Published in final edited form as:

Mayo Clin Proc Digit Health. 2024 March ; 2(1): 119–128. doi:10.1016/j.mcpdig.2024.01.003.

Appropriateness of Ophthalmology Recommendations From an Online Chat-Based Artificial Intelligence Model

Prashant D. Tailor, MD,

Timothy T. Xu, MD,

Blake H. Fortes, MD,

Raymond Iezzi, MD,

Timothy W. Olsen, MD,

Matthew R. Starr, MD,

Sophie J. Bakri, MD,

Brittni A. Scruggs, MD, PhD,

Andrew J. Barkmeier, MD,

Sanjay V. Patel, MD,

Keith H. Baratz, MD,

Ashlie A. Bernhisel, MD,

Lilly H. Wagner, MD,

Andrea A. Tooley, MD,

Gavin W. Roddy, MD, PhD,

Arthur J. Sit, MD,

Kristi Y. Wu, MD,

Erick D. Bothun, MD,

Sasha A. Mansukhani, MBBS,

Brian G. Mohney, MD,

John J. Chen, MD, PhD,

Michael C. Brodsky, MD,

Deena A. Tajfirouz, MD,

Kevin D. Chodnicki, MD,

Wendy M. Smith, MD,

Lauren A. Dalvin, MD

Correspondence: Address to Lauren A Dalvin, MD, 200 First Street SW, Rochester, MN 55905 (Dalvin.lauren@mayo.edu; Twitter: @LADalvinMD).

POTENTIAL COMPETING INTERESTS

The authors report no competing interests.

SUPPLEMENTAL MATERIAL ONLINE

Supplemental material can be found online at https://www.mcpdigitalhealth.org/. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Department of Ophthalmology (P.D.T., T.T.X., B.H.F., R.I., T.W.O., M.R.S., S.J.B., B.A.S., A.J.B., S.J.B., B.A.S., A.J.B., S.V.P., K.H.B., A.A.B., L.H.W., A.A.T., G.W.R., A.J.S., K.Y.W., E.D.B., S.A.M., B.G.M., J.J.C., M.C.B., D.A.T., K.D.C., W.M.S., L.A.D.) and Department of Neurology (J.J.C.), Mayo Clinic, Rochester, MN; and Department of Ophthalmology, Duke University, Durham, NC (K.Y.W.).

Abstract

Objective: To determine the appropriateness of ophthalmology recommendations from an online chat-based artificial intelligence model to ophthalmology questions.

Patients and Methods: Cross-sectional qualitative study from April 1, 2023, to April 30, 2023. A total of 192 questions were generated spanning all ophthalmic subspecialties. Each question was posed to a large language model (LLM) 3 times. The responses were graded by appropriate subspecialists as appropriate, inappropriate, or unreliable in 2 grading contexts. The first grading context was if the information was presented on a patient information site. The second was an LLM-generated draft response to patient queries sent by the electronic medical record (EMR). Appropriate was defined as accurate and specific enough to serve as a surrogate for physician-approved information. Main outcome measure was percentage of appropriate responses per subspecialty.

Results: For patient information site-related questions, the LLM provided an overall average of 79% appropriate responses. Variable rates of average appropriateness were observed across ophthalmic subspecialties for patient information site information ranging from 56% to 100%: cataract or refractive (92%), cornea (56%), glaucoma (72%), neuro-ophthalmology (67%), oculoplastic or orbital surgery (80%), ocular oncology (100%), pediatrics (89%), vitreoretinal diseases (86%), and uveitis (65%). For draft responses to patient questions via EMR, the LLM provided an overall average of 74% appropriate responses and varied by subspecialty: cataract or refractive (85%), cornea (54%), glaucoma (77%), neuro-ophthalmology (63%), oculoplastic or orbital surgery (62%), ocular oncology (90%), pediatrics (94%), vitreoretinal diseases (88%), and uveitis (55%). Stratifying grades across health information categories (disease and condition, risk and prevention, surgery-related, and treatment and management) showed notable but insignificant variations, with disease and condition often rated highest (72% and 69%) for appropriateness and surgery-related (55% and 51%) lowest, in both contexts.

Conclusion: This LLM reported mostly appropriate responses across multiple ophthalmology subspecialties in the context of both patient information sites and EMR-related responses to patient questions. Current LLM offerings require optimization and improvement before widespread clinical use.

Chat generative pretrained transformer (ChatGPT) (OpenAI) is an online artificial intelligence (AI)-driven natural language processing model released in November 2022, which represents a generational advancement in machine-human interaction.[1] Trained to act as a novel chatbot technology that responds to text-based queries spanning general knowledge inquiries to complex conversational questions, ChatGPT is a large language model (LLM) trained on deep learning architecture and text-based big data. This technology enables the LLM to comprehend complex linguistic patterns and generate text responses reminiscent of human conversations.

In medicine, AI in the form of chatbot technology is a promising avenue to increase efficiency of clinicians via assistance with medical documentation, automated responses to electronic medical record-based patient portal inquiries, and laboratory testing or imaging screening and interpretation.[2,3] Prior investigations of LLMs have already yielded promising findings in helping users access relatively accurate medical information.[4–6] Large language models have been implemented in Bing (Microsoft's online search engine), and are changing how patients find and answer medical questions online.[7] They have even been shown to have higher quality and better empathy than physician responses.[8] Rather than searching for information online, users can pose questions to the LLM and directly receive answers to inquiries, albeit typically without sources or with inaccurate or false citations and potentially variable accuracy.

Given LLMs' potential to revolutionize how patients engage with broadly accessible online medical information, there is a need to assess the appropriateness and accuracy of an LLM's responses to ophthalmologic inquiries. This study qualitatively evaluated the appropriateness of a popular LLM's responses to simple and complex ophthalmology-related clinical inquiries from all specialties within ophthalmology.

PATIENTS AND METHODS

This was a cross-sectional study performed in April 2023 assessing the appropriateness of an LLM, ChatGPT-4 (OpenAI), in responding to simple and complex ophthalmology-related questions. For each ophthalmic subspecialty (comprehensive or cataract, cornea, glaucoma, neuro-ophthalmology, oculoplastic or orbital surgery, ocular oncology, pediatrics, vitreoretinal diseases, and uveitis), ~20 clinical questions (range: 20 to 25 questions) for each subspecialty area were generated and reviewed by Mayo Clinic faculty in the department of ophthalmology in Rochester, Minnesota. Subspecialists were encouraged to generate questions on the basis of common patient questions they might receive in clinic or via the patient portal. There were no other specific requirements for the expert opinion-generated questions, but reviewers were asked to generate a mix of common and nuanced questions, including inquiries related to risk factor counseling, disease etiology and pathogenesis, test result interpretation, medication counseling, and clinical experience. A total of 25 expert reviewers participated in the study; 22 reviewers both wrote and graded questions, and 3 reviewers wrote questions that were graded by same-subspecialty colleagues.

Consistent with methodology from prior literature, each question was posed to the LLM 3 times.[5] Each question generated 3 unique responses from the LLM. Responses to each question were graded by at least 1 subspecialist in his or her area of expertise. Based on the expert reviewer's clinical judgment, reviewers graded each set of responses as appropriate, inappropriate, or unreliable on the basis of the response's content in 2 grading contexts.[5,6] The first context was as if the information was presented on a patient information site that patients might find by a web search of their question (ie, mayoclinic.org or a similar institutional website). The second context was as an AI-generated draft response to a patient question sent to the physician via the electronic medical record (EMR) through the EMR portal. The latter context was added specifically to assess the potential for an LLM to

provide appropriate draft responses from physicians to patients. We defined appropriate responses as a response that was accurate and specific enough to serve as a surrogate for physician-approved information. In terms of grading responses, if any of the 3 AI-generated responses to a question were deemed to have inaccurate or inappropriate information per the grader, the question was graded as inappropriate in that specific context. If all 3 responses to a question were appropriate per the grader, the question in that context was graded as appropriate. If the 3 responses were inconsistent (ie, variable content across all 3 responses) but included appropriate content, then the entire question was graded as unreliable. For each subspecialty, we calculated the mean and median percentage of appropriate responses for both the patient information site and EMR draft response contexts. To calculate the overall percentage of appropriate responses, we calculated the average percent appropriate across all subspecialties for both contexts.

To investigate the performance of the LLM in responding to ophthalmology-related queries, we categorized each question within the various subspecialties into 4 distinct health information categories: disease and condition, risk and prevention, surgery-related, and treatment and management. It is important to note that if a question was applicable to multiple categories it was included in all applicable categories. We employed χ^2 tests to assess whether there were significant differences in the LLM's performance across these 4 categories, in both of the contexts previously detailed. Statistical testing across all subspecialties was completed with the Kruskal-Wallis test and pairwise comparisons were performed with Dunn's test with Bonferroni correction to adjust for multiple comparisons. All statistical testing was performed in Python (version 3.9).

RESULTS

In total, 192 questions were assessed by LLM and graded by 22 subspecialists across 9 subspecialties (Tables 1–5). By subspecialty, the number of graders was highest for vitreoretinal diseases (n=4) and neuro-ophthalmology (n=4) and lowest for ocular oncology (n=1) and uveitis (n=1) (Table 1). Aggregate grading for both contexts stratified by the 4 health information categories is detailed in Table 2. Individual grades for subspecialty questions are shown in Tables 3–5 and Supplemental Tables 6–11 (available online at https://www.mcpdigitalhealth.org/).

In the context of a patient information site, the LLM provided appropriate responses 79% of the time (Table 1). For draft responses to patient questions via EMR, the appropriate response rate was 74%. The top performing subspecialties in the context of a patient information site were: ocular oncology (100%), cataract or refractive surgery (92%; range, 80%−100%), pediatric ophthalmology (89%; range, 81%−100%), and vitreoretinal diseases (86%; range, 70%−100%) (Tables 1 and 3 and Supplemental Tables 8–10). The top performing subspecialties in the context of LLM-generated draft responses to patient questions through EMR were pediatric ophthalmology (94%; range, 81–100%), ocular oncology (90%), vitreoretinal diseases (88%; range, 75%−100%), and cataract or refractive surgery (85%; range, 70%−100%) (Tables 1 and 3 and Supplemental Tables 8–10). The worst performing subspecialties in the context of a patient information site were cornea (56%; range, 43%−67%), uveitis (65%), and neuro-ophthalmology (67%; range, 58%−79%)

(Tables 1 and 4 and Supplemental Tables 6 and 11). Similarly, the worst performing subspecialties in the context of LLM-generated draft responses to patient questions through EMR portal were cornea (54%; range, 38%−67%), uveitis (55%), oculoplastic or orbital surgery (62%; range, 56%−68%), and neuro-ophthalmology (67%; range, 58%−79%) (Tables 1 and 4 and Supplemental Tables 6 and 11). There were no significant differences in appropriateness rates in the context of a patient information site (P=.44) and draft responses to patient questions through EMR (P=.43). Pairwise comparison did not show any significant differences.

Glaucoma, pediatric ophthalmology, and vitreoretinal diseases reported higher appropriateness rates in terms of EMR draft responses to patient questions vs patient information site. Oculoplastic or orbital surgery reported the largest difference between both contexts (80% for patient information site vs 62% for draft responses to patient questions through EMR). In terms of individual graders across all subspecialties, multiple graders in different subspecialties (vitreoretinal disease, cataract or refractive surgery, ocular oncology, and pediatric ophthalmology) graded responses as 100% appropriate (Table 1). Conversely, the same grader in cornea issued the LLM the worst performance for both contexts at 43% and 38% respectively (Tables 1 and 5). Generally, inappropriate responses were related to inappropriate management recommendations (eg, incorrectly recommending crosslinking), incorrect factual information (eg, stating the wrong gene), and missing crucial information (eg, obtaining neuroimaging) (Tables 3–5 and Supplemental Tables 6–11). Unreliable responses lacked information in 1 or 2 of the 3 query attempts that would make the question appropriate but did not include anything that would warrant the grade of inappropriate (Tables 3–5 and Supplemental Tables 6–11).

The analysis of reviewer grades stratified by health information categories revealed notable variations in the assessment of both contexts (Table 2). For the context of a patient information site, the proportion of content graded as appropriate was highest in the disease and condition category (72.29%) and the lowest in surgery-related (54.72%). Conversely, the inappropriate content was most prevalent in the surgery-related (26.42%) and treatment and management (26.37%) categories. In contrast, the context of an LLM-generated response to a patient question showed a slightly lower yet insignificant percentage of appropriate content across all categories, with the highest in disease and condition (68.67%) and the lowest in surgery-related (50.94%). The incidence of inappropriate responses was generally higher, particularly in surgery-related (33.96%) and treatment and management (28.57%). The differences in health information categories in both contexts did not reach statistical significance, as indicated by the P-values (P=.28 for the patient information site and P=.78 for LLM-generated draft responses to patient questions through EMR).

DISCUSSION

We report robust aggregate appropriateness of an LLM across multiple ophthalmic subspecialties both in the context of a patient information site (56%−100%) and as responses to EMR patient messages to physicians (54%−90%). These results represent an important benchmark in ophthalmology for both patient information and education, as patients will inevitably use the LLMs to make medical decisions regarding their ophthalmic care. It is

essential for ophthalmologists to understand how these models work and to understand both their strengths and inherent weaknesses.

Large language models are AI models designed to understand and generate natural language text that can be used in numerous ways such as building a website from a notebook sketch, creating jokes, and performing at human levels on standardized tests.[9–11] The LLMs are trained on large volumes of text data across multiple fields and sources such as websites, articles, and text.[10,12] A key point is that the LLM's knowledge is entirely based on the information on which it was trained. It will not know new information after the training date unless it is updated.[10] Training enables the model to learn language in terms of structure, grammar, and phrases.[10] Training often is split into pretraining and fine-tuning.[12] In pretraining, the LLM learns language and knowledge from the sources, whereas in fine-tuning, the model is further refined on individualized tasks or data.[12] This fine-tuning enables subspecialization of the model that can be built toward specific tasks.[12] This training process requires astronomical computational resources as it needs to train on billions of parameters.[10,12]

Another important concept is how responses are generated to users, particularly in a question-answer context. When a user inputs text into a LLM it uses a technique called tokenization where the text is broken down into smaller parts called tokens, which can be as small as individual characters.[12] Tokenization enables LLMs to both process the user's input and understand it. The LLMs then discern the context of the question by weighing the importance of different words in the question to identify key information to generate an appropriate answer.[12] Once it has processed the input, the LLM leverages the information from training to generate responses.[12] To create contextual and coherent responses to users, LLMs are predictive, meaning they generate text by predicting a token at a time, which is conditionally based on the previously generated text until a complete response is produced.[10,12]

When aggregated across all ophthalmology subspecialties, there were nearly equivalent rates of appropriate responses when comparing draft messages to patient questions vs patient information sites. However, multiple graders noted that the information provided by the LLM was neither specific nor personalized enough for a response back to a patient and that the message length was verbose. There is inherently more subjectivity in online patient message responses by physicians as each physician has a unique electronic bedside manner, so it is not surprising that there was variability. Despite this, the mean physician appropriateness for online patient messages across all subspecialties was 74% of questions and minimum appropriateness was 54%. This reports an important proof of concept to augment physician efficiency, as implementation of a medically optimized and validated LLM could help physicians reduce electronic message burden through automatically generated responses and message drafts to patients.

The analysis of performance across different health information categories highlights the variable quality of health information. Notably, the disease and condition category consistently showed the highest appropriateness for both contexts. This could suggest that this LLM is more reliable when addressing general disease and condition information,

possibly because of the availability of structured and well-researched data in these areas. However, the marked decrease in appropriateness for surgery-related topics underscores potential gaps in nuanced or procedural knowledge. The higher rates of inappropriate and unreliable content in this category may reflect the complexity and variability inherent in surgical procedures, which might be challenging for LLMs to accurately interpret and convey. The lack of relevant statistical difference across all categories in both contexts suggests a broadly similar performance level; however, the subtle variations in content accuracy and reliability across different categories highlight the need for careful consideration when utilizing these sources for patient education and information dissemination.

In terms of inappropriate responses for a patient information website, the LLM consistently overgeneralized ophthalmic treatments or procedures, specifically inappropriately equating 1 surgery with the wrong procedure. An example of this was in the oculoplastic or orbital surgery section, in which the LLM would respond to eyelid surgery or blepharoplasty, which our subspecialists marked as inappropriate, as all eyelid surgery is not a blepharoplasty. Another example was in cornea where the model would inappropriately state that crosslinking or photorefractive keratectomy was indicated when it was not. Both the glaucoma and cornea sections had more questions with this error than others. This was surprising to some degree, as the model was quite robust at answering very niche subspeciality concepts and topics like ocular oncology. Another consistent error was related to questions regarding most common causes/treatments/medications where the model would appropriately state multiple common causes but then would list exceptionally rare causes rather than more appropriate prevalent causes. An example of this was in uveitis; when asked about common causes of posterior uveitis, the LLM did not list syphilis but listed more esoteric and rare causes. Another consistent theme leading to either inappropriate or inconsistent responses was the omission of critical information across the 3 responses. An example of this was a lack of neuroimaging recommendations for cranial nerve III palsy, which could have significant fatal clinical ramifications. All these errors would induce patient misinformation and could potentially cause harm. Overall, the LLM performed better on subspecialties where there were more common questions (ie, cataract or refractive surgery, pediatric ophthalmology, and vitreoretinal diseases) than nuanced questions (ie, uveitis, neuro-ophthalmology, and cornea).

The implications of this LLM's performance for clinical practice, patient education, and research are far reaching, and with that comes legal and ethical implications.[13] The LLMs can considerably improve patient education far beyond basic general ophthalmology questions as illustrated by our findings. Overall, our subspecialty experts were impressed at the level of conciseness, detail, and accuracy in the LLM's responses. Multiple individual graders and multiple subspecialties reported results similar to or better than previously reported LLM appropriateness rates in preventative cardiology (84%) and breast radiology (88%).[5,6] Niche fields such as ocular oncology reported excellent results, which is pertinent, as information and questions on diseases in these niche fields are often comparatively sparse. Furthermore, LLMs could improve patient education as they allow patients to theoretically directly access appropriate information without having to navigate and use judgment on the appropriateness of a plethora of online sources. This could benefit

care by reducing misinformation and improving physician appointments, as appropriately informed patients can better make informed decisions about their care. An additional benefit compared with typical online search engines is the complexity of questions that can be answered, particularly between 2 treatment options as the LLMs created a pros and cons list dynamically. Furthermore, the LLM can be designed to modify the education level of the responses, further improving the access to care.

Despite these benefits, there are significant caveats to these LLMs. The LLMs are trained from a variety of sources as mentioned previously, but the most concerning issue is the lack of transparency regarding the information used for training. Inherently, if the training information used to train the model is biased or outdated, then the underlying model will produce biased results, which is a critical concern regarding patient health information. Furthermore, this LLM and others do not provide any sources for the information generated, and there have been many episodes of LLM hallucinations or fabrication of completely incorrect information.[10] Finally, as stated previously, inappropriate information can adversely impact patient outcomes. Despite the numerous potential benefits, it is imperative that these caveats be addressed.

Strengths of this study include comprehensive analysis across all ophthalmology subspecialties, with multiple graders in most subspecialties, with the latest generation of a popular LLM. There are several limitations to this study. First, LLMs are not meant for medical use. Although this is the case, this will not stop patients from using it as we have already seen patients present based on the recommendation of a LLM. Second, appropriateness is inherently subjective, as practice patterns vary among ophthalmologists. This is particularly true in questions that specify the best. Furthermore, there may be selection bias in the question creation and grading by subspecialists. Subspecialty experts may be both harsher in their assessments than others and have a bias toward what the correct response is to their own question. Both our neuro-ophthalmology and cornea graders selected more nuanced questions, and they have specific well-known expertise (ie, myelin oligodendrocyte glycoprotein antibody disease and Fuchs endothelial corneal dystrophy) in these questions. Finally, 2 subspecialties in niche areas reported only 1 expert grader.[10]

CONCLUSION

In conclusion, this LLM reported relatively high rates of appropriateness across multiple ophthalmology subspecialties, although several subspecialties, specifically neuro-ophthalmology, uveitis, and cornea, reported relatively poor performance. Compared with other fields, we report slightly worse overall LLM performance, which is likely related to greater number of questions and graders, as some individual subspecialty performances were similar or better.[5,6] The LLMs have relevant potential benefits in both improving patient education and augmenting physician workflows; however, we must start to address fundamental concerns with LLMs, such as lack of transparency, bias associated with training data, and incorrect responses. Adoption of LLMs by consumers is rapidly increasing, and its widespread use by patients in ophthalmology is inevitable. Further training is required of LLMs, as the current available models are not yet sufficient to accurately replace physician-provided educational information and patient message responses. Ophthalmologists must

take an active role with oversight and future research of these models to both maximize beneficial clinical applications of LLMs and prevent patient misinformation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors thank Elizabeth A Bradley, MD for her clinical expertise.

Grant Support:

Leonard and Mary Lou Hoeft Career Development Award Fund in Ophthalmology Research, Grant Number P30 CA015083 from the National Cancer Institute, and CTSA Grant Number KL2 TR002379 from the National Center for Advancing Translational Science (NCATS). The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Abbreviations and Acronyms:

AI	artificial intelligence
ChatGPT	chat generative pretrained transformer
CRAO	acute central retinal artery occlusion
EMR	electronic medical record
LLM	large language models
SLT	selective laser trabeculoplasty

REFERENCES

1. Introducing ChatGPT. OpenAI blog Posted November, 2022. 30, https://openai.com/blog/chatgpt. Accessed April 4, 2023.
2. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233–1239. [PubMed: 36988602]
3. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med. 2023;388(13):1201–1208. [PubMed: 36988595]
4. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol. 2023;228(6):696–705. [PubMed: 36924907]
5. Sarraju A, Bruemmer D, Van Iterson E, et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA. 2023;329(10):842–844. [PubMed: 36735264]
6. Haver HL, Ambinder EB, Bahl M, et al. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. Radiology. 2023;307(4):e230424. [PubMed: 37014239]
7. Mehdi Y Confirmed: the new Bing runs on OpenAI's GPT-4. Microsoft blog Posted online March 14, 2023. https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4. Accessed April 4, 2023.
8. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–596. [PubMed: 37115527]

9. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. [PubMed: 36812645]
10. Achiam J, Adler S, Agarwal S, et al. GPT-4 Technical Report. Preprint Posted online December. 2023;19:arXiv:2303.08774. 10.48550/arXiv.2303.08774.
11. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol. 2023;141(6):589–597. [PubMed: 37103928]
12. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Preprint Posted online August 2, 2023. arXiv:1706. 03762. 10.48550/arXiv.1706.03762.
13. Bressler NM. What artificial intelligence chatbots mean for editors, authors, and readers of peer-reviewed ophthalmic literature. JAMA Ophthalmol. 2023;141(6):514–515. [PubMed: 37103930]

TABLE 1.

Appropriateness of Responses by a LLM by Ophthalmic Subspecialty

Subspecialty	Patient Information Site			LLM-Generated Draft Responses to Patient Questions by EMR		
	Mean Percent Appropriate	Median Percent Appropriate	Range Percent Appropriate (Number of Graders, n)	Mean Percent Appropriate	Median Percent Appropriate	Range Percent Appropriate (Number of Graders, n)
1. Cataract/refractive	92%	95%	80%–100% (n=3)	85%	85%	70%–100% (n=3)
2. Cornea	56%	57%	43%–67% (n=3)	54%	57%	38%–67% (n=3)
3. Glaucoma	72%	70%	50%–95% (n=3)	77%	70%	65%–95% (n=3)
4. Neuro-ophthalmology	67%	65%	58%–79% (n=4)	63%	65%	42%–79% (n=4)
5. Oculoplastic and orbital surgery	80%	80%	76%–84% (n=2)	62%	62%	56%–68% (n=2)
6. Ocular oncology	100%	100%	100%–100% (n=1)	90%	90%	90%–90% (n=1)
7. Pediatric ophthalmology	89%	85%	81%–100% (n=3)	94%	100%	81%–100% (n=3)
8. Vitreoretinal diseases	86%	88%	70%–100% (n=4)	88%	88%	75%–100% (n=4)
9. Uveitis	65%	65%	65%–65% (n=1)	55%	55%	55%–55% (n=1)
Total	79%	80%	56%–100% (n=24)	74%	70%	54%–94% (n=24)

EMR, electronic medical record; LLM, large language model.

TABLE 2.

Comparison of Reviewer Grades Stratified by Health Information Categories

Patient Information Site					
Grade Type	**Disease and Condition (n=83)**	**Risk and Prevention (n=24)**	**Surgery-Related (n=53)**	**Treatment and Management (n=91)**	***P***
Appropriate	72.29%	70.83%	54.72%	61.54%	.28
Inappropriate	19.28%	8.33%	26.42%	26.37%	
Unreliable	8.43%	20.83%	18.87%	12.09%	
AI-Generated Draft Response to Online Patient Questions to Physicians					
Grade Type	**Disease and Condition (n=83)**	**Risk and Prevention (n=24)**	**Surgery-Related (n=53)**	**Treatment and Management (n=91)**	***P***
Appropriate	68.67%	58.33%	50.94%	60.44%	.78
Inappropriate	22.89%	25.00%	33.96%	28.57%	
Unreliable	8.43%	16.67%	15.09%	10.99%	

Mayo Clin Proc Digit Health. Author manuscript; available in PMC 2024 April 04.

TABLE 3.

Evaluation of Cataract or Refractive Surgery Responses From an Online Artificial Intelligence Model by Comprehensive Ophthalmologists

Questions	Reviewer Grade—Patient Information Site	Reviewer Grade—Draft Responses to Online Patient Questions to Physicians
1. Do I need to wear glasses after cataract surgery?	A, A, A	A, A, A
2. What are the risks of cataract surgery?	A, A, A	A, A, I
3. What are the steps of cataract surgery?	A, A, A	A, A, A
4. Can I get cataract surgery in both eyes on the same day?	A, A, A	A, A, A
5. When is it time to get cataract surgery?	A, A, A	U, A, A
6. What age will I develop cataracts?	A, A, A	A, A, A
7. Why do I need laser procedure after cataract surgery?	A, A, A	A, A, A
8. What is the success rate of LASIK surgery?	I, A, U	I, A, U
9. What is the success rate of PRK surgery?	A, A, A	A, A, A
10. How long will LASIK or PRK surgery last?	A, A, A	A, A, A
11. Is LASIK surgery better than PRK surgery?	A, A, I	A, A, I
12. What are intraocular Collamer lenses?	A, A, A	A, A, A
13. What is SMILE surgery?	A, A, U	A, A, U
14. What are the biggest risks of laser refractive surgery?	A, A, U	A, A, U
15. Is a toric lens worth the money?	A, A, A	A, A, A
16. Is a multifocal lens worth the money?	A, A, A	A, A, A
17. Is an extended depth of field lens worth the money?	A, A, A	A, A, A
18. What if I am not happy with my vision after cataract surgery?	A, A, A	A, A, A
19. What is monovision?	A, A, A	I, A, U
20. How should I decide which intraocular lens to choose?	A, A, A	A, A, A

A, appropriate; I, inappropriate; PRK, photorefractive keratectomy; U, unreliable.

TABLE 4.

Evaluation of Cornea Responses From an Online Artificial Intelligence Model by Cornea Surgeons

Questions	Reviewer Grade—Patient Information Site	Reviewer Grade—AI-Generated Draft Responses to Online Patient Questions to Physicians
1. What causes Fuchs corneal dystrophy?	I, U, A	I, U, A
2. Is Descemet stripping endothelial keratoplasty or Descemet membrane endothelial keratoplasty a better option for Fuchs dystrophy?	I, U, U	I, U, U
3. What are the treatments for keratoconus?	I, I, I	I, I, I
4. What are the symptoms of keratoconus?	A, A, A	A, A, A
5. Is epithelium-on or epithelium-off crosslinking better for keratoconus?	A, A, A	A, A, A
6. Is there a risk if I wear contact lenses overnight?	A, A, A	I, A, A
7. How often should I replace my contact lenses?	A, A, I	A, A, I
8. Is there any treatment for anterior basement membrane dystrophy?	I, A, U	I, A, U
9. How long will I need to be on steroid eye drops after getting shingles in my eye?	I, I, A	I, I, A
10. How long should I take Valtrex or Acyclovir after a herpes simplex corneal infection?	U, A, I	U, A, I
11. Can I lose vision from a bacterial ulcer of the cornea?	A, A, A	A, A, A
12. What causes astigmatism after a corneal transplant?	A, A, A	A, A, A
13. What surgical procedure is best at preventing a pterygium from coming back?	I, I, I	I, I, I
14. What causes neurotrophic keratitis?	U, A, A	U, A, A
15. How should I treat my dry eye if lubricating drops and Restasis aren’t working?	I, I, I	I, I, I
16. How long will my cornea transplant last?	U, U, A	U, U, A
17. What causes dry eyes?	A, A, A	A, A, A
18. What is the best eye drop to use for dry eyes?	I, A, U	I, A, U
19. What is the best treatment for dry eyes?	A, A, A	A, A, A
20. What should I expect for a crosslinking procedure?	A, A, U	A, A, U
21. Will a corneal scar eventually go away?	I, A, A	I, A, A

A, appropriate; I, inappropriate; U, unreliable.

TABLE 5.

Evaluation of Glaucoma Responses From an Online Artificial Intelligence Model by Glaucoma Surgeons.

Questions	Reviewer Grade—Patient Information Site	Reviewer Grade—AI-Generated Draft Responses to Online Patient Questions to Physicians
1. What lifestyle changes can I do for my glaucoma?	A, A, I	A, A, A
2. How long do I need to take glaucoma drops?	A, A, A	A, A, A
3. Can I get my vision back that I have lost from glaucoma?	A, A, A	A, A, A
4. I have glaucoma. Do I need to worry about my family members also having glaucoma?	A, A, U	A, A, U
5. Can you get LASIK if you have glaucoma?	U, A, I	U, A, I
6. I have high intraocular pressure, do I need to be treated for glaucoma?	A, A, A	A, A, A
7. Will glaucoma drops improve vision?	A, A, A	A, A, A
8. Why do glaucoma drops make my eyes uncomfortable?	U, A, A	U, A, A
9. Will eye injections help my glaucoma?	U, A, I	U, A, I
10. Can stem cells treat my glaucoma?	A, A, A	A, A, A
11. What is this laser that can lower my eye pressure?	U, A, U	U, A, U
12. Is glaucoma surgery guaranteed to lower my pressure?	A, A, U	A, A, U
13. What is this glaucoma stent that can be done at the time of cataract surgery?	U, A, A	U, A, A
14. Is a tube better than a trabeculectomy for glaucoma?	I, A, A	I, A, A
15. Can my eye pressure be too low?	I, A, U	I, A, U
16. Can I check my eye pressure at home?	A, A, U	A, A, U
17. How will I know if my glaucoma is getting worse?	A, A, A	A, A, A
18. What happens if I miss doses of my glaucoma drops	A, A, A	A, A, A
19. Can your eye explode from high eye pressure?	A, A, U	A, A, A
20. What are the advantages and disadvantages of SLT vs glaucoma drops as initial therapy for glaucoma?	A, A, U	A, A, A

A, appropriate; I, inappropriate; U, unreliable.

Mayo Clin Proc Digit Health. Author manuscript; available in PMC 2024 April 04.