion of references. The number of answers classified as concordant with the AAO PPP was significantly different across the five LLMs (P = 0.006). GPT-o1, GPT-4o, and Claude 3.5 Sonnet showed very similar performance, respectively scoring 135 (91.8%), 133 (90.5%), and 136 (92.5%) concordant answers (P = 0.81); Gemini 1.5 Pro scored 124 concordant answers (84.4%), while the worst performing model was DeepSeek-R1-Lite-Preview, with 119 concordant answers (81.0%). GPT-o1 was the LLM with the least number of discordant answers (n = 3; 2.0%), while the highest amount was recorded for DeepSeek-R1-Lite-Preview (n = 10; 6.8%). Figure 1 shows the number of concordant, undetermined, and discordant answers for each LLM.

Some of the undetermined and discordant answers were deemed to be potentially harmful to patients. The highest number of harmful answers was reported for Gemini 1.5 Pro, with 6 instances (4.1%), followed by DeepSeek-R1-Lite-Preview, with 5 instances (3.4%). The other models were safer, with 3 harmful answers (2.0%) for Claude 3.5 Sonnet, and 2 (1.4%) from both GPT-o1 and GPT-4o (Fig. 2). Concerning references, Gemini 1.5 Pro was the most transparent model, with 86 answers citing peer-reviewed articles (58.5%). Surprisingly, including references did not improve agreement with the AAO PPP guidelines (84.88% concordant answers with references vs. 83.61% concordant answers without references; P = 0.82). Other models provided significantly fewer references (P < 0.001). GPT-o1 cited studies in 23 answers (15.6%), DeepSeek-R1-Lite-Preview in 15 (10.2%), while both GPT-4o and Claude 3.5 Sonnet in 14 (9.5%) (Fig. 3). Similarly to Gemini 1.5 Pro, references did not correlate with significantly higher numbers of concordant responses, in any LLM.

Overall, 112 questions required an open answer, while the remaining 35 were confirmatory by design (14 with an expected positive answer and 21 with a negative one). Three LLMs had better performance in confirmatory

Question	Is aspirin associated with an increased risk of age-related macular degeneration?	
AAO PPP	A meta-analysis of 10 studies found that the use of aspirin was not associated with an increased risk of AMD. Therefore, patients who have been instructed by a physician to use aspirin should continue to use it as prescribed	Adherence to AAO PPP
GPT o1	No definitive causal link has been established between aspirin use and an increased risk of age-related macular degeneration (AMD). While some observational studies have raised questions about a possible association, the overall body of evidence and large, controlled studies have not confirmed that aspirin use leads to a higher incidence of AMD or its progression. *Key Points*: *1. Conflicting Observational Findings*: A few observational studies suggested a slight association between aspirin use and AMD. However, these findings have not been consistently replicated and may be influenced by factors like age, general health status, or concurrent conditions for which aspirin is prescribed. *2. Lack of Conclusive Evidence*: Larger studies and meta-analyses have not found a definitive, dose-dependent, or temporal relationship that would indicate aspirin use directly contributes to AMD. The discrepancies across various studies highlight the complexity of determining a true causal relationship. *3. Clinical Recommendations*: Aspirin is often used for cardiovascular protection. Given the lack of strong evidence linking aspirin to increased AMD risk, most eye care professionals do not recommend stopping aspirin therapy for fear of AMD progression. Decisions about aspirin use should continue to be based on cardiovascular risk assessments rather than concerns about AMD. *Conclusion*: Current evidence does not support changing aspirin use solely on the basis of potential AMD risk. Patients should follow medical advice for aspirin therapy based on their individual health profiles rather than concerns about AMD.	C
GPT4o	Yes, some studies have suggested that regular aspirin use may be associated with an increased risk of age-related macular degeneration (AMD), particularly the "wet" form of the condition. AMD is a leading cause of vision loss in older adults, and it comes in two forms: *dry AMD* and *wet AMD*. The "wet" form is more severe and involves the growth of abnormal blood vessels in the retina. Here are the key points from research on the relationship between aspirin use and AMD: *1. Wet AMD Risk*: A few studies, including a well-known study from the European Eye Study and the Beaver Dam Eye Study, found that regular aspirin users, especially those using it long-term, were at a higher risk of developing wet AMD compared to non-users. For example, one study found that individuals who took aspirin daily for over 10 years had a slightly higher incidence of wet AMD. *2. Dry AMD Risk*: The evidence connecting aspirin use to dry AMD is less clear. Some studies haven't found a significant association between aspirin and an increased risk of the dry form. *3. Conflicting Evidence*: Despite these findings, other research has shown no strong connection between aspirin use and AMD. This has led to some debate in the medical community about the exact nature of this potential risk. *4. Potential Mechanisms*: It's not entirely clear how aspirin might increase the risk of wet AMD. Aspirin is an anti-inflammatory and blood-thinning medication, and some researchers have speculated that its effect on blood vessels could contribute to the development of abnormal vessels in the retina. *5. Weighing the Risks*: It's important to note that aspirin is often prescribed to reduce the risk of heart attack and stroke. For people with cardiovascular conditions, the benefits of aspirin may outweigh the risks related to AMD. Given the mixed results from different studies, it's crucial for individuals, especially those with a higher risk of AMD, to consult with their doctors before making any changes to aspirin use based on concerns about eye health.	U
Claude 3.5-Sonnet	Based on several epidemiological studies, there is evidence suggesting that regular aspirin use may be associated with an increased risk of age-related macular degeneration (AMD), particularly the neovascular (wet) form. The strongest evidence comes from: 1. The Beaver Dam Eye Study found that regular aspirin use (at least twice weekly for more than 3 months) was associated with increased risk of developing neovascular AMD over a 10-year period. 2. The European Eye Study showed a nearly doubled risk of neovascular AMD among regular aspirin users. 3. The Blue Mountains Eye Study demonstrated an independent association between regular aspirin use and incident neovascular AMD over a 15-year period. However, it's important to note that: - The absolute risk increase appears to be relatively small - The benefits of aspirin for cardiovascular prevention may outweigh the potential increased AMD risk for many patients - The association doesn't definitively prove causation - Patients shouldn't stop prescribed aspirin without consulting their healthcare provider For patients with cardiovascular conditions requiring aspirin therapy, the established benefits generally outweigh the potential small increase in AMD risk. Individual risk-benefit discussions with healthcare providers are recommended. Since you're asking about research findings, I should mention that while I aim to accurately represent the scientific literature, you should verify these specific studies with current medical sources, as I may not have perfect recall of the exact details.	D
Continued		

Question	Is aspirin associated with an increased risk of age-related macular degeneration?	
Gemini 1.5 Pro	The association between aspirin and age-related macular degeneration (AMD) has been a topic of research and some debate, but the current consensus leans towards *no significant increase in risk*. Here's a breakdown of the evidence: *Early Studies*: Some initial observational studies suggested a potential link between aspirin use and AMD, particularly wet AMD. However, these studies had limitations, as they couldn't fully account for other factors that might influence both aspirin use and AMD development. *Large-Scale Clinical Trials*: More recent and robust studies, including the Age-Related Eye Disease Studies (AREDS and AREDS2), have provided stronger evidence. These large-scale trials found *no significant association* between aspirin use and progression to late AMD or its subtypes. *Randomized Controlled Trials*: These trials, considered the gold standard in research, have also indicated that aspirin does not increase the risk of developing AMD. *Important Considerations*: *Medical Advice*: It's crucial to consult your doctor before making any decisions about aspirin use, especially if you have AMD or are at risk for it. They can assess your individual situation and advise on the best course of action. *Benefits of Aspirin*: Aspirin has proven benefits in preventing cardiovascular disease, which is a major cause of morbidity and mortality. *Risk Factors for AMD*: While aspirin may not be a significant risk factor, other factors like age, smoking, genetics, and diet play a more substantial role in AMD development. *In conclusion*: The current evidence suggests that aspirin use does not significantly increase the risk of age-related macular degeneration. If your doctor recommends aspirin for other health reasons, you generally don't need to avoid it due to concerns about AMD. However, always discuss your individual situation with your healthcare provider. Here are some resources for further reading: *AREDS: Aspirin Safe for Patients With AMD - American Academy of Ophthalmology*: [https://www.aao.org/eyenet/article/areds-aspirin-safe-for-patients-with-amd*Does* *Aspirin Increase the Risk of Permanent Vision Loss from Macular Degeneration?*: [https://www.brightfocus.org/macular/article/aspirin-and-macular*Does* *aspirin increase risk of AMD? - Macular Society*: [https://www.macularsociety.org/research/features/aspirin-risk/	C
DeepSeek R1-Lite-Preview	The question of whether aspirin is associated with an increased risk of age-related macular degeneration (AMD) is complex and requires a nuanced answer based on current research. Here is a summary of the key points: *1. Aspirin's Role and Uses*: Aspirin is a nonsteroidal anti-inflammatory drug (NSAID) commonly used for pain relief, fever reduction, and as an antiplatelet agent to prevent heart attacks and strokes. *2. AMD Overview*: AMD is a degenerative eye disease affecting the macula, leading to central vision loss. It has two forms: dry AMD and the more severe wet AMD, characterized by abnormal blood vessel growth. *3. Potential Association*: Some studies suggest a possible link between regular aspirin use and an increased risk of wet AMD. However, this association is not definitively established, and further research is needed to confirm causation. *4. Mechanisms and Considerations*: The exact mechanisms by which aspirin might influence AMD development are unclear. Aspirin's effects on blood clotting and inflammation could potentially impact eye health, but this is speculative. *5. Risk Factors and Confounders*: Other risk factors for AMD, such as age, smoking, genetics, and diet, may confound the relationship between aspirin use and AMD risk. *6. Dosage and Frequency*: The impact of aspirin on AMD risk may depend on dosage and frequency of use, though this is not well-defined. *7. Risk-Benefit Balance*: Aspirin has established benefits for heart health, so any potential increased risk of AMD must be weighed against these benefits on an individual basis. *8. Consistency of Evidence*: The evidence on aspirin and AMD risk is not entirely consistent across studies, indicating that the relationship may not be straightforward. *9. Guidelines and Recommendations*: Medical organizations such as the American Academy of Ophthalmology or the American Heart Association may provide guidance on aspirin use in the context of AMD risk, though specific recommendations may not be available. In conclusion, while some studies suggest a potential association between aspirin use and increased risk of AMD, the evidence is not conclusive. Individuals should consult with their healthcare providers to discuss their specific risks and benefits before making decisions about aspirin use.	U

Table 1. Example of a question with concordant ("C"), undetermined ("U"), and discordant answers ("D") compared to the American Academy of Ophthalmology Preferred Practice Patterns (AAO PPP).

Answers		GPT-o1	GPT-4o	Claude 3.5 Sonnet	Gemini 1.5 Pro	DeepSeek R1-Lite-Preview	*P* value
Concordant	Total	135 (91.8%)	133 (90.5%)	136 (92.5%)	124 (84.4%)	119 (81%)	0.006*
	R+	23 (17%)	13 (9.8%)	10 (7.4%)	73 (58.9%)	13 (10.9%)	< 0.001*
Undetermined	Total	9 (6.1%)	8 (5.4%)	4 (2.7%)	16 (10.9%)	18 (12.2%)	0.096
	R+	0 (0%)	1 (0.8%)	1 (0.7%)	10 (8.1%)	2 (1.7%)	0.002*
	H+	2 (22.2%)	0 (0%)	1 (25.0%)	3 (18.8%)	0 (0%)	0.102
Discordant	Total	3 (2%)	6 (4.1%)	7 (4.8%)	7 (4.8%)	10 (6.8%)	0.406
	R+	0 (0%)	0 (0%)	3 (42.9%)	3 (42.9%)	0 (0%)	0.028
	H+	1 (33.3%)	2 (33.3%)	2 (28.6%)	3 (42.9%)	5 (50%)	0.932
All	R+	23 (15.6%)	14 (9.5%)	14 (9.5%)	86 (58.5%)	15 (10.2%)	< 0.001*
	H+	3 (2.0%)	2 (1.4%)	3 (2.0%)	6 (4.1%)	5 (3.4%)	0.471

Table 2. Number of concordant, undetermined, and discordant answers compared to the American academy of ophthalmology preferred practice patterns. Each category includes the number of references ("R+") and potentially harmful answers ("H+"). The *P* values indicate the statistical significance of differences across all five llms, calculated using the Chi-square test of independence or fisher's exact test, as appropriate. * A *P* value < 0.016 was considered statistically significant.

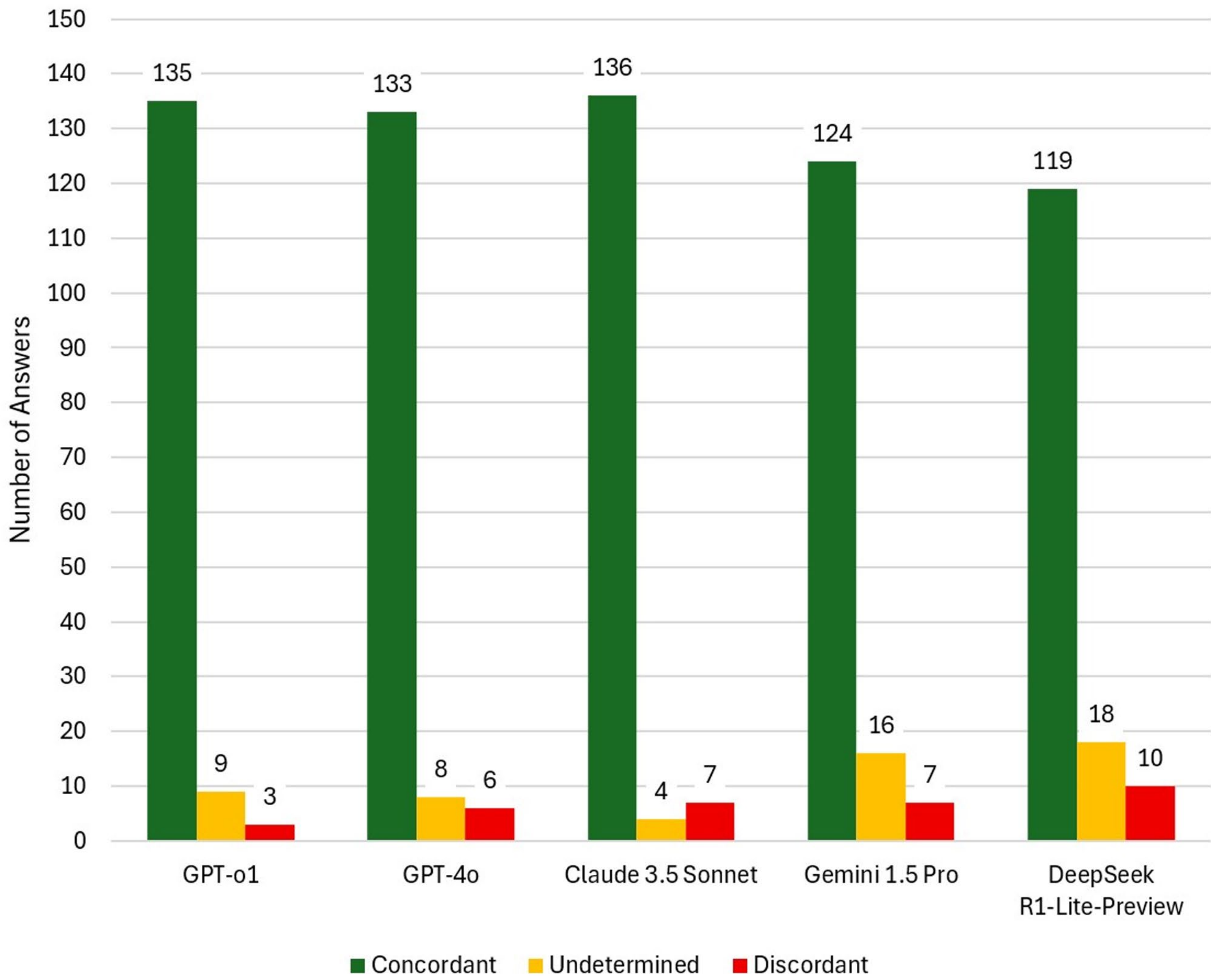

Fig. 1. Number of concordant (green), undetermined (yellow), and discordant answers (red) compared to the American Academy of Ophthalmology Preferred Practice Patterns®.

questions (GPT-o1, 90.2% vs. 97.1%, $P = 0.30$; Gemini 1.5 Pro, 81.3% vs. 94.3%, $P = 0.11$; DeepSeek-R1-Lite-Preview, 79.5% vs. 85.7%, $P = 0.47$), while two showed worse results (GPT-4o, 92.0% vs. 85.7%, $P = 0.32$; Claude 3.5 Sonnet, 92.9% vs. 91.4%, $P = 0.72$); however, the differences were not statistically significant. Interestingly, in case of confirmatory questions with positive ground truth, all five LLMs were always concordant with the AAO PPP, while performance worsened in confirmatory questions with negative ground truth (Table 3).

The word count of each answer was significantly different across LLMs ($P < 0.001$). The most verbose chatbot was GPT-o1 (321.7 ± 164.2 words [95% CI, 295.0-348.5]), closely followed by GPT-4o (296.4 ± 97.4 words [95% CI, 280.6-312.3]) and Gemini 1.5 Pro (291.7 ± 87.71 [95% CI, 277.4–306.0]). DeepSeek-R1-Lite-Preview presented a lower count of words (245.1 ± 116.5 [95% CI, 226.2-264.1]), while the most concise model was Claude 3.5 Sonnet (180.9 ± 45.01 [95% CI, 173.6-188.2]). GPT-4o was the only model to present a significant difference in word count between concordant and non-concordant answers (respectively, 286.7 ± 91.5 words vs. 405.9 ± 98.7, $P < 0.001$).

Discussion

Generative AI has the potential to support healthcare systems by providing quick access to information and assisting clinicians in both administrative and clinical-related tasks[14–17]. LLMs will also increasingly influence medical education, answering patients' queries and enhancing their understanding of ocular diseases and available treatment options[18,19]. However, there are still several barriers to the widespread adoption of this technology in daily clinical practice and telemedicine. Key priorities include eliminating or greatly reducing hallucinations, while improving the overall accuracy of the models and their adherence to official medical guidelines[20].

In this study five publicly available LLMs have been prompted to answer 147 questions based on the AAO PPP documents. Answers were classified as concordant, undetermined, or discordant compared to the "Highlighted findings and recommendations for care" of the AAO PPP. The best performing models were GPT-o1, GPT-4o,

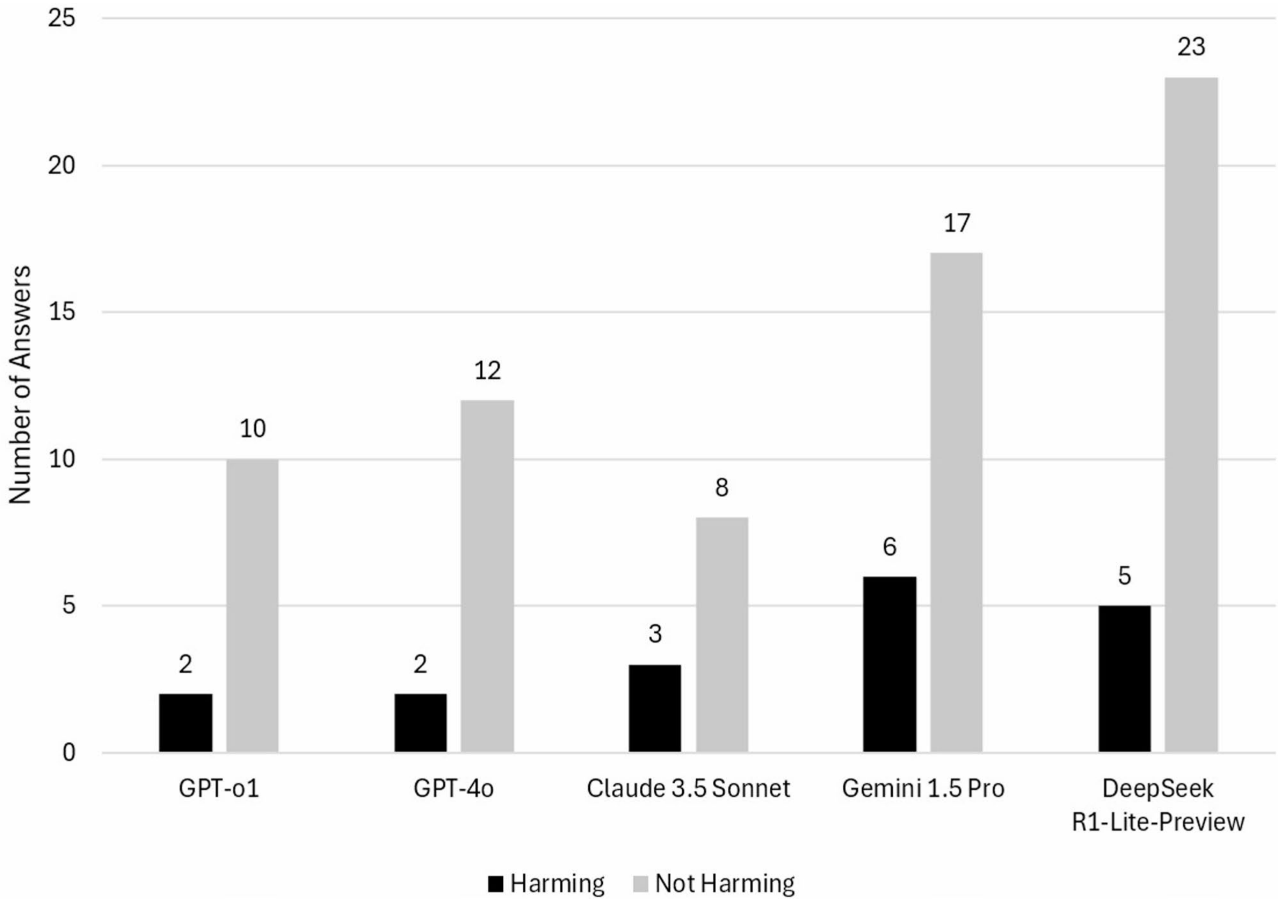

Fig. 2. Number of answers potentially harmful to the patients.

and Claude 3.5 Sonnet, with approximately 90% concordant answers. This might be considered a promising result; however, it is still arguably insufficient to consistently rely on LLMs for clinical decision-making. Gemini 1.5 Pro and DeepSeek-R1-Lite-Preview obtained even lower percentages, respectively 84.4% and 81.6%. While achieving higher agreement with guidelines is crucial, a greater concern is the impact of discordant, incomplete, or inaccurate answers, which may directly or indirectly harm patients. The percentage of potentially harmful answers across all LLMs ranged from 1.4 to 4.1%. LLMs yielded potential for harm due to omitted clinical signs and/or inadequate treatment recommendations.

Notably, confirmatory questions with positive ground truth were always answered correctly by all LLMs, while performance was worse for confirmatory questions with negative ground truth. This may suggest a possible confirmation bias of LLMs; however, our sample size of confirmatory questions was limited, and the study might not have been adequately powered to detect significant differences.

Another key aspect in evaluating answers from LLMs is the inclusion of references to peer-reviewed articles. All queried LLMs were severely lacking in this regard. Gemini 1.5 Pro provided the highest percentage of answers with references (58.5%). Citations were even less common in other LLMs, appearing in approximately 10–15% of the responses. Furthermore, the inclusion of references did not correlate with higher agreement rate with the AAO PPP. Arguably, current LLMs may be still limited by their core principles of function. LLMs primarily operate by predicting the most likely next word in a sentence based on the submitted prompt and prior context. They are trained on a huge quantity of texts to recognize and reproduce natural language; however, given the sheer volume of data, it is not feasible to perform a comprehensive quality check on the entire training dataset, which may end up including inaccurate or inappropriate information. In addition, GPT-4o and Gemini 1.5 Pro can also browse the internet to gather more data. Currently, it is not clearly documented how LLMs deal with conflicting information from their training dataset or from websites they browse, and what are the specific criteria used to prioritize one source over another.

Analyzing the word count of LLM responses may offer insights into how models manage uncertainty or knowledge gaps. Short answers may be incomplete or unclear. On the other hand, models may attempt to compensate for uncertainty or lack of accurate information through excessive verbosity. While longer responses may provide detailed explanations, they may also introduce ambiguity or irrelevant data. In our study, Claude 3.5 Sonnet was the most concise model, whereas models by OpenAI generated the longest answers. Notably, GPT-4o generated significantly longer answers in case of non-concordant responses. Other LLMs did not display significant differences in word count between concordant and non-concordant answers.

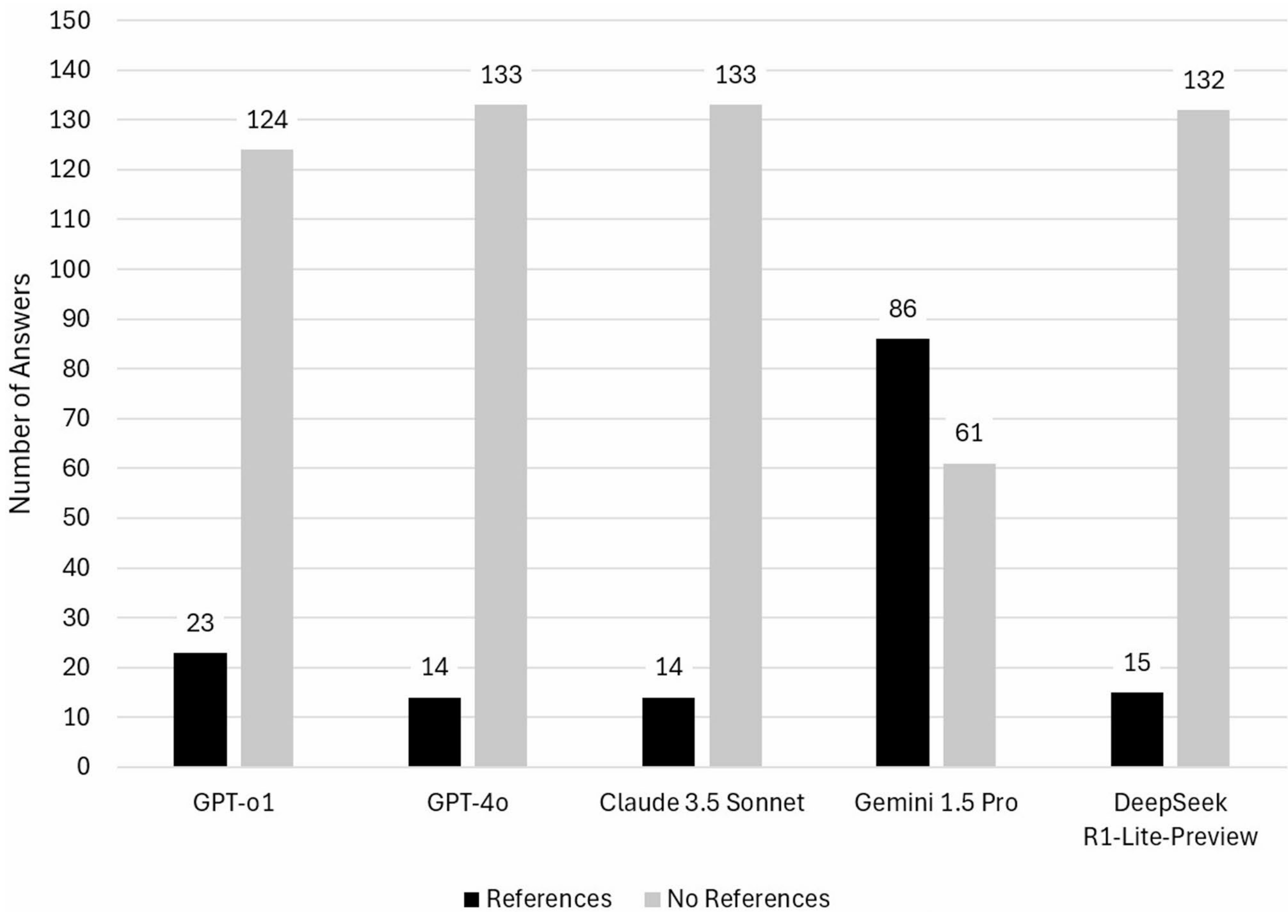

Fig. 3. Number of answers containing references to peer-reviewed articles.

Answers		GPT-o1	GPT-4o	Claude 3.5 Sonnet	Gemini 1.5 Pro	DeepSeek R1-Lite-Preview
Open	C	101 (90.2%)	103 (92%)	104 (92.9%)	91 (81.3%)	89 (79.5%)
	U	9 (8%)	6 (5.4%)	4 (3.6%)	16 (14.3%)	14 (12.5%)
	D	2 (1.8%)	3 (2.7%)	4 (3.6%)	5 (4.5%)	9 (8%)
Confirmatory (All)	C	34 (97.1%)	30 (85.7%)	32 (91.4%)	33 (94.3%)	30 (85.7%)
	U	0 (0%)	2 (5.7%)	0 (0%)	0 (0%)	4 (11.4%)
	D	1 (2.9%)	3 (8.6%)	3 (8.6%)	2 (5.7%)	1 (2.9%)
P value		0.30	0.32	0.72	0.11	0.47
Confirmatory (+)	C	14 (100%)	14 (100%)	14 (100%)	14 (100%)	14 (100%)
	U	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	D	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Confirmatory (-)	C	20 (95.2%)	16 (76.2%)	18 (85.7%)	19 (90.5%)	16 (76.2%)
	U	0 (0%)	2 (9.5%)	0 (0%)	0 (0%)	4 (19%)
	D	1 (4.8%)	3 (14.3%)	3 (14.3%)	2 (9.5%)	1 (4.8%)
P value		1.00	0.07	0.26	0.51	0.07

Table 3. Number of concordant, undetermined, and discordant answers compared to the American Academy of Pphthalmology Preferred Practice Patterns. Results are presented separately for "open" and "confirmatory" questions (positive and negative ground truth). On top, *P* values indicate the statistical significance of the difference between concordant answers for open and confirmatory questions, in each of the five llms. On bottom, *P* values indicate the statistical significance of the difference between concordant answers for confirmatory questions with positive ground truth (+) and negative ground truth (-), in each of the five llms. All P values were calculated using the Chi-square test of independence or fisher's exact test, as appropriate. * A *P* value < 0.01 was considered statistically significant.

A new feature of some LLMs is the enhancement of logical reasoning thanks to reinforcement learning. GPT-o1 generates a chain of thought before each answer, optimizing performance in complex reasoning tasks[21]. Similarly, DeepSeek provides a dedicated tool called "DeepThink" (R1 model) which is based on reinforcement learning[22]. Despite these advancements, our results did not reveal significant differences in the performance of GPT-o1, GPT-4o, and Claude 3.5 Sonnet. In addition, DeepSeek was the worst performer in our study. Publishers' benchmarks highlight the capabilities of GPT-o1 and DeepSeek in logical reasoning and math-related problems[23,24]; however, the medical questions submitted in our study likely did not offer opportunities for these strengths to emerge.

A possible solution to improve the reliability of generative AI is the development of specialized LLMs tailored towards medical practice. Luo et al. developed ChatZOC, a retrieval-augmented LLM framework for ophthalmology based on the Chinese model Baichuan-13B[25]. This LLM was fine-tuned on a comprehensive ophthalmic dataset including over 30,000 pieces of ophthalmic knowledge[25]. ChatZOC outperformed the baseline Baichuan-13B model in answering ophthalmology questions; however, results were not superior to the ChatGPT-4 model. Zheng et al. developed another LLM called MOPH by adopting the open-source LLM ChatGLM2-6B. Prompt engineering was used to enhance MOPH's semantic understanding by creating domain-specific prompts, leveraging Chinese ophthalmic databases, peer-reviewed articles, and AAO EyeWiki. MOPH demonstrated an adherence of 83.3% in respect to Chinese guidelines[26]. Training and running such models requires information technology expertise, as well as high computational resources and economical investments. On the other hand, local deployment of LLMs minimizes potential privacy concerns, and allows better control over the model.

An alternative and simpler approach may be the creation of "custom GPTs". These are versions of the default GPT-4o model which include user-defined "Instructions" and extra "Knowledge" consisting of uploaded documents that the LLM can use as reference to answer the user's prompts. Similar custom models called "Gems" can be designed in Gemini. In this case the LLM retrieves data from documents uploaded in Google Drive.

This study presents several limitations that deserve mentioning. Firstly, we did not include a control group of experienced ophthalmologists to answer the questions submitted to the LLMs. In addition, the classification of answers as concordant, undetermined, discordant, and harmful, although collaboratively performed by multiple investigators, is still partially subjective. Moreover, while we compared agreement between answers to "open" and "confirmatory" questions, we did not systematically formulate each question in both formats. This approach could have provided valuable insight into the potential impact of different prompting strategies. Finally, a key limitation of studies in the field of generative AI is that LLM performance may change rapidly due to frequent model updates and changes in training data cutoffs.

To our knowledge this is the first study to assess the agreement of LLMs with ophthalmology guidelines (PubMed, search query: "ophthalmology" AND ("agreement" OR "concordance" OR "adherence") AND ("guidelines" or "practice patterns") AND "large language model", 2022–2025). However, other authors have already investigated LLMs' adherence to guidelines in other medical fields. Nwachukwu et al. have evaluated whether commercially available LLMs (ChatGPT-4, Gemini, Mistral-7B, and Claude-3 Sonnet) provide treatment recommendations concordant with the clinical guidelines published by the American Academy of Orthopaedic Surgeons for the management of rotator cuff tears and anterior cruciate ligament injuries. In their study, agreement with guidelines was most frequently observed with ChatGPT-4 (79.2%) and least frequently with Mistral7B (58.3%). Conversely, discordant recommendations were most frequently recorded with Gemini (12.5%) and least frequently with ChatGPT-4 (2.1%)[27]. More recently, Fast et al. developed a benchmark called "Autonomous Medical Evaluation for Guideline Adherence" designed to evaluate the adherence of 17 LLMs to medical guidelines across 20 diagnostic scenarios in 13 specialties. The benchmark included 135 questions and more than one thousand weighted scoring elements. ChatGPT-4 scored highest (41.9/50), followed closely by Llama-3 70B and WizardLM-2-8 × 22B [28].

Assessing and optimizing the reliability of LLMs in medicine should be a priority to expand the applicability of generative AI. Physicians and developers should combine efforts to create domain-specific LLMs, tested and approved by international medical associations; however, significant challenges remain for their full-fledged implementation in clinical practice. Medical questions used for testing purposes often provide a clear and unequivocal clinical scenario. In contrast, real-life diagnostic and therapeutic workflows often rely on the identification of subtle signs or specifics symptoms which can be ambiguously described by the patients. The complexity of such clinical scenarios cannot be fully captured in an online conversation with a chatbot. LLMs may serve best as supportive tools rather than standalone solutions in medical practice.

Data availability

Data are available upon reasonable request by contacting Andrea Taloni (taloni.ophthalmology@gmail.com).

Received: 9 February 2025; Accepted: 9 June 2025
Published online: 01 July 2025

References

1. Taloni, A. et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. *Sci. Rep.* **13**, 18562. https://doi.org/10.1038/s41598-023-45837-2 (2023).
2. Haddad, F. & Saade, J. S. Performance of ChatGPT on ophthalmology-related questions across various examination levels: Observational study. *JMIR Med. Educ.* **10**, e50842. https://doi.org/10.2196/50842 (2024).
3. Sakai, D. et al. Performance of ChatGPT in board examinations for specialists in the Japanese Ophthalmology Society. *Cureus* **15**, e49903. https://doi.org/10.7759/cureus.49903 (2023).

4. Gill, G. S. et al. Comparison of Gemini advanced and ChatGPT 4.0's performances on the ophthalmology resident ophthalmic knowledge assessment program (OKAP) examination review question banks. *Cureus* **16**, e69612. https://doi.org/10.7759/cureus.69612 (2024).
5. Gill, G. S., Blair, J. & Litinsky, S. Evaluating the performance of ChatGPT 3.5 and 4.0 on StatPearls oculoplastic surgery text- and image-based exam questions. *Cureus* **16**, e73812. https://doi.org/10.7759/cureus.73812 (2024).
6. Cohen, S. A. et al. 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.* **39**, 472–479. https://doi.org/10.1080/08820538.2024.2326058 (2024).
7. Bernstein, I. A. et al. Comparison of ophthalmologist and large language model Chatbot responses to online patient eye care questions. *JAMA Netw. Open* **6**, e2330320. https://doi.org/10.1001/jamanetworkopen.2023.30320 (2023).
8. Carlà, M. M., Giannuzzi, F., Boselli, F. & Rizzo, S. Testing the power of Google DeepMind: Gemini versus ChatGPT 4 facing a European ophthalmology examination. *AJO Int.* **1**, 100063. https://doi.org/10.1016/j.ajoint.2024.100063 (2024).
9. Alkaissi, H. & McFarlane, S. I. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* https://doi.org/10.7759/CUREUS.35179 (2023).
10. Chen, A. & Chen, D. O. Accuracy of Chatbots in citing journal articles. *JAMA Netw. Open* **6**, e2327647. https://doi.org/10.1001/jamanetworkopen.2023.27647 (2023).
11. Mugaanyi, J. et al. Evaluation of large language model performance and reliability for citations and references in scholarly writing: Cross-disciplinary study. *J. Med. Internet Res.* **26**, e52935. https://doi.org/10.2196/52935 (2024).
12. McGowan, A. et al. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Res.* **326**, 115334. https://doi.org/10.1016/j.psychres.2023.115334 (2023).
13. About Preferred Practice Patterns (PPPs). American Academy of Ophthalmology.
14. Singh, S., Djalilian, A. & Ali, M. J. ChatGPT and ophthalmology: Exploring its potential with discharge summaries and operative notes. *Semin. Ophthalmol.* **38**, 503–507. https://doi.org/10.1080/08820538.2023.2209166 (2023).
15. GPT-4 and Ophthalmology Operative Notes | Annals of Biomedical Engineering. https://doi.org/10.1007/s10439-023-03263-5. Accessed 13 Jan 2025.
16. Chotcomwongse, P., Ruamviboonsuk, P. & Grzybowski, A. Utilizing large language models in ophthalmology: The current landscape and challenges. *Ophthalmol. Ther.* **13**, 2543–2558. https://doi.org/10.1007/s40123-024-01018-6 (2024).
17. Huang, A. S. et al. Assessment of a large language model's responses to questions and cases about glaucoma and retina management. *JAMA Ophthalmol.* **142**, 371–375. https://doi.org/10.1001/jamaophthalmol.2023.6917 (2024).
18. Özer Özcan, Z., Doğan, L. & Yilmaz, I. E. Artificial doctors: Performance of Chatbots as a tool for patient education on keratoconus. *Eye Contact Lens* **51**, e112. https://doi.org/10.1097/ICL.0000000000001160 (2025).
19. Doğan, L., Özer Özcan, Z. & Edhem Yılmaz, I. The promising role of chatbots in keratorefractive surgery patient education. *J. Fr. Ophtalmol.* **48**, 104381. https://doi.org/10.1016/j.jfo.2024.104381 (2025).
20. Liu, Y., Yao, Y., Ton, J.-F. et al. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment (2024).
21. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/. Accessed 10 Jan 2025.
22. DeepSeek-AI, Liu, A., Feng, B. et al. DeepSeek-V3 Technical Report (2024).
23. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/. Accessed 14 Jan 2025.
24. Introducing DeepSeek-V3 | DeepSeek API Docs. https://api-docs.deepseek.com/news/news1226. Accessed 14 Jan 2025.
25. Luo, M.-J. et al. Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. *JAMA Ophthalmol.* **142**, 798–805. https://doi.org/10.1001/jamaophthalmol.2024.2513 (2024).
26. Zheng, C. et al. Development and evaluation of a large language model of ophthalmology in Chinese. *Br. J. Ophthalmol.* **108**, e324526. https://doi.org/10.1136/bjo-2023-324526 (2024).
27. Nwachukwu, B. U. et al. Currently available large language models do not provide musculoskeletal treatment recommendations that are concordant with evidence-based clinical practice guidelines. *Arthroscopy* https://doi.org/10.1016/j.arthro.2024.07.040 (2024).
28. Fast, D. et al. Autonomous medical evaluation for guideline adherence of large language models. *NPJ Digit. Med.* **7**, 358. https://doi.org/10.1038/s41746-024-01356-6 (2024).

Acknowledgements

The American Academy of Ophthalmology Preferred Practice Pattern® is a registered trademark of the American Academy of Ophthalmology (AAO). This research was neither funded nor promoted by the AAO, but was conceived independently by the authors.

Author contributions

Conceptualization, A.T., G.C., V.S., and G.G.; Methodology, A.T., AC.S., G.A., G.C., LML.B., MA.R., and G.G.; Validation, A.T., G.A., MA.R., G.C., LML.B., A.S., V.S., and G.G.; Formal Analysis, A.T., AC.S., G.C., LML.B., A.S., MA.R., and G.G.; Investigation, A.T., AC.S., G.A., LML.B., A.S., and MA.R.; Data Curation, A.T., AC.S., G.A., LML.B., A.S., and MA.R.; Writing – Original Draft Preparation, A.T., AC.S., G.C., and MA.R.; Writing – Review & Editing, A.S., G.C., V.S., and G.G.; Visualization, A.T., AC.S., G.A., MA.R., G.C., LML.B., A.S., V.S., and G.G.; Supervision, A.T., V.S., and G.G.; Project Administration, A.T., V.S., and G.G. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

The research did not involve humans or animals.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-06404-z.

Correspondence and requests for materials should be addressed to G.G.