# ChatGPT in Ophthalmology: Past, Present, and Future

Amit Toledano MD[1,3], Ehud Raz Gatt MD[2,3], Asaf Laks MD[1,3], Biana Dubinsky-Pertzov MD MPH[1,3], Adi Einan-Lifshitz MD[1,3], Eran Pras MD[1,3], and Asaf Shemer MD[1,3]

[1]Department of Ophthalmology, Shamir Medical Center (Assaf Harofeh), Zerifin, Israel
[2]Wolfson Medical Center, Holon, Israel
[3]Gray Faculty of Medical and Health Sciences, Tel Aviv University, Tel Aviv, Israel

**ABSTRACT** **Background:** The rapid evolution of large language models warrants updated benchmarking in ophthalmology to determine whether newer versions offer clinically meaningful improvements over earlier models and human comparators.
**Objectives:** To evaluate the diagnostic accuracy of ChatGPT-4o and ChatGPT-5 in ophthalmic cases and to compare it with previously reported results of ChatGPT-3.5, residents, and specialists.
**Methods:** This retrospective cohort study was conducted in one academic tertiary medical center. We reviewed data of patients admitted to the ophthalmology department from June 2022 to January 2023. We then created two clinical cases for each patient. The first was according to medical history alone (Hx). The second added the clinical examination (Hx and Ex). For each case, we asked for the three most likely diagnoses from ChatGPT-4o and ChatGPT-5. We then compared the accuracy rates (at least one correct diagnosis) with previous results of ChatGPT-3.5, residents, and specialists.
**Results:** A total of 63 cases were analyzed, first using history alone and then with examination findings. Based on history alone, GPT-5 and GPT-4o correctly identified 73% and 70% of cases, respectively, outperforming GPT-3.5 (54%, $P < 0.05$) and approaching the accuracy of residents (75%) and attending physicians (71%, $P < 0.05$). When physical examination was included, diagnostic accuracy rose to 94% for GPT-5 and 89% for GPT-4o, surpassing GPT-3.5 (68%, $P < 0.05$) and closely matching or exceeding human performance (residents 94%, attendings 87%).
**Conclusions:** ChatGPT-4o and ChatGPT-5 significantly outperformed GPT-3.5 and achieved diagnostic accuracy similar or even higher to clinicians in diagnosing ophthalmology cases.

*IMAJ* 2026; 28: 232–236

**KEY WORDS:** artificial intelligence (AI), ChatGPT-5, ChatGPT-4, diagnostic accuracy, ophthalmology

The use of artificial intelligence (AI) is advancing quickly, and large language models (LLMs) such as Generative Pre-trained Transformer (ChatGPT) are becoming increasingly relevant. In a previous study [1], we evaluated the diagnostic performance of ChatGPT-3.5 in ophthalmology and compared it to the results from residents and attending physicians, which showed overall inferiority to both, especially in more complex patient scenarios [1].

ChatGPT is trained to provide human-like answers to questions or tasks given by the user. Since the release of ChatGPT-4 in March 2023, various studies have shown improved reasoning and accuracy [2]. The relatively short time between the release of GPT-3.5 and GPT-4o, and recently GPT-5, highlights the pace at which this technology is evolving.

Several studies have raised concerns regarding the usage in ChatGPT and its integration into clinical practice, noting ethical, data security and hallucinations risk [3,4]. The current trend suggests that ChatGPT can be used as a support tool for education, research, and clinical assistance, but not as a replacement for physicians due to the risk of dehumanization in care [5].

On 7 August 2025, OpenAI released ChatGPT-5. This model was reported to have stronger performance and fewer hallucinations [6]. To the best of our knowledge, there are no published articles that discuss the diagnostic capabilities of ChatGPT-5 in medicine, and specifically in the field of ophthalmology.

## PATIENTS AND METHODS

### STUDY DESIGN AND DATA SOURCE

A retrospective cohort study included all adult patients (age > 18 years) admitted to the ophthalmology department from June 2022 to January 2023. Only admissions of ≥ 3 days were included.

### CLINICAL SCENARIOS

Each case was divided into a historic component (age,

co-morbidities, medications, chief complaint) and an examination component (full ophthalmic examination findings). Both sections were created according to the medical records. Patient information was presented in bulk for the history part and clinical examination findings were presented after. The dataset was identical to our prior research assessing ChatGPT-3.5 diagnostic capabilities [1].

### CASE PRESENTATION AND PROMPTING

Each case was introduced to GPT-4o and GPT-5 in two stages. In the first stage, only the patient's history (Hx) was provided and the model was asked to give the three most likely diagnoses. In the second stage, the same case was re-entered with the addition of the full ophthalmic examination (Hx + Ex), and the model was asked again to provide the three most likely diagnoses. The same two-step process was applied for GPT-4o and 5.

### OUTCOME MEASURES

The primary outcome was the accuracy of diagnosis, defined as the inclusion of a correct diagnosis (as determined by discharge summary) among the three diagnoses listed by ChatGPT (GPT-4o and GPT-5). These results were then compared to ChatGPT-3.5, the residents' diagnoses, and the attendings' diagnoses.

Secondary outcomes included comparing the primary diagnosis (the first diagnosis that ChatGPT provided) and further analyses of each subspecialty and rare cases.

### STATISTICAL ANALYSIS

Categorical variables (correct vs. incorrect responses) were summarized as frequencies and percentages, and 95% confidence intervals (95%CI) were calculated. Overall differences in accuracy rates among the five groups were assessed using the chi-square test of independence. Pairwise comparisons between groups were performed using chi-square tests for 2 × 2 contingency tables. All tests were two-tailed, and a *P*-value of $< 0.05$ was considered statistically significant. Statistical analyses were performed using IBM Statistical Package for the Social Sciences statistics software, version 23 (SPSS, IBM Corp, Armonk, NY, USA).

### ETHICS APPROVAL

This study was approved by the institutional research committee and was performed in accordance with the ethical standards of the 1964 Helsinki declaration and its later amendments.

## RESULTS

A total of 63 cases were included in the final analysis. The patient population had a mean age of 51.2 ± 17.7 years, with a slight male predominance (54%). Diagnoses represented the full spectrum of ophthalmology, with retinal (33.3%), corneal (25%), and neuro-ophthalmologic (16%) conditions being most common.

When presented with prompts based on patient history alone, GPT-5 correctly identified 73% of the cases (46 of 63) and GPT-4o correctly identified 70% (44 of 63), outperforming GPT-3.5 (54%, 34 cases; $P < 0.05$) and approaching the accuracy of residents (75%) and attending physicians (71%, $P > 0.05$) [Table 1, Figure 1]. From the correctly diagnosed cases, the primary suggestion was the correct one (the first of three suggestions) in 74% (34 of 46 cases) for GPT-5 and 75% (33 of 44 cases) for GPT-4o, compared with 50% in GPT-3.5 (17 of 34 cases). Overall, GPT-5's primary diagnosis suggestion was correct in 54% of cases, almost like GPT-4o (52%, 33 of 63 cases), and a marked improvement over GPT-3.5 (27%; $P < 0.05$), closely approaching the attending physicians' performance (55%) and exceeding that of the residents (42%) [Figure 2].

With the addition of physical examination findings, GPT-5's diagnostic accuracy rose to 94% (59 of 63) and GPT-4o approached 89% (56 of 63), surpassing GPT-3.5 (68%, 43 of 63 cases) and reaching similar results of the residents and attending physicians (94% and 87%, respectively; $P < 0.05$) [Table 1, Figure 1]. Among correctly diagnosed cases, GPT-5's primary suggestion aligned with the correct diagnosis in 90% of cases (53 of 59), while GPT-4o succeeded in its primary suggestion in 98% of correctly diagnosed cases (55 of 56). In comparison, GPT-3.5's correct diagnosis was the primary suggestion in only 53%. Overall, the primary sugges-

**Table 1.** Diagnosis accuracy rates: ChatGPT-5, ChatGPT-4o, ChatGPT-3.5, residents, and attendings

| | ChatGPT-3.5 | ChatGPT-4o | ChatGPT-5 | Residents | Attendings |
|---|---|---|---|---|---|
| Patient history alone | 34 (54%) | 44 (70%) | 46 (73%) | 47 (75%) | 45 (71%) |
| Patient history and clinical examination | 43 (68%) | 56 (89%) | 59 (94%) | 59 (94%) | 55 (87%) |

tion was the correct diagnosis based on both history and physical examination in 84% for GPT-5 and 87% for GPT-4o. This correct conclusion was substantially higher than that of GPT-3.5 (37%) and even exceeded the average rates observed among residents (79%) and attending physicians (77%) [Figure 2].

**Figure 1.** Diagnostic accuracy rates: ChatGPT-5, ChatGPT-4o, ChatGPT-3.5, residents, and attending physicians

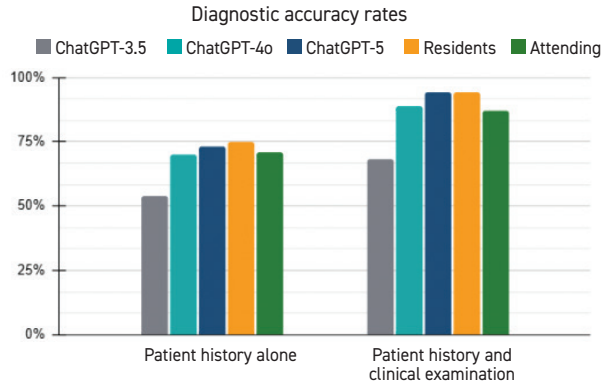
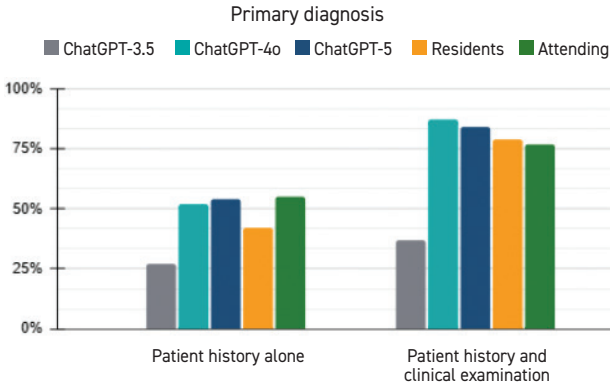



**Figure 2.** Primary diagnosis accuracy based on history alone and based on history and physical examination




Similar to the earlier model, GPT-5 and GPT-4o both demonstrated their highest diagnostic accuracy for retinal cases (95–100% in GPT-5 and 81–100% in GPT-4o), corneal cases (88–100% in GPT-5 and 81–88% in GPT-4o), and neuro-ophthalmology cases (70–100% in both GPT-5 and GPT-4o). Across nearly all subspecialties, both models outperformed GPT-3.5 based on both history alone and combined history and physical examination [Figure 3]. Although diagnostic accuracy for glaucoma cases remained relatively lower based on patient history alone in the newer versions (0% in GPT-5 and 40% in GPT-4o vs. 20% in GPT-3.5), a significant improvement was observed with physical examination inclusion in the last two versions (80% vs. 40% in GPT-3.5).

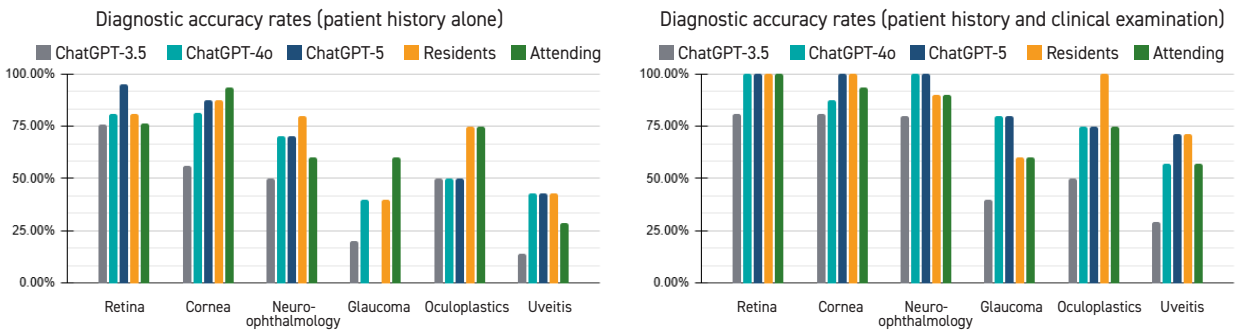
GPT-5 and GPT-4o also demonstrated greater recognition of uncommon ophthalmic conditions. GPT-4o identified non-arteritic anterior ischemic optic neuropathy (NAION) in 100% of cases and GPT-5 in 50–100% of cases (vs. 0–50% for GPT-3.5) and improved performance in Vogt-Koyanagi-Harada disease (33–66% in both GPT-4o and GPT-5 vs. 0–33% in GPT-3.5).

Only a few misclassifications by GPT-4o and GPT-5 occurred with history-only input while GPT-3.5 provided the correct diagnosis. All were resolved when full examination data were included.

## DISCUSSION

LLMs have demonstrated strong diagnostic performance on medical examinations and clinical vignettes, complementing AI's proven effectiveness in imaging

**Figure 3.** Diagnostic accuracy rates stratified by subspecialty based on history alone and based on both history and physical examination: ChatGPT-5, ChatGPT-4o, ChatGPT-3.5, residents, and attending physicians



analysis [7,8]. While prior studies have explored earlier LLMs such as GPT-3.5, most have not evaluated GPT-4o [1,9-13] and, to the best of our knowledge, none have evaluated the latest model, GPT-5. In our previous study [1] GPT-3.5 demonstrated a relatively low diagnostic accuracy, highlighting the limitations of earlier models of ChatGPT. Therefore, we evaluated GPT-4o and ChatGPT-5, the most current advanced versions of ChatGPT.

Our primary analysis demonstrated that both GPT-5 and GPT-4o achieved statistically significant improvement in diagnostic performance compared to GPT-3.5, particularly when provided with additional details of the clinical examination.

Furthermore, GPT-4o and GPT-5 outperformed GPT-3.5 in complex and rare conditions such as NAION and Vogt-Koyanagi-Harada disease, highlighting their enhanced ability to recognize uncommon ophthalmic entities.

Previous studies have shown that earlier models, such as GPT-3.5, consistently underperformed compared to human respondents in ophthalmology-related assessments. For example, GPT-3.5 lagged while GPT-4 achieved performance on par with or even surpassed human averages across multiple ophthalmology subspecialties [11-13]. These findings align with our current results, in which GPT-4o approached or matched human-level performance, even in complex clinical scenarios.

A notable finding in our study was GPT-4o's performance in uveitis cases, which was the lowest among all ophthalmic subspecialties assessed. This trend was similarly observed in a previous study [2] and may point to a potential bias or limitation in the current training data with LLMs. Notably, GPT-5 correctly diagnosed 71% of uveitic cases when provided with both patient history and physical examination findings, suggesting a potential improvement in LLM performance for complex inflammatory conditions.

Another interesting finding was that overall diagnosis accuracy rate for residents was slightly higher than that of the attendings, which could be explained by the fact that residents may be actively preparing for board examinations or other assessments, thus encouraging them to more easily present possible differential diagnoses for a given case. Other studies have also shown that younger physicians may generate broader differential lists and can, at times, outperform senior clinicians on vignette-based diagnostic tasks [14,15]. However, ultimately, the differences between the two groups were non-significant.

Newer LLMs, including GPT4-o, surpass earlier models through genuine multimodal processing [16]. Although a recent review highlighted the versatility of GPT-4o across the clinical workflow [17], some LLMs still show inconsistency in diagnostic reliability and heavily depend on input structure and prompt design [3].

So far, several studies have reviewed the performance of ChatGPT, mostly using question banks or cases treated in an outpatient setting [2,9,12]. One noteworthy strength of our study is the focus on hospitalized cases, which are typically more complex compared to some question bank cases or outpatient cases.

Our study has several limitations. First, although LLMs can leverage imaging analysis [18], our design was text-only and excluded ocular imaging, which may have underrepresented their full diagnostic capacity. Second, the residents and attending cohorts were relatively small. A larger sample could better detect subtle intergroup differences. Third, a single physician obtained the history and examination for all cases, thus clinicians (and models) based their diagnoses on documentation not generated by themselves, which may deviate from routine practice and influence diagnostic accuracy.

## CONCLUSIONS

Both ChatGPT-4o and ChatGPT-5 markedly outperformed GPT-3.5 in solving complex ophthalmology cases. These results reinforce the view that LLMs can serve as meaningful diagnostic adjuncts, provided their domain-specific limitations are recognized. Looking ahead, future work should incorporate ocular imaging into the text inputs, test model-led adaptive history taking, and deliberately probe domains flagged as weak to uncover potential blind spots.

**Correspondence**

**Dr. A. Toledano**
Dept. of Ophthalmology, Shamir Medical Center (Assaf Harofeh), Zerifin 70300, Israel
**Phone:** (972-8) 977-9620
**Fax:** (972-8) 977-9627
**Email:** dr.toledanoamit@gmail.com

**References**

1. Shemer A, Cohen M, Altarescu A, et al. Diagnostic capabilities of ChatGPT in ophthalmology. *Graefes Arch Clin Exp Ophthalmol* 2024; 262 (7): 2345-52.

2. Shvartz E, Zur O, Attal L, Nujeidat Z, Plopsky G, Bahir D. GPT-4o vs. residency exams in ophthalmology–a performance analysis. *J Med Artif Intell* 2025; 8: 48.
3. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* 2024; 30: 2613-22.
4. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. *J Med Internet Res* 2023; 25: e48009.
5. Dossantos J, An J, Javan R. Eyes on AI: ChatGPT's transformative potential impact on ophthalmology. *Cureus* 2023; 15: e40765.
6. OpenAI. Introducing ChatGPT-5. Open AI 2025. [Available from https://openai.com/index/introducing-gpt-5/?utm_source=chatgpt.com]. [accessed 7 September 2025].
7. Hashemian H, Peto T, Ambrósio R, et al. Application of artificial intelligence in ophthalmology: an updated comprehensive review. *J Ophthalmic Vis Res* 2024; 19: 354-67.
8. Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Intern Med* 2024; 184: 581-3.
9. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci* 2023; 3 (4): 100324.
10. Chen JS, Reddy AJ, Al-Sharif E, et al. Analysis of ChatGPT responses to ophthalmic cases: can ChatGPT think like an ophthalmologist? *Ophthalmol Sci* 2024; 5 (1): 100600.
11. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a large language model's responses to questions and cases about glaucoma and retina management. *JAMA Ophthalmol* 2024; 142: 371-5.
12. Taloni A, Borselli M, Scarsi V, 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* 2023; 13: 18562.
13. Cai LZ, Shaheen A, Jin A, et al. Performance of generative large language models on ophthalmology board-style questions. *Am J Ophthalmol* 2023; 254: 141-9.
14. St-Onge C, Landry M, Xhignesse M, et al. Age-related decline and diagnostic performance of more and less prevalent clinical cases. *Adv Health Sci Educ Theory Pract* 2016; 21 (3): 561-70.
15. Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. *JAMA Netw Open* 2019; 2: e190096.
16. Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv preprint [Available from https://arxiv.org/pdf/2303.08774].
17. Li H, Fu J-F, Python A. Implementing large language models in health care: clinician-focused review with interactive guideline. *J Med Internet Res* 2025; 27: e71916.
18. Ting DSW, Cheung CY-L, Lim G, et al. Development and Validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017; 318: 2211.

**Capsule**

## Noxious neutrophils and autoimmunity

Neuromyelitis optica spectrum disorder is an autoimmune disease of the central nervous system driven by B cells and antibodies specific to the channel protein aquaporin-4. **Li** et al. found that bone marrow from individuals with this disease exhibited an aberrant shift toward granulopoiesis, a feature that was also observed in individuals who experienced a relapse after treatment. This increase in granulopoiesis was associated with enrichment of a type of neutrophils in the blood. In an early-phase clinical trial, the authors found that neutralization of B cell–activating factor with belimumab resulted in reduced disease severity. Together, these findings implicate neutrophils in the pathogenesis of neuromyelitis optica spectrum disorder and highlight a therapeutic target.

*Sci Transl Med* 2026; 18 (840): eaeb4775
Eitan Israeli

**Capsule**

## Intercepting tumor development

Pancreatic ductal adenocarcinoma (PDAC) is the most common form of pancreatic cancer and is frequently driven by mutations in the *KRAS* gene. Pancreatic intraepithelial neoplasias are precancerous precursor lesions of PDAC that typically harbor the *KRAS* activating mutation. **Than** et al. tested the effects of small-molecule *KRAS* inhibitor drugs as a chemopreventative strategy for pancreatic cancer. Using a well-studied model of pancreatic cancer, mice with pancreatic intraepithelial neoplasia lesions were treated with *KRAS* inhibitors before tumor formation. A series of different dosing regimens were tested, and early intervention was found to extend survival. Three-dimensional mapping of the tumor microenvironment suggested no major distortions in tissue architecture after treatment. The design of clinical trials to determine whether similar effects are observed in humans will be important next steps.

*Science* 2026; 391: 1161
Eitan Israeli