



Analysis of ChatGPT Responses to Ophthalmic Cases: Can ChatGPT Think like an Ophthalmologist?

Jimmy S. Chen, MD,^{1,2} Akshay J. Reddy, BS,³ Eman Al-Sharif, MD,^{1,4} Marissa K. Shoji, MD,¹ Fritz Gerald P. Kalaw, MD,^{1,2} Medi Eslani, MD,¹ Paul Z. Lang, MD,¹ Malvika Arya, MD,¹ Zachary A. Koretz, MD, MPH,¹ Kyle A. Bolo, MD,¹ Justin J. Arnett, MD,¹ Aliya C. Roginiel, MD, MPH,¹ Jiun L. Do, MD, PhD,¹ Shira L. Robbins, MD,¹ Andrew S. Camp, MD,¹ Nathan L. Scott, MD,¹ Jolene C. Rudell, MD, PhD,¹ Robert N. Weinreb, MD,^{1,2} Sally L. Baxter, MD, MSc,^{1,2} David B. Granet, MD, MHCM¹

Objective: Large language models such as ChatGPT have demonstrated significant potential in question-answering within ophthalmology, but there is a paucity of literature evaluating its ability to generate clinical assessments and discussions. The objectives of this study were to (1) assess the accuracy of assessment and plans generated by ChatGPT and (2) evaluate ophthalmologists' abilities to distinguish between responses generated by clinicians versus ChatGPT.

Design: Cross-sectional mixed-methods study.

Subjects: Sixteen ophthalmologists from a single academic center, of which 10 were board-eligible and 6 were board-certified, were recruited to participate in this study.

Methods: Prompt engineering was used to ensure ChatGPT output discussions in the style of the ophthalmologist author of the Medical College of Wisconsin Ophthalmic Case Studies. Cases where ChatGPT accurately identified the primary diagnoses were included and then paired. Masked human-generated and ChatGPT-generated discussions were sent to participating ophthalmologists to identify the author of the discussions. Response confidence was assessed using a 5-point Likert scale score, and subjective feedback was manually reviewed.

Main Outcome Measures: Accuracy of ophthalmologist identification of discussion author, as well as subjective perceptions of human-generated versus ChatGPT-generated discussions.

Results: Overall, ChatGPT correctly identified the primary diagnosis in 15 of 17 (88.2%) cases. Two cases were excluded from the paired comparison due to hallucinations or fabrications of nonuser-provided data. Ophthalmologists correctly identified the author in 77.9% ± 26.6% of the 13 included cases, with a mean Likert scale confidence rating of 3.6 ± 1.0. No significant differences in performance or confidence were found between board-certified and board-eligible ophthalmologists. Subjectively, ophthalmologists found that discussions written by ChatGPT tended to have more generic responses, irrelevant information, hallucinated more frequently, and had distinct syntactic patterns (all $P < 0.01$).

Conclusions: Large language models have the potential to synthesize clinical data and generate ophthalmic discussions. While these findings have exciting implications for artificial intelligence-assisted health care delivery, more rigorous real-world evaluation of these models is necessary before clinical deployment.

Financial Disclosures: The author(s) have no proprietary or commercial interest in any materials discussed in this article. *Ophthalmology Science* 2025;5:100600 © 2024 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



Supplemental material available at www.ophtalmologyscience.org.

Advances in artificial intelligence (AI) within domains such as deep learning^{1,2} and natural language processing^{3,4} have transformed medicine. Recently, the advent of generative AI, or AI trained to create original content, has resulted in exciting opportunities for image and text synthesis in medicine.^{5,6} In particular, large language models (LLMs) are a type of generative AI trained on

textual data in an unsupervised manner that learn to generate novel, relevant, and human-like outputs in response to user inputs. The human-like interactions and veracity of the responses from these LLMs have garnered them the nickname of "chatbots."⁷ To date, LLMs have been successful in taking licensing medical exams,^{8,9} question and answering,¹⁰⁻¹² writing notes and discharge

summaries,^{13,14} writing manuscripts,¹⁵ responding to patient concerns,^{16,17} and evaluating patient cases.¹⁸

Ophthalmology is a high-volume surgical subspecialty poised to benefit from LLMs. Prior studies have demonstrated that LLMs such as ChatGPT (OpenAI) and Bing-Chat (Microsoft) perform comparably to human respondents on ophthalmology board exam practice questions.^{19–21} ChatGPT has also shown aptitude in accurately answering questions in retinal disease,²² glaucoma,²³ oculoplastics,^{24,25} and myopia,²⁶ demonstrating the potential to increase access to information and improve patient and trainee education. Additionally, ChatGPT has demonstrated success in writing operative notes²⁷ and abstracts.²⁸ While prior work evaluating ChatGPT's ability to generate assessment and plans have been evaluated in other medical domains,^{18,29} there remains a gap in knowledge regarding how LLMs can generate assessments and plans in ophthalmic clinical scenarios.

To address this gap in knowledge, we sought to use prompt engineering, or designing prompts to guide and improve ChatGPT outputs, assessing the performance of ChatGPT out-of-the-box on publicly available ophthalmic cases. The purpose of this study was twofold: (1) to assess the accuracy of assessments and plans generated by ChatGPT and (2) to evaluate ophthalmologists' abilities to distinguish between responses generated by human clinicians versus ChatGPT.

Methods

The University of California, San Diego and ChatGPT were given permission by the Medical College of Wisconsin to use their Ophthalmic Case Studies. This study was deemed exempt upon review by the University of California, San Diego Institutional Review Board and adheres to the Declaration of Helsinki.

Chat-GPT Generated Discussions

All 17 ophthalmic cases from the Medical College of Wisconsin Ophthalmology Case Studies were extracted.³⁰ These cases were written by a board-certified ophthalmologist for medical student education. Each ophthalmic case consisted of a complete history and ophthalmic examination (eye vitals, a slit lamp examination, a dilated fundus examination, and ancillary testing including gonioscopy, OCT, and fluorescein angiography, if applicable). Detailed discussion regarding the case diagnosis, disease definition, pertinent examination findings for the case diagnosis and its differential diagnoses, and treatment were described for each case. Cases span multiple areas of ophthalmology and were not limited to specific subspecialty domains.

All cases were input into ChatGPT version 3.5. We chose version 3.5 because (1) it is publicly available without a subscription fee and (2) to allow for performance comparisons as technical optimizations for ChatGPT are currently limited to version 3.5. Prior to case input, prompt engineering was performed to ensure ChatGPT understood the requested task. Specifically, the prompt was designed by J.S.C. and A.J.R. with the following requirements: (1) to act like an expert ophthalmologist and (2) write its response in a format identical to the clinician-written discussion. In other words, we designed a prompt requesting ChatGPT to generate a discussion in the style of the ophthalmologist writing the Medical College of Wisconsin cases. The prompt was designed

such that the requirements were described first, and the original history and ophthalmic examination were provided subsequently. The same prompt was provided to ChatGPT for all cases. Cases were provided nonsequentially to ChatGPT; in other words, a new chat session was open for each case, thus, ChatGPT was not given an opportunity to learn from prior cases. Prompts were also tested across different windows on different days. Upon manual review, ChatGPT responses were found to be similar and repeatable across retesting. An example of the history, examination, and paired discussions is shown in Table 1 and 2. The final prompt used to elicit discussions from ChatGPT is available in Table S3 (available at www.opthalmologyscience.org), and the labeled discussions generated by clinicians and ChatGPT are available in Table S4 (available at www.opthalmologyscience.org). The survey form used to elicit clinician responses is available in Table S5 (available at www.opthalmologyscience.org).

Each ChatGPT-generated discussion was manually reviewed by an ophthalmologist (J.S.C.) for accuracy of diagnosis. Cases where the primary diagnosis was identical between the ChatGPT-generated and clinician-generated discussions and no hallucination, or fabrication of data that was not user-provided, occurred in the primary diagnosis were included in the survey. Cases where the ChatGPT-generated diagnosis was incorrect were also manually reviewed.

Ophthalmologist Recruitment

Board-eligible and board-certified ophthalmologists at the University of California, San Diego were recruited to participate in an evaluation distinguishing ChatGPT-generated and clinician-generated discussion for each included case between September 1, 2023 and December 15, 2023 via a faculty-wide email. Each ophthalmologist was provided the patient history and examination and was masked to the authorship of 2 provided discussions side-by-side for each case in randomized order (Fig 1). Participating ophthalmologists were ensured to have not seen the cases before. The ophthalmologists subsequently were given a link to an anonymized Google Forms survey with the following questions for each case: (1) Which discussion was generated by ChatGPT? (2) How confident are you in your answer (using a 5-point Likert scale from 1 = very unconfident to 5 = very confident)? and (3) A free-text option to comment on their choices. Subjective feedback could include anything from factors justifying their selection or comments on the clinical content of the discussion.

Statistical Analysis

Data analysis was performed using R version 4.0.5 (R Foundation for Statistical Computing). For all included and excluded cases, the Flesch–Kincaid Grade score was calculated, which is a score for calculating the grade level difficulty of specific texts.³¹ These scores were calculated for each clinician-generated and ChatGPT-generated response. Sums and percentages were calculated for the total correct responses for each case as well as the correctly identified discussions for each ophthalmologist. Mean Likert scores assessing confidence in response were calculated for each case, and an R^2 was calculated to assess the correlation between response confidence and accuracy. Ophthalmologists' comments for included cases and reasons for exclusion were compiled and tallied for similar comments and themes. Comments were manually reviewed and summarized by common themes and observations among each ophthalmologist. Overall, the following 5 themes were identified: (1) irrelevance or incompleteness of content, (2) genericness or vagueness of content, (3) inaccurate content (i.e., medically incorrect statements), (4) hallucination of data not presented (i.e., describing data that does not exist), and (5) strange

Table 1. Example History and Exam from the Medical College of Wisconsin Case Studies

Example History	Example Exam	
Patient history	Ocular exam	
HPI:	Visual acuity (cc):	
A 72-yr-old female retired CEO reports that	OD: 20/30	
approximately 2 wks ago she noticed the onset of	OS: 20/100	
decreased vision in her left eye associated with a	IOP (tonoapplanation):	
“black spot” close to the center of her vision. The spot	OD: 16 mmHg	
is present with both near and far vision and does not	OS: 17 mmHg	
move away. The vision has been slightly blurred for a	Pupils:	
while. She denies any recent trauma, eye pain,	Equal, round, and reactive to light, no APD OU	
redness, or discharge. Peripheral vision is normal. She	Extraocular movements:	
has not had an eye exam in 5 yrs.	Full OU. No nystagmus.	
Past ocular history:	Confrontational visual fields:	
No prior eye surgeries. No hx of eye trauma, amblyopia,	Full to finger counting OU.	
or strabismus.	External:	
Ocular medications:	Normal, both sides	
None	Slit lamp:	
Past medical history:	Lids and lashes	Normal OU
Hypertension, hypercholesterolemia, coronary artery	Conjunctiva/sclera	Normal OU
disease s/p myocardial infarction s/p stenting 5 yrs	Cornea	Clear OU
prior.	Anterior chamber	Deep and quiet OU
Surgical history:	Iris	Normal OU
C-sections x 2, appendectomy 15 yrs ago	Lens	1+ NSC OU
Past family ocular history:	Anterior vitreous	Clear OU
No evidence of blinding diseases, glaucoma, or macular	Dilated fundus examination:	
degeneration	OD	Clear view, CDR 0.5 with sharp optic disc
Social history:		margins, flat macula, scattered large soft
Past 50 pack/yr smoking history. No alcohol or drug use.		drusen within the arcades, normal vessels,
Currently retired and living with her husband		and peripheral retina
Medications:	OS	Clear view, CDR 0.6 with sharp optic disc
Aspirin, Plavix, Lipitor, lisinopril, metoprolol		margins, flat macula, scattered soft drusen
Allergies:		within the arcades, 1 disc area of
ROS:		subretinal hemorrhage in macular area
Denies recent illness or any new CNS, heart, lungs, GI,		near fovea
skin, or joint symptoms.	Amsler grid:	
	OD: Normal	
	OS: Blurry spot near the center of	
	the grid with wavy lines	

APD OU = no afferent pupillary defect in both eyes; CDR = cup-to-disc ratio; CEO = Chief Executive Officer; CNS = central nervous system; GI = gastrointestinal; HPI = history of present illness; IOP = intraocular pressure; NSC = nuclear sclerotic cataract; OD = right eye; OS = left eye; OU = both eyes; ROS = review of systems.

syntax or tone (i.e., nonhuman voice or grammatical errors). The number of occurrences for each theme that was commented on by an ophthalmologist was quantified for each case. Statistically significant differences were assessed via the Mann–Whitney U test due to sample size, with $P \leq 0.05$ thresholded for significance.

Results

Identification of ChatGPT-Generated versus Clinician-Generated Discussions

ChatGPT correctly identified the primary diagnosis in 15 of the 17 (88.2%) available cases and hallucinated additional diagnostic factors in the primary diagnosis in 2 cases. Excluded cases included misdiagnoses of the following: esotropia and amblyopia (which ChatGPT labeled as only esotropia), globe rupture (which ChatGPT labeled as a corneal laceration), leukocoria (which ChatGPT correctly diagnosed but hallucinated a cataract), and papilledema

secondary to intracranial mass (which ChatGPT correctly diagnosed but hallucinated pituitary adenoma). Thus, 13 cases were included in the survey. Sixteen ophthalmologists, of which 10 were board-eligible and 6 were board-certified, completed the survey and blind identification of all discussions from the 13 cases. Of the 16 ophthalmologists, 7 (43.8%) were female. Four of the ophthalmologists (25.0%) were glaucoma trained, 4 (25.0%) were retina trained, 3 (18.8%) were pediatrics trained, 2 (12.5%) were ophthalmology trained, 2 (12.5%) practiced comprehensive ophthalmology, and 1 was cornea trained (6.2%).

Overall, ophthalmologists correctly identified the author of the discussions in $77.9\% \pm 26.6\%$ of the cases, of which board-eligible ophthalmologists correctly identified the authors in $79.2\% \pm 23.5\%$ of cases and board-certified ophthalmologists correctly identified the authors in $75.6\% \pm 33.5\%$ of cases. There was no statistically significant difference between the performances of the board-eligible and board-certified ophthalmologists ($P = 0.8$).

Table 2. Example Paired Generated Discussions by the Medical College of Wisconsin Ophthalmologist versus ChatGPT

Case 10—Response 1 [Ophthalmologist]	Case 10—Response 2 [ChatGPT]
<p>Diagnosis: Wet age-related macular degeneration</p> <p>Discussion: Differential diagnosis: The appearance of bilateral soft drusen and an area of subretinal hemorrhage in the left eye is consistent with the diagnosis of wet age-related macular degeneration. The subretinal hemorrhage is most likely due to choroidal neovascularization (CNV). Other causes of subretinal hemorrhages and decreased vision include retinal arterial macroaneurysms, diabetic retinopathy, hypertensive retinopathy, Valsalva retinopathy, and other causes of CNV (high myopia, angioid streaks, presumed ocular histoplasmosis, or trauma).</p> <p>Definition: Age-related macular degeneration (AMD) is a degenerative retinal disease. Two main types exist although many classification schemes are in use. Dry AMD is characterized by the appearance of hard or soft drusen and/or areas of retina pigmented epithelium (RPE) loss or hypertrophy. Large areas of RPE loss, termed geographic atrophy, are mainly responsible for decrease vision in this type of AMD. Wet AMD, comprising 15% of all AMD cases, results in severe central visual loss if not treated promptly. In wet AMD pathologic choroidal neovascular membranes develop under the retina, RPE, or both resulting in destruction of retinal architecture and formation of fibrovascular scars and vision loss. Risk factors for the development of AMD are genetic and environment and include mutations in the complement factor H gene, family history of AMD, age, and White race. Many associations with other diseases exist including smoking and hypertension. High risk populations for progression from dry to wet AMD include those with large drusen, many intermediate-sized drusen, noncentral geographic atrophy, or advanced AMD in the other eye.</p> <p>Examination: Patients with AMD are recommended to monitor their vision, each eye separate, with the Amsler grid (a series of horizontal and vertical lines forming a grid of squares). Appearance of a blind or discolored spot or waviness of the straight lines (metamorphopsia) could signal alterations in the photoreceptor layer and should be evaluated for the development of wet AMD. Fluorescein angiogram (FA) is the gold standard to identify areas of abnormal vasculature within or under the retina consistent with CNV. Typical findings include areas of hyperfluorescence ± leakage. The OCT technique allows for identification of intra or subretinal fluid in the macula. In patients with CNV this corresponds to extravasation of fluid from leaky capillaries. Response to therapy is seen with improvement of vision, decrease of fluid in OCT, and decrease of leakage and/or hyperfluorescence in FA.</p> <p>Treatment: The age-related eye disease study (AREDS) vitamin formulation is recommended for patients with severe forms of dry AMD. These vitamins (a specific combination of vitamins A, C, E, zinc, and copper) have been shown to decrease the progression of AMD. For vision-threatening wet AMD, the most effective therapy is anti-VEGF agents. Prompt treatment can stabilize or improve visual acuity in two-thirds of patients. Other less successful treatment modalities include photodynamic therapy and laser photocoagulation.</p>	<p>Diagnosis: Wet age-related macular degeneration (AMD) with subretinal hemorrhage in the left eye</p> <p>Discussion: Differential diagnosis: The primary diagnosis in this case is wet age-related macular degeneration (AMD) with subretinal hemorrhage in the left eye. The sudden onset of decreased vision associated with a black spot in the center of vision, blurred vision, and the presence of subretinal hemorrhage are consistent with the characteristics of wet AMD. Other potential causes of central vision loss, such as central retinal artery or vein occlusion, diabetic macular edema, or macular hole, should also be considered. However, the clinical findings, including drusen and subretinal hemorrhage, align with the diagnosis of wet AMD.</p> <p>Definition: Wet age-related macular degeneration is an advanced form of AMD characterized by the growth of abnormal blood vessels beneath the macula. These blood vessels can leak fluid and blood into the macula, leading to vision loss. Drusen, which are yellow deposits under the retina, are often present in AMD.</p> <p>The primary risk factor for AMD is age, and it is more common in individuals over the age of 50. Other risk factors include a family history of AMD, smoking, hypertension, and genetic predisposition. The patient's history of smoking and older age aligns with common risk factors for AMD.</p> <p>Examination: The patient's ocular examination revealed decreased visual acuity of 20/100 in the left eye, along with normal intraocular pressure in both eyes. Pupils were equal, round, and reactive to light, with no afferent pupillary defect. Extraocular movements were full, and no nystagmus was observed. Confrontational visual fields were full to finger counting in both eyes. The slit lamp examination showed normal lids, lashes, conjunctiva, sclera, cornea, iris, and lens. The dilated fundus examination revealed a clear view, cup-to-disc ratios of 0.5 and 0.6 with sharp optic disc margins, and flat maculae in both eyes. In the left eye, scattered soft drusen were present within the arcades, along with a subretinal hemorrhage in the macular area near the fovea. The Amsler grid test showed a blurry spot near the center of the grid with wavy lines in the left eye.</p> <p>Treatment: The primary treatment for wet AMD involves anti-VEGF therapy. These medications are injected into the eye to inhibit the growth of abnormal blood vessels, reduce leakage, and preserve or improve vision. Multiple injections are typically needed over time, and close monitoring is essential to assess treatment response and determine the need for further injections. Early intervention with anti-VEGF therapy offers the potential to stabilize or even improve vision in wet AMD. Treatment aims to reduce the leakage of fluid and blood into the macula, thereby preserving central vision and preventing further visual deterioration. Regular follow-up appointments and ongoing management are crucial to ensure the best possible visual outcomes for the patient.</p>

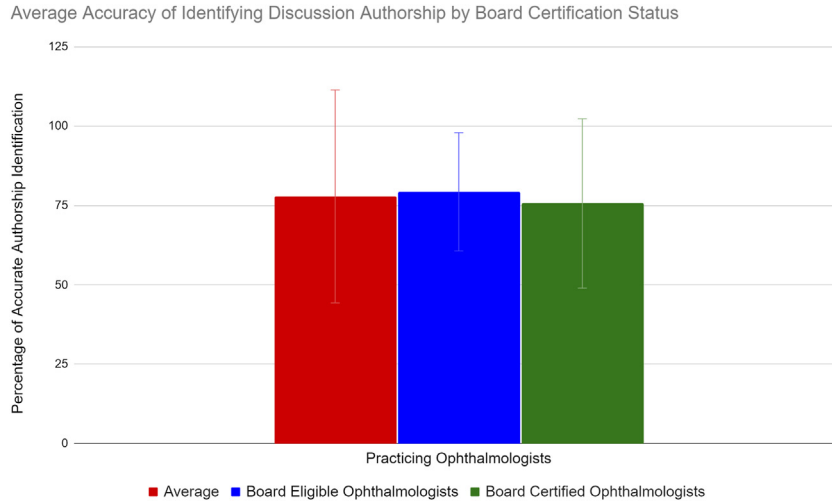


Figure 1. Percentage of accurate identification of ChatGPT vs. ophthalmologist discussion authorship by board-eligible and board-certified ophthalmologists.

The Flesch–Kincaid reading level scores between the ophthalmologist-written (14.5 ± 1.4) and ChatGPT-generated discussions (13.9 ± 1.3) were also fairly similar ($P = 0.31$). Ophthalmologists had the most difficulty distinguishing between authors for corneal cases (traumatic corneal abrasion and chemical burn) (accuracy 56.25%–62.5%), and also relatively lower accuracy (69%) in distinguishing authorship for cases on proliferative diabetic retinopathy, giant cell arteritis, and rhegmatogenous retinal detachments. These data are shown in Table 6 and Figure 1.

Analysis between Confidence and Accuracy

Ophthalmologists generally felt somewhat confident that they could identify the author of the discussions (mean Likert score 3.6 ± 1.0), which was similar among board-eligible ophthalmologists (3.6 ± 0.9) and board-certified ophthalmologists (3.4 ± 1.0) [$P = 0.38$]. There was a

weak correlation between response confidence and accuracy ($R^2 = 0.32$), as shown in Figure 2.

Ophthalmologist Impressions of ChatGPT Responses

Compared with human-generated discussions, the ophthalmologists generally felt that ChatGPT-generated discussions contained more irrelevant/incomplete information (1.0 ± 1.5 vs. 3.2 ± 2.3 ophthalmologist comments per case, respectively, $P < 0.01$), contained more generic content (0.8 ± 0.6 vs. 3.9 ± 2.3 ophthalmologist comments per case, respectively, $P < 0.01$), hallucinated data more frequently (0.0 ± 0.0 vs. 0.3 ± 0.8 ophthalmologist comments per case, respectively, $P < 0.01$), and had odd syntax or nonnatural tone (0.8 ± 1.0 vs. 3.5 ± 2.2 ophthalmologist comments per case, respectively, $P < 0.01$). There was no statistically significant difference between the number of

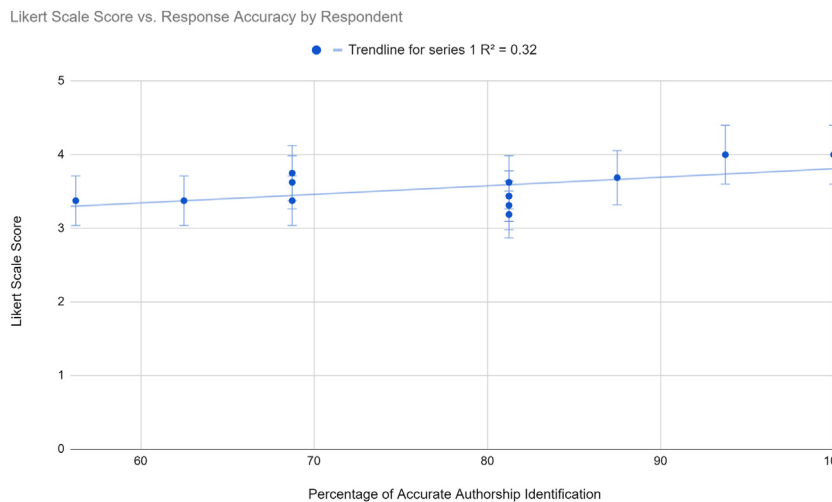


Figure 2. Correlation between Likert scale confidence score and response accuracy regarding authorship by respondent.

Table 6. Performance of Ophthalmologists by Case

Case Designation	Case Diagnosis	Overall Percentage of Correctly Distinguishing Discussion Authors	Percentage of Correctly Distinguishing Discussion Authors by Board-Eligible Ophthalmologists	Percentage of Correctly Distinguishing Discussion Authors by Board-Certified Ophthalmologists	Human-Authored Flesch–Kincaid Score	ChatGPT-Authored Flesch–Kincaid Score
Total case average		77.88 ± 33.5	79.23 ± 23.51	75.64 ± 26.63	14.48 ± 1.40	13.91 ± 1.26
Case 1	Suspected acute angle closure glaucoma right eye	87.5	100	66.67	14.31	14.00
Case 2	Acute conjunctivitis of both eyes	81.25	80	83.33	15.86	15.13
Case 3	Traumatic corneal abrasion	56.25	40	83.33	14.23	13.39
Case 4	Postoperative endophthalmitis	93.75	90	100	14.59	15.79
Case 6	Traumatic hyphema	100	100	100	13.71	13.12
Case 7	Rhegmatogenous retinal detachment resulting from a peripheral retinal tear	68.75	80	50	13.03	12.85
Case 10	Wet age-related macular degeneration	81.25	90	66.67	14.11	11.54
Case 11	Chemical (sulfuric acid from car battery) burn to eye	62.5	70	50	12.45	13.23
Case 13	Central retinal vein occlusion (CRVO) left eye	81.25	80	83.33	14.18	13.85
Case 14	Proliferative diabetic retinopathy in both eyes	68.75	70	83.33	17.53	16.03
Case 15	Grave's ophthalmopathy	81.25	80	83.33	16.14	13.73
Case 16	Giant cell arteritis (GCA) or temporal arteritis	68.75	60	83.33	14.99	13.24
Case 17	Primary open-angle glaucoma	81.25	60	83.33	13.05	14.95

Table 8. Feedback Given by Respondents to Both ChatGPT-Generated and Human-Generated Discussions Grouped by 5 Common Themes: (1) Irrelevant/Incomplete Content, (2) Generic Content, (3) Inaccurate Statements, (4) Hallucinations, and (5) Abnormal Syntax or Tone

Theme	Representative Comments
Irrelevant or incomplete content	<p>“Tangential discussion, extremely broad diagnosis not applicable to this patient” (Case 1)</p> <p>“... [The] Exam for response 2 is very methodical and narrow and answers the question directly and not much more than [Response 1]- the giveaway for me is mentioning exam findings such as pupils are round and reactive, which a clinician may not usually deem pertinent enough to mention in a quick clinical summary. Response 1 also gives prognosis data that 2/3 of patients will have better vision, which is important for counseling patients - it sounds like it's coming from someone who's actually counseled patients before.” (Case 10)</p>
Generic content	<p>“Generic examination, but other sections were more difficult to differentiate. Response 1 [ChatGPT] overall was less descriptive” (Case 11)</p> <p>“The differential diagnosis appears generic in the ChatGPT-generated output.” (Case 4)</p>
Inaccurate statements	<p>“The exam section summarized the findings again without providing further details on what to expect when examining these patients. Also, it mentioned in the treatment section that we sometimes patch to prevent further trauma and I think this is inaccurate.” (Case 3)</p> <p>“This statement is inaccurate: “Smoking and genetic factors may also contribute to the development of the condition [Giant Cell Arteritis]. and this is probably the response generated by GPT” (Case 16)</p>
Hallucinations	<p>“Response 2 referenced an image (suggesting that it was the original discussion associated with the case).” (Case 13)</p> <p>“Case 14 lists both eyes as having PDR when only one eye has PDR” (Case 14)</p>
Syntax or tone criticism	<p>“Response 1 has a more “teaching” (human?) tone - particularly the exam section as well as the treatment section particularly when it talks about the success rate of surgery, like the clinician is talking to residents about something that may come up on a test, rather than just answering the question at hand.” (Case 7)</p> <p>“Use of dashes as punctuation is more likely a human.” (Case 7)</p> <p>“Weird vague concluding statements at the end of certain paragraphs like, ‘Early intervention can greatly improve the prognosis and prevent severe complications, including vision loss.’” (Case 16)</p>

comments regarding inaccurate statements between human-generated and ChatGPT-generated discussions (0.6 ± 1.0 vs. 0.7 ± 0.9 , respectively, $P = 0.89$) (Table S7, available at www.ophtalmologyscience.org).

Representative comments highlight the above data. For example, one ophthalmologist noted the following differences between human-generated (response 1) and ChatGPT-generated discussions (response 2) for case 7: “Response 1 has a more “teaching” (human?) tone - particularly the exam section as well as the treatment section particularly when it talks about the success rate of surgery, like the clinician is talking to residents about something that may come up on a test, rather than just answering the question at hand. The definition for response 1 also goes through a more thorough differential of the different types of retinal detachments whereas response 2 goes through less likely differentials such as CRAO/CRVO and AMD.” Another ophthalmologist commented on ChatGPT’s generated discussion (response 2) with the following comments for case 10: “...[The] Exam for response 2 is very methodical and narrow and answers the question directly and not much more than [Response 1]- the giveaway for me is mentioning exam findings such as pupils are round and reactive, which a clinician may not usually deem pertinent enough to mention in a quick clinical summary. Response 1 also gives prognosis data that 2/3 of patients will have better vision, which is important for counseling patients - it sounds like it's coming from someone who's actually counseled patients before.” A full list of comments is supplied in Table 8.

Discussion

In this study, we used ChatGPT to generate ophthalmic discussions of ophthalmic diseases, including assessment

and plans, and assessed the ability of ophthalmologists to distinguish whether these discussions were written by a human or ChatGPT. Our study has 2 key findings: (1) ChatGPT can generate ophthalmologist-style discussions, although its writing style may limit its utility in ophthalmic writing and (2) ophthalmologists demonstrate moderate performance in distinguishing between ChatGPT versus ophthalmologist-generated discussions.

ChatGPT has demonstrated both domain aptitude in ophthalmology knowledge^{19,20,22} as well as scientific writing abilities.^{15,27,28} However, few studies have demonstrated the application of its medical knowledge to write in a clinical style.¹⁸ Here, we find that ChatGPT can correctly identify the diagnosis in 88.2% of cases comprising common ophthalmic diagnoses. Additionally, ChatGPT was able to write at a similar reading level to ophthalmologists and generate generally appropriate differential diagnoses, discuss pertinent examination findings, and discuss treatment similar to an ophthalmologist (Table 1 and 2), essentially simulating clinician diagnostic processes.³² Although subjective perceptions of content accuracy were generally similar to that of the ophthalmologist, there were significant concerns regarding the writing style and nuances regarding its presentation of ophthalmic reasoning, particularly with regard to irrelevant and generic information. For example, 8 of the 16 ophthalmologists commented that ChatGPT outputs irrelevant differential diagnoses for case 1, specifically scleritis and uveitis for angle closure glaucoma (Table 8). Similarly, ophthalmologists often commented on how easily they were able to distinguish human-like writing by the clinician on the basis of syntax patterns recycled by ChatGPT (i.e., “The primary diagnosis is...”) and sentence architecture. These findings are consistent with prior concerns

regarding the generic nature of ChatGPT writing²⁷ and the accuracy of its scientific writing.²⁸ In fact, prior AI models trained to discern ChatGPT- versus human-written scientific writing demonstrated similar findings regarding the importance of syntactic difference (i.e., sentence-level diversity, use of “popular” words) in identifying work by ChatGPT.^{33,34} Altogether, these data highlight the need to exercise caution when employing ChatGPT for scientific writing at this stage in its development, particularly in specialized fields like ophthalmology.

Despite linguistic limitations in ChatGPT’s writing, ChatGPT’s ability to synthesize information in a clinician-like style introduces exciting potential applications to improve eye care. In this study, the ophthalmologists’ performance in distinguishing human-generated versus ChatGPT-generated discussion was highly variable, ranging from 56% to 100% (Fig 1), with modest confidence. Similar results have demonstrated ophthalmologists’ difficulty distinguishing AI versus human responses to eye forum questions.¹⁷ Altogether, these findings suggest that LLMs have the potential to synthesize ophthalmic data and output this data synthesis in a written manner similar to ophthalmologists, akin to clinical reasoning and note writing. With increasing reliance on tools such as electronic health records in the digital era of medicine, LLMs have the potential to augment the capability and reach of ophthalmologists with regard to documentation (note writing, patient letter writing, and discharge summaries), data summarization, addressing patient concerns, and scientific reasoning.^{35–37} While these LLM-based systems are not designed to and should not replace human clinicians, integration of LLMs into clinician workflows to provide basic ophthalmic knowledge and reduce time spent in the electronic health record (i.e., chart review, note writing) may allow ophthalmologists to spend more time on face-to-face patient counseling and more complex human aspects of patient care.³⁸ Additionally, LLMs optimized for ophthalmic care delivery and communication may potentially improve health care access for patients with limited access to ophthalmologists. Importantly, meticulous prompt engineering, or design of prompts to elicit responses from and guide LLMs, likely played a role in the performance of ChatGPT both in terms of diagnostic accuracy and similar discussion-writing. Prior literature has demonstrated that prompt engineering plays a key role in optimizing the output of LLMs.^{39–41} Thus, thoughtful inputs, LLM optimization for clinical use, and rigorous real-world evaluations would be needed before clinical deployment of these models.

Although there are numerous benefits of LLMs, the potential for patient harm remains a key consideration. An important limitation of LLMs is their tendency to hallucinate and fabricate data,^{42,43} as demonstrated by the results in our study (Table S7, available at www.ophthalmologyscience.org). Interestingly, our study did not find significant differences in perceptions of accurate information in human- versus ChatGPT-generated discussions, likely due to incorporation of “completeness”

of data (i.e., complete differentials), a subjective perception, into this definition of accuracy. Regardless, both hallucinations and fabrications introduce a concern regarding the ethical use of AI, specifically its potential to harm the patient.⁴⁴ For example, LLMs may hallucinate psychiatric diseases such as major depressive or bipolar disorder for a patient, resulting in unnecessary and undue stigmatization. Similarly, fabrication of data may result in deception or misinformation for patients and the broad public, an issue that has arisen with fake news articles and conspiracy theories in journalism.^{45,46} Within medicine, misdiagnosis or inaccurate advice suggested by ChatGPT may generate more patient anxiety and has important legal implications if the concerns are vision- or health-threatening. These concerns highlight the need for ophthalmologist supervision and collaboration with these LLMs and reiterate the need for rigorous testing to mitigate the risks of patient harm from these unsupervised models.

Our study has additional limitations. First, our study was limited with regard to the number of cases analyzed and number of ophthalmologists participating, and thus it may be difficult to draw generalizable conclusions from our analyses. Further evaluation of ChatGPT on a larger set and diverse set of cases, including rare diseases, as well as involving ophthalmologists in various practice settings and stages in their careers is needed. Second, these cases were designed by an ophthalmologist as representative vignettes initially designed for educational purposes. It is currently unclear how ChatGPT would perform in real-world clinical scenarios. Third, the performance of ChatGPT was not compared with that of other LLMs such as Claude and Bard, which have demonstrated variable performance across several ophthalmic tasks. Finally, we did not explore other more computationally intensive methods for ChatGPT optimization such as fine-tuning and parameter customization. More research is needed to understand how fine-tuning ChatGPT affects its performance within ophthalmology-specific tasks.

Large language models such as ChatGPT have significant potential to generate ophthalmic discussions. Although these findings demonstrate that LLMs have the potential to write and synthesize clinical data similar to writing and synthesizing clinical information, significant caution and more research is needed to understand how LLM-generated writing could both benefit and harm patients before widespread clinical deployment. Ultimately, LLMs have already transformed the way humans interact with information and have started to affect health care delivery. Multidisciplinary collaboration between clinicians, informaticians, and policy leaders will be essential to develop thoughtful, ethical, and innovative implementations of LLMs.

Acknowledgments

The authors would like to thank Dr Judy Hoggatt, Dr Heather Stiff, and the Medical College of Wisconsin (MCW) Department of Ophthalmology and Visual Sciences for their publicly available Ophthalmic Case Studies for Medical Students. This manuscript was submitted with their permission and consent for publication.

Footnotes and Disclosures

Originally received: January 20, 2024.

Final revision: August 9, 2024.

Accepted: August 13, 2024.

Available online: August 23, 2024. Manuscript no. XOPS-D-24-00024.

¹ Viterbi Family Department of Ophthalmology, Shiley Eye Institute, University of California, San Diego, La Jolla, California.

² UCSD Health Department of Biomedical Informatics, University of California San Diego, La Jolla, California.

³ School of Medicine, California University of Science and Medicine, Colton, California.

⁴ Surgery Department, College of Medicine, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s):

R.N.B.: Consultant – Topcon, AbbVie, Alcon.

S.L.B.: Equipment (nonfinancial support) – Optomed and Topcon, outside of the submitted work.

HUMAN SUBJECTS: Human subjects were included in this study. The University of California, San Diego and ChatGPT were given permission by the Medical College of Wisconsin to use their Ophthalmic Case Studies. Reviewed by legal counsel on March 21, 2024. This study was deemed

exempt upon review by the University of California, San Diego Institutional Review Board and adheres to the Declaration of Helsinki.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Chen, Reddy, Weinreb, Baxter, Granet

Data collection: Chen, Reddy, Al-Sharif, Shoji, Kalaw, Eslani, Lang, Arya, Koretz, Bolo, Arnett, Roginiel, Do, Robbins, Camp, Scott, Rudell, Baxter

Analysis and interpretation: Chen, Reddy

Obtained funding: Study was performed as part of regular employment duties at the University of California, San Diego. No additional funding was provided.

Overall responsibility: Chen, Reddy, Al-Sharif, Shoji, Kalaw, Eslani, Lang, Arya, Koretz, Bolo, Arnett, Roginiel, Do, Robbins, Camp, Scott, Rudell, Weinreb, Baxter, Granet

Abbreviations and Acronyms:

AI = artificial intelligence; **LLM** = large language model.

Keywords:

Artificial Intelligence, ChatGPT, Large language models, Medical Education, Ophthalmology.

Correspondence:

David B. Granet, MD, MHCM, Anne F. Ratner Chair in Pediatric Ophthalmology, Ratner Eye Center, Shiley Eye Institute, 9415 Campus Point Dr MC0946, La Jolla, CA, 92093. E-mail: dgranet@health.ucsd.edu.

References

- Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–2410.
- Brown JM, Campbell JP, Beers A, et al. For the imaging and informatics in retinopathy of prematurity (i-ROP) research consortium. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol*. 2018;136(7):803–810.
- Baxter SL, Klie AR, Radha SB, et al. Text processing for detection of fungal ocular involvement in critical care patients: cross-sectional study. *J Med Internet Res*. 2020;22(8):e18855.
- Applications of natural language processing in ophthalmology: present and future - PMC. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9393550/>. Accessed January 1, 2024.
- Chen JS, Coyner AS, Chan RVP, et al. Deepfakes in ophthalmology: applications and realism of synthetic retinal images from generative adversarial networks. *Ophthalmol Sci*. 2021;1(4):100079.
- Bellemo V, Burlina P, Yong L, et al. Generative adversarial networks (GANs) for retinal fundus image synthesis. In: Carneiro G, You S, eds. *Computer Vision – ACCV 2018 Workshops*. Berlin, Germany: Springer International Publishing; 2019:289–302.
- What's the best chatbot for me? Researchers put LLMs through their paces. *Nature*. <https://www.nature.com/articles/d41586-023-03023-4>. Accessed January 1, 2024.
- 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.
- Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of Large Language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312.
- Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum - PMC. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10086829/>. Accessed January 1, 2024.
- Kuşçu O, Pamuk AE, Sütay Süslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? *Front Oncol*. 2023;13:1256459.
- Çağlar U, Yıldız O, Meric A, et al. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. *J Pediatr Urol*. 2024;20:26.e1–26.e5.
- Nguyen J, Pepping CA. The application of ChatGPT in healthcare progress notes: a commentary from a clinical and research perspective. *Clin Transl Med*. 2023;13(7):e1324.
- Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health*. 2023;5(3):e107–e108.
- Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. *Am J Cancer Res*. 2023;13(4):1148–1154.
- 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.
- Bernstein IA, Zhang Y, Govil D, et al. Comparison of ophthalmologist and Large Language model chatbot responses to online patient eye care questions. *JAMA Netw Open*. 2023;6(8):e2330320.
- Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res*. 2023;25:e48659.

19. Antaki F, Touma S, Milad D, et al. Evaluating the performance of ChatGPT in ophthalmology. *Ophthalmol Sci.* 2023;3(4):100324.
20. Cai LZ, Shaheen A, Jin A, et al. Performance of generative Large Language models on ophthalmology board—style questions. *Am J Ophthalmol.* 2023;254:141–149.
21. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol.* 2023;141(6):589–597.
22. Momenai B, Wakabayashi T, Shahlaee A, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina.* 2023;7:862–868.
23. Delsoz M, Raja H, Madadi Y, et al. The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. *Ophthalmol Ther.* 2023;12(6):3121–3132.
24. Cox A, Seth I, Xie Y, et al. Utilizing ChatGPT-4 for providing medical information on blepharoplasties to patients. *Aesthet Surg J.* 2023;43(8):NP658–NP662.
25. Al-Sharif E, Penteador R, Dib El Jalbout N, et al. Evaluating the accuracy of ChatGPT and Google BARD in fielding oculoplastic patient queries: a comparative study on artificial versus human intelligence. *Ophthalmic Plast Reconstr Surg.* 2023;40:303–311.
26. Biswas S, Logan NS, Davies LN, et al. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. *Ophthalmic Physiol Opt.* 2023;43:1562–1570.
27. ChatGPT and Ophthalmology: Exploring its potential with discharge summaries and operative notes: seminars in ophthalmology: Vol 38, No 1. <https://www.tandfonline.com/doi/abs/10.1080/08820538.2023.2209166?journalCode=isio20>; 2023. Accessed August 4, 2023.
28. Hua HU, Kaakour AH, Rachitskaya A, et al. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. *JAMA Ophthalmol.* 2023;141:819–824.
29. Hirose T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case—derived clinical vignettes: diagnostic accuracy evaluation. *JMIR Med Inform.* 2023;11:e48808.
30. Ophthalmic Case Studies | Ophthalmology and Visual Sciences | Medical College of Wisconsin. <https://www.mcw.edu/departments/ophthalmology-eye-institute/education/ophthalmic-case-studies>. Accessed January 1, 2024.
31. Kincaid J, Fishburne R, Rogers R, Chissom B. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch reading Ease Formula) for navy enlisted personnel. *Inst Simul Train*; 1975. <https://stars.library.ucf.edu/istlibrary/56>. Accessed February 28, 2024.
32. Gupta A, Quinn M, Saint S, et al. The variability in how physicians think: a case-based diagnostic simulation exercise. *Diagn Berl Ger.* 2021;8(2):167–175.
33. Desaire H, Chua AE, Isom M, et al. Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Rep Phys Sci.* 2023;4(6):101426.
34. JMIR Medical Education - Differentiating ChatGPT-Generated and Human-Written Medical Texts: Quantitative Study. <https://mededu.jmir.org/2023/e48904/>. Accessed January 2, 2024.
35. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res.* 2023;25:e48568.
36. Ting DSJ, Tan TF, Ting DSW. ChatGPT in ophthalmology: the dawn of a new era? *Eye.* 2023;38:1–4.
37. Tan TF, Thirunavukarasu AJ, Campbell JP, et al. Generative artificial intelligence through ChatGPT and other Large Language models in ophthalmology: clinical applications and challenges. *Ophthalmol Sci.* 2023;3(4):100394.
38. Baxter SL, Gali HE, Chiang MF, et al. Promoting quality face-to-face communication during ophthalmology encounters in the electronic health record era. *Appl Clin Inform.* 2020;11(1):130–141.
39. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. *Ann Biomed Eng.* 2023;51(12):2629–2633.
40. Kleinig O, Gao C, Kovoov JG, et al. How to use large language models in ophthalmology: from prompt engineering to protecting confidentiality. *Eye.* 2023;38:649–653.
41. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res.* 2023;25:e50638.
42. Salvagno M, Taccone FS, Gerli AG. Artificial intelligence hallucinations. *Crit Care.* 2023;27(1):180.
43. Beutel G, Geerits E, Kielstein JT. Artificial hallucination: GPT on LSD? *Crit Care.* 2023;27:148.
44. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. *ACM Comput Surv.* 2023;55(12), 248:1–248:38.
45. Moran C. ChatGPT is making up fake Guardian articles. Here's how we're responding. The Guardian. <https://www.theguardian.com/commentisfree/2023/apr/06/ai-chatgpt-guardian-technology-risks-fake-article>; 2023. Accessed January 2, 2024.
46. Hsu T, Thompson SA. Disinformation researchers raise alarms about A.I. Chatbots. The New York Times. <https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html>; 2023. Accessed January 2, 2024.