Türkiye Klinikleri Oftalmoloji Dergisi
Turkiye Klinikleri Journal of Ophthalmology

ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA

DOI: 10.5336/ophthal.2025-110083

Assessment of the Performance of ChatGPT-4o, Claude 3.5 Sonnet, and Google Gemini 2.0 Flash on Image-Based Ocular Oncology and Pathology Questions: A Cross Sectional Research

ChatGPT-4o, Claude 3.5 Sonnet ve Google Gemini 2.0 Flash'in Görüntü Tabanlı Oküler Onkoloji ve Patoloji Sorularındaki Performansının Değerlendirilmesi: Kesitsel Araştırma

Süleyman DEMİR[a], İsmail Cem TÜRKEŞ[b]

[a]Adana 5 Ocak State Hospital, Clinic of Ophthalmology, Adana, Türkiye
[b]Gülhane Training and Research Hospital, Clinic of Ophthalmology, Ankara, Türkiye

ABSTRACT Objective: This study aimed to evaluate the diagnostic performance of 3 flagship models from 3 different companies, Chat Generative Pre-trained Transformer-4 Omni (ChatGPT-4o), Claude 3.5 Sonnet, and Gemini 2.0 Flash, on image-based questions in ocular oncology and pathology to investigate potential differences between these models, and their clinical utility. **Material and Methods:** Fifty multiple-choice, image-based questions were randomly selected from 312 questions in the field of ocular oncology and pathology from the OphthoQuestions (www.ophthoquestions.com) database. ChatGPT-4o, Claude 3.5 Sonnet and Gemini 2.0 Flash models, which have the ability to process images in large language models (LLMs), were included in the study. Cochran's Q test was applied to compare the performance of the 3 LLMs and McNemar's test was used in pairwise comparisons. **Results:** There was a statistically significant difference between all 3 LLMs (p=0.001, Cochran's Q test). Claude 3.5 sonnet showed the highest accuracy by correctly identifying 84% of the questions, followed by ChatGPT-4o with 80% and Gemini 2.0 with 62%. In the pairwise comparisons, Claude 3.5 sonnet and ChatGPT-4o were found to be statistically superior to Gemini 2.0 Flash model (p=0.002, p=0.004, respectively). There was no significant difference between Claude 3.5 and ChatGPT-4o (p=0.727, McNemar test). **Conclusion:** Our results indicate that Claude 3.5 Sonnet and GPT-4o outperform Gemini 2.0 Flash in diagnostic accuracy for ocular oncology and pathology. While LLMs show promise in this field, they require evaluation with larger datasets, and their accuracy must be improved before they can be clinically implemented

Keywords: ChatGPT-4o; Claude-3.5 Sonnet; Gemini 2.0 Flash; ocular oncology; image processing

ÖZET Amaç: Bu çalışmanın amacı, 3 farklı şirketten 3 önde gelen büyük dil modeli Chat Generative Pre-trained Transformer-4 Omni (ChatGPT-4o), Claude-3.5 Sonnet ve Gemini 2.0 Flash'ın oküler onkoloji ve patoloji alanındaki görüntü tabanlı sorularda tanısal performansını değerlendirmek ve bu modeller arasındaki potansiyel farkları ile pratik kullanım için uygunluklarını araştırmaktır. **Gereç ve Yöntemler:** OphthoQuestions (www.ophthoquestions.com) veritabanından oküler onkoloji ve patoloji alanında yer alan 312 sorudan rastgele seçilen 50 çoktan seçmeli, görüntü tabanlı soru kullanılmıştır. Soruların yanıtları, doğru yanıt anahtarıyla karşılaştırılmış ve doğru ya da yanlış olarak kaydedilmiştir. Çalışmaya, güncel ve görüntü işleme özelliği olan büyük dil modelleri [large language models (LLMs)] ChatGPT-4o, Claude 3.5 Sonnet ve Gemini 2.0 Flash dâhil edilmiştir. 3 LLMs modellerinin performansını karşılaştırmak için Cochran's Q testi, 2'li karşılaştırmalar için ise McNemar testi kullanılmıştır. **Bulgular:** Üç LLMs arasında istatistiksel olarak anlamlı bir fark bulunmuştur (p=0,001, Cochran's Q testi). Claude 3.5 Sonnet, soruların %84'ünü doğru cevaplayarak en yüksek doğruluk oranını göstermiş, bunu %80 doğrulukla ChatGPT-4o ve %62 doğrulukla Gemini 2.0 Flash takip etmiştir. İkili karşılaştırmalarda, Claude 3.5 Sonnet ve ChatGPT-4o, Gemini 2.0 Flash modeline karşı istatistiksel olarak üstün bulunmuştur (sırasıyla p=0,002, p=0,004). Claude 3.5 Sonnet ile ChatGPT-4o arasında ise anlamlı bir fark bulunmamıştır (p=0,727, McNemar test). **Sonuç:** Sonuçlarımız, Claude 3.5 Sonnet ve GPT-4o'nun oküler onkoloji ve patoloji alanında tanısal doğruluk açısından Gemini 2.0 Flash'ı geride bıraktığını göstermektedir. LLMs bu alanda umut vaat etse de, daha büyük veri setleriyle eğitilmesi ve doğruluklarının klinik kullanım için iyileştirilmesi gerekmektedir.

Anahtar Kelimeler: ChatGPT-4o; Claude 3.5 Sonnet; Gemini 2.0 Flash; oküler onkoloji; görüntü işleme

Correspondence: Süleyman DEMİR
Adana 5 Ocak State Hospital, Clinic of Ophthalmology, Adana, Türkiye
E-mail: ssuleyman810@gmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Ophthalmology.

Received: 01 Mar 2025 ***Received in revised form:*** 09 May 2025 ***Accepted:*** 11 May 2025 ***Available online:*** 03 Jul 2025

Ocular pathology plays a crucial role in diagnosis and treatment, as a vital bridge between laboratory and clinical practice.[1] Ocular pathology covers a range of diseases, including benign and malignant neoplasms, inflammatory diseases and degenerative disorders. This field also involves the analysis of tissue samples, which facilitates a deeper understanding of disease mechanisms and is necessary for proper diagnosis and treatment planning. Ocular pathology also has an important role in monitoring treatment responses and evaluating disease progression, while in the field of ocular oncology, the evaluation of excised tumor specimens is critical in determining prognosis and treatment.[2,3]

Large language models (LLMs) have revolutionized artificial intelligence (AI) due to their logical reasoning and natural language understanding capabilities.[4] Generative AI has been very popular after OpenAI's ChatGPT-3.5 (OpenAl, California, USA) was released in November 2022. The release in 2023 of more complex models such as ChatGPT-4.0, which combines text, speech and image processing, has once again proven the potential of AI.[5,6] Models such as Claude 3.5-Sonnet (Anthropic, USA), Chat Generative Pre-trained Transformer-4 Omni (ChatGPT-4o) and Google Gemini 2.0 Flash (Alphabet Inc, California, USA) demonstrate the rapid development of AI by exhibiting increasingly human-like mental processes.

LLMs have proven to be capable of synthesizing information from a wide variety of sources.[7] Due to their relative novelty, the role of these LLMs in medicine remains unclear. Large deep learning models have recently become a subject of research.[8] It is highly debatable whether such systems can replace the knowledge and experience of expert doctors. Despite the limitations of experience-based medical decision-making that allows doctors to make complex diagnoses, there is a renewed interest in evaluating the performance of LLMs in routine physician examinations.[9]

LLMs can improve ocular pathology and oncology by enhancing diagnosis and treatment strategies.[10] In ocular pathology, LLMs can process both histopathology images and text. This capability aids in achieving accurate and rapid diagnoses. Similarly, in ocular oncology, LLM's effectiveness in analyzing complex data will help the doctor reach an accurate diagnosis and make quick decisions about different diagnoses that will be individual for patients based on their case history.[11,12]

However, the diagnostic potential of the recently launched models ChatGPT-4o, Gemini 2.0 Flash, and Claude 3.5 Sonnet, all featuring image analysis capabilities, has yet to be fully explored. Therefore, this study aims to assess and compare the diagnostic performance of three flagship models from three different companies: ChatGPT-4o, Gemini 2.0 Flash and Claude 3.5 Sonnet. Our aim is to recognize the diagnostic capabilities of LLMs with image processing, to investigate potential differences between these models and their suitability for practical use.

MATERIAL AND METHODS

This article does not include any studies or data with human or animal participants. As the study evaluates the performance of AI models and does not use identifiable human data, ethics committee approval is not required. Therefore, compliance with the Declaration of Helsinki does not apply to this article.

A total of 50 image-based questions related to the field of ocular oncology and pathology were randomly selected from a pool of 4,550 questions available on OphthoQuestions (www.ophthoquestions.com) in January 2025. The site's built-in randomization function was used to ensure unbiased selection and no particular topic or question was favored. The selected questions were related to the field of ocular oncology and pathology. Of the 312 ocular oncology and pathology questions in the question bank, 50 multiple-choice image-based questions were included. Sample questions and answers provided by the 3 LLMs are shown in Figure 1, Figure 2 and Figure 3. Each query was conducted within the same chat session. Each question posed to the LLM was recorded during the same chat.

The LLMs' answers to the questions were categorised as correct or incorrect after comparison with the answer keys. ChatGPT-4o and Google Gemini 2.0 Flash, Claude 3.5- Sonnet models are the most

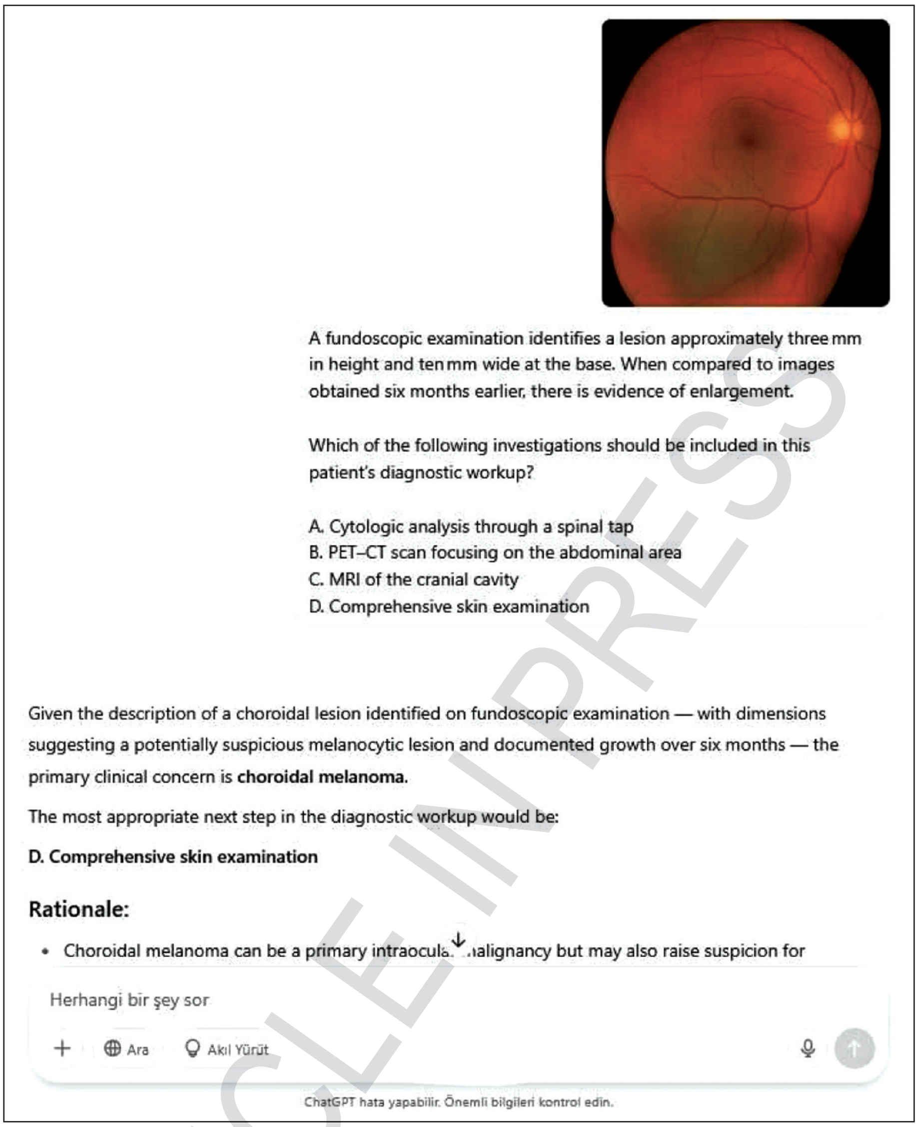

FIGURE 1: A sample of a question and answer to the ChatGPT-4o model

PET-CT: Positron emission tomography-computed tomography; MRI: Magnetic resonance imaging

modern LLMs built by 3 different companies and they can process images as well as text.[13-15]

STATISTICAL ANALYSIS

The data's normality was evaluated using the Kolmogorov-Smirnov test. Percentages were used to represent accuracy and compliance rates, as well as to summarize descriptive data. To determine if there was a difference in categorical values between the groups, Cochran's Q test was utilized. The McNemar test was employed to compare categorical data between the groups. A p value of 0.05 was considered statistically significant. In pairwise comparisons, $p<0.01$ was considered statistically significant after Bonferoni correction. The data was analyzed using SPSS version 25.0 (IBM, New York, USA).

RESULTS

In this study, we evaluated the diagnostic accuracy of 3 LLMs ChatGPT-4o, Claude 3.5 Sonnet and Gemini 2.0 Flash using image-based diagnostic questions. Our results revealed significant performance differences between the models. Our study found that

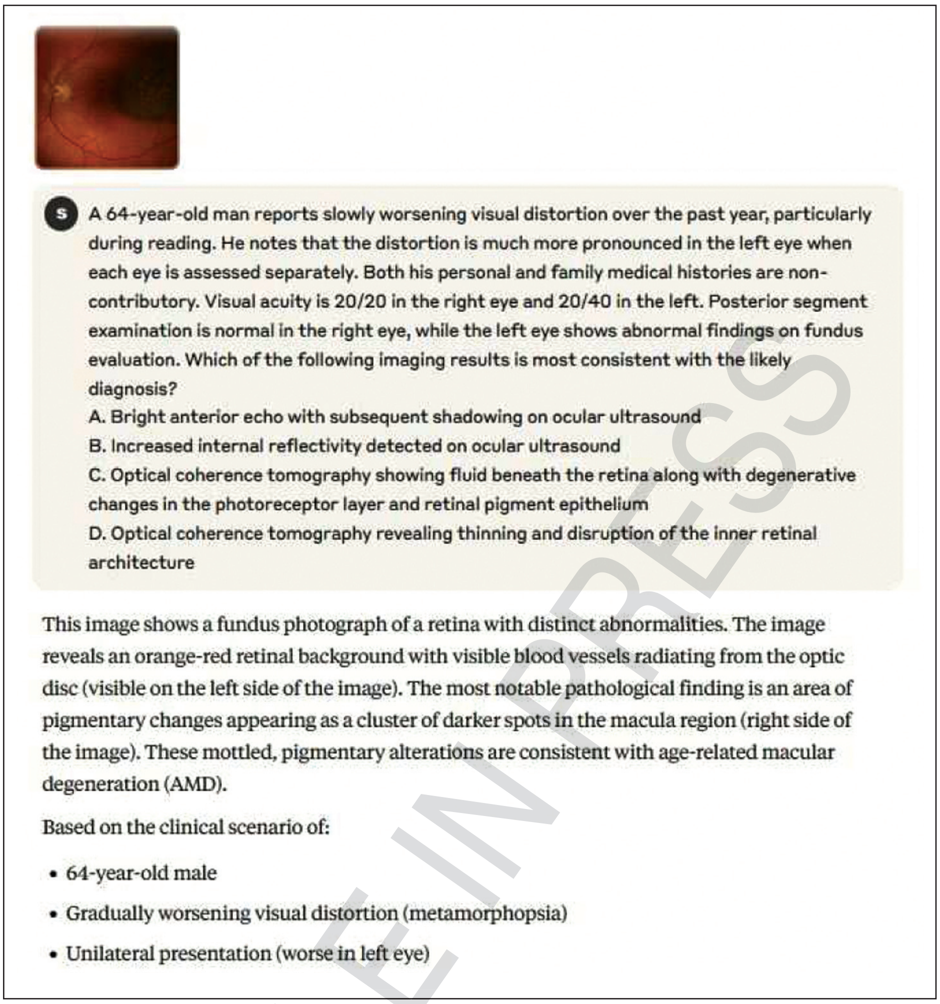

FIGURE 2: A sample of a question and answer to the Claude-3.5 Sonnet model

AMD: Age-related macular degeneration

Claude 3.5 Sonnet correctly answered 42 correct answers (84%) of the 50 image-based case questions in ocular oncology and pathology, followed by ChatGPT-4o with 40 correct answers (80%) and Gemini 2.0 Flash with 31 correct answers (62%). Table 1 presents the correct and incorrect answers and percentages of each 3 LLMs to the questions.

Statistical analysis using Cochran's Q test found statistically significant differences among the 3 models (p=0.001). Pairwise McNemar tests showed statistically significant differences between Claude 3.5 Sonnet and Gemini 2.0 Flash (p=0.002) and between ChatGPT-4o and Gemini 2.0 Flash (p=0.004). These findings indicate that both Claude 3.5 Sonnet and ChatGPT-4o exhibited significantly higher diagnostic accuracy than Gemini 2.0 Flash. The difference in performance between the Claude 3.5 Sonnet and ChatGPT-4o LLMs was not statistically significant (p=0.727, McNemar test). Table 2 presents the statistical analysis of all 3 LLMs and their statistical analysis in pairwise comparisons.

DISCUSSION

This study aims to examine the capabilities, advantages, and limitations of the most recent LLMs, Claude 3.5 Sonnet, ChatGPT-4o, and Gemini 2.0 Flash, in the field of ocular pathology and oncology, particularly in their ability to comprehend image-

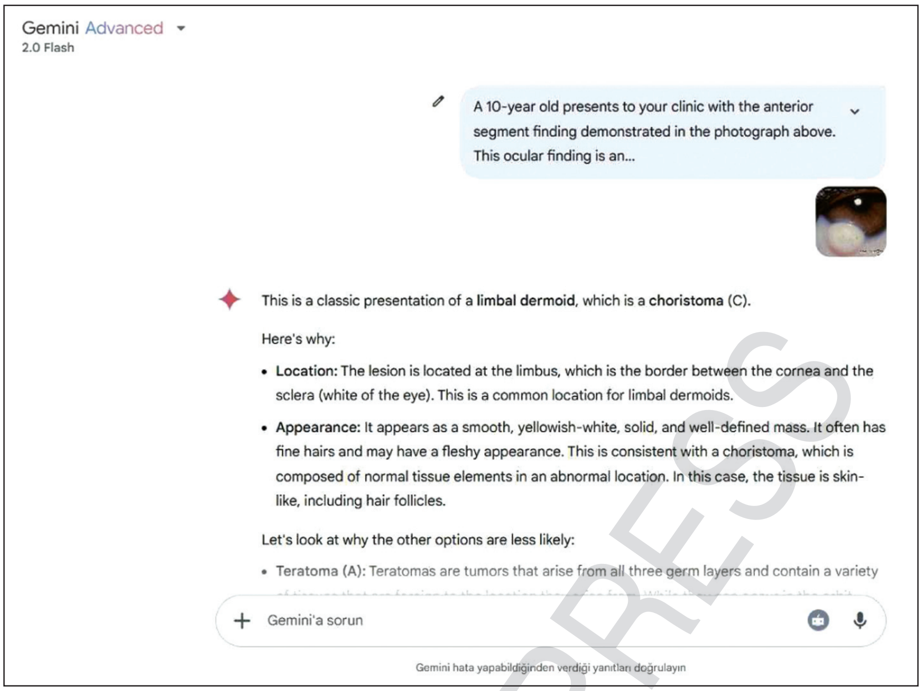

FIGURE 3: A sample of a question and answer to the Google Gemini-2.0 Flash model

TABLE 1: The correct and incorrect answers and percentages of each 3 LLMs to the image-based questions

LLMs	GPT-4o n (%)	Claude 3.5 n (%)	Gemini 2.0 n (%)
Correct diagnosis	40 (80%)	42 (%84)	31 (62%)
Incorrect diagnosis	10 (20%)	8 (16%)	19 (38%)
Image-based total question	50 (100%)	50 (100%)	50 (100%)

LLMs: Large language models; GPT: Generative pre-trained transformer

TABLE 2: The statistical analysis of all 3 LLMs and their statistical analysis in pairwise comparisons

	McNemar's test		
Cochran's Q test	GPT-4o vs Claude 3.5	GPT-4o vs Gemini 2.0	Claude 3.5 vs Gemini 2.0
p=0.001	0.727	**0.004***	**0.002***

*A p value of <0.01 was considered statistically significant.
GPT: Generative pre-trained transformer

based diagnostic inquiries. The results show that although all 3 LLMs show a certain level of diagnostic competence, there are significant performance differences between them. In our study, Claude 3.5 Sonnet achieved the highest diagnostic success rate in image-based questions, answering 42 out of 50 questions correctly (84%), followed by ChatGPT-4o with 40 correct answers (80%) and Gemini 2.0 Flash with 31 correct answers (62%). The statistical analysis revealed that the differences between the models were significant (p=0.001, Cochran's Q test). In pairwise analyses, Claude 3.5 Sonnet and ChatGPT-4o outperformed Gemini 2.0 flash in interpreting image-based questions (p=0.002 and p=0.004, respectively; McNemar test). However, in a pairwise comparison, the difference between Claude 3.5 and ChatGPT-4o was not statistically significant (p=0.727), and the 2 models performed similarly in comprehending image-based questions. When we evaluate the results of this study, the varying efficiency levels of these models raise concerns about the standardization of AI-driven diagnostic tools. In this study, Claude 3.5 Sonnet and ChatGPT-4o performed similarly in ocular oncology and pathology, while Gemini 2.0 flash performed worse than the other 2 LLMs in image-based questions. These findings suggest that LLMs vary in their effectiveness when analyzing medical images. We suggest that this discrepancy may be due to differences in the algorithms of the language models.

In other studies conducted in the literature, a study by Fujimoto et al. in the field of anesthesiology compared ChatGPT-4, Claude 3 Opus, and Gemini 1.0.[15] The results showed that ChatGPT-4 and Claude 3 Opus outperformed Gemini 1.0. In another study, Kim et al. conducted a performance analysis study in which ChatGPT-4 and Claude 3 Opus were used to evaluate the questions of the Dental Licensing Examination, and LLMs such as Claude 3 Opus and ChatGPT-4 achieved 85.4% of human performance on average.[16] Schmidl et al. investigated the use of the newly published LLM Claude 3 Opus in comparison to ChatGPT 4.0 for diagnosing and planning therapy for primary head and neck squamous cell carcinoma, finding that Claude 3 Opus performed better in diagnostic accuracy, while both models provided similar treatment recommendations congruent with multidisciplinary tumor board guidelines.[17] Chen et al. in a study of ChatGPT-4o and Claude 3-Opus classification of thyroid nodules, found that LLMs showed low accuracy and poor agreement with pathological results.[7] In the same study, Chen et al. recommended caution when using LLMs for clinical diagnosis.[7] In a study comparing Claude 3 Opus and ChatGPT-4 for melanoma detection using dermatoscopic images, Liu et al. found that both language models showed similar accuracy, with Claude 3 Opus outperforming ChatGPT-4 in distinguishing malignant tumors ($p<0.001$).[18] In a study conducted by Sensoy et al. comparing 3 LLMs (Bing, ChatGPT, and Bard) using 36 questions related to ophthalmic pathologies and intraocular tumors, no statistically significant difference was found in the accuracy rates of the 3 chatbots ($p=0.705$, Pearson chi-square test).[19]

ChatGPT-4o and Claude 3.5 Sonnet outperformed Google Gemini 2.0 Flash in areas requiring high precision, such as medical image analysis in our study. This difference is due to the image processing capabilities of the models. GPT-4o processes visual, text and audio data in a single model, first compressing the images with a Variational Autoencoder-based encoder and then producing high-quality outputs with a diffusion decoder.[20] The visual processing details of Claude 3.5 Sonnet and Gemini 2.0 Flash are not fully publicly available, but it is known that Claude 3.5 Sonnet uses a ResNet-based visual encoder, which allows it to analyze complex visual content and generate meaningful text. In contrast, the Gemini 2.0 Flash model prioritizes speed and low latency, which can limit the ability for deep contextual analysis and lead to superficial and erroneous results in visual analysis. This probably explains the differences observed in our study.

In ophthalmic oncology and pathology, the implementation of LLMs poses significant challenges regarding patient privacy, fairness, bias, transparency, and the ethical use of AI. The reliability and verifiability of these models, which might influence clinicians' diagnostic decisions, are critical.[21] Overcoming these ethical and technical barriers could transform LLMs into a valuable tool for more accurate diagnosis, optimizing patient care and improving the diagnostic process in ocular tumors and pathological evaluations.[22,23] Furthermore, it is critical to provide a clear explanation of the query instructions issued to LLMs, as their sensitivity to formulation can considerably effect the study's results, reproducibility, and the validity of comparisons, which should not be ignored.[24,25]

In summary, these findings highlight the potential of LLMs in ocular pathology and oncology, but also reveal significant performance differences. For image-based questions in ocular oncology and pathology, Claude 3.5 Sonnet and ChatGPT-4o showed higher diagnostic accuracy, while Gemini 2.0 Flash showed lower diagnostic accuracy. Given the complexity of ophthalmic oncology and pathology, where precise imaging interpretation is crucial, the discrepancies and differences between these LLMs need to be validated by further studies. Although our findings show significant differences between the three LLMs, it should be noted that using only 50 image-based questions may limit the representation of the full spectrum of ocular oncology and pathology cases. This field covers a wide range of lesions, from benign conjunctival tumors to highly aggressive uveal melanomas, many of which have different histopathological and imaging features. In conclusion, it should be noted that our results are limited for the clinical practical application of LLMs in the field of ocular oncology and caution should be taken when adapting these accuracy rates to scenarios in the field of ocular oncology. To address this, future studies should leverage multicenter datasets to provide more robust, clinically relevant results. Larger studies with different disease pictures and hybrid approaches combining AI and human skills should be the main focus of future research in ophthalmology to improve diagnostic accuracy and generalizability.

LIMITATIONS

This study has several limitations. The primary limitation is that the sample size of 50 image-based questions may not fully represent the diversity of ocular oncology and pathology cases. While there are significant differences between the large language models, the study did not evaluate how these models perform with different types of ocular images. The generalizability of this study is limited because the selected cases have precise diagnostic features and do not fully represent areas involving the complexity of ocular oncology, such as uveal melanomas.

Although LLMs have made significant progress in the field of ocular oncology, they still face some serious challenges in image processing. Ocular oncological images often contain complex structures and heterogeneity. The ability of the LLMs to distinguish between the limited or irregular borders of tumors and the surrounding healthy tissue depends on the coverage and quality of the training data. In addition, different techniques, contrasts, noises, and artifacts in the images may also affect the accuracy of the LLMs. In this context, considering that LLMs are usually trained with text data and have limited capacity to interact directly with visual data, advanced image processing algorithms and multimodal learning approaches are needed to accurately analyze such images. As a result, a wider variety of data and more in-depth training are required for the effective use of LLMs in oncology imaging. This makes their ability to address image-based questions in a vital field such as ocular oncology and pathology extremely limited for practical application.

CONCLUSION

To best of our knowledge, this is the first study to assess the image-based diagnostic capabilities of ChatGPT-4o, Claude 3.5 Sonnet and Gemini 2.0 Flash LLMs, which are the latest models in ophthalmology. Our results show that Claude 3.5 Sonnet and GPT-4o outperform Gemini 2.0 Flash in terms of diagnostic accuracy. Although LLMs show potential for ocular oncology, these models need to be tested with larger datasets and their accuracy needs to be improved before they can be used in clinical applications. Future research should prioritize increasing model accuracy, establishing improved interpretability. By addressing these challenges, LLMs can play a transformative role in the field of ocular oncology and pathology and support physicians in making more precise and timely decisions.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

***Idea/Concept:** Süleyman Demir; **Design:** Süleyman Demir; **Control/Supervision:** Süleyman Demir, İsmail Cem Türkeş; **Data Collection and/or Processing:** Süleyman Demir, İsmail Cem Türkeş; **Analysis and/or Interpretation:** Süleyman Demir, İsmail Cem Türkeş; **Literature Review:** Süleyman Demir, İsmail Cem Türkeş; **Writing the Article:** Süleyman Demir, İsmail Cem Türkeş; **Critical Review:** Süleyman Demir, İsmail Cem Türkeş; **References and Fundings:** Süleyman Demir, İsmail Cem Türkeş; **Materials:** Süleyman Demir, İsmail Cem Türkeş.*

REFERENCES

1. Grossniklaus HE. Ophthalmic pathology: history, accomplishments, challenges, and goals. Ophthalmology. 2015;122(8):1539-42. PMID: 26210596.
2. Apple DJ, Werner L, Mamalis N, Olson RJ. The "demise" of diagnostic and research ocular pathology: temporary or forever? Trans Am Ophthalmol Soc. 2003;101:127-37; discussion 137-9. PMID: 14971571; PMCID: PMC1358982.
3. Klintworth GK. Ashton lecture. Ophthalmic pathology from its beginning to the high technology of this millennium. Eye (Lond). 2001;15(Pt 5):569-77. PMID: 11702964.
4. Zhang C, Chen J, Li J, Peng Y, Mao Z. Large language models for human–robot interaction: a review. Biomimetic Intell Robot. 2023;3(4):100131. https://www.sciencedirect.com/science/article/pii/S2667379723000451
5. Kalyan KS. A survey of GPT-3 family large language models including ChatGPT and GPT-4. Natural Language Processing Journal. 2024;6:100048. https://www.sciencedirect.com/science/article/pii/S2949719123000456
6. Demir S. A comparative analysis of GPT-3.5 and GPT-4.0 on a multiple-choice ophthalmology question bank: a study on artificial intelligence developments. Rom J Ophthalmol. 2024;68(4):367-71. PMID: 39936066; PMCID: PMC11809821.
7. Chen Z, Chambara N, Wu C, Lo X, Liu SYW, Gunda ST, et al. Assessing the feasibility of ChatGPT-4o and Claude 3-Opus in thyroid nodule classification based on ultrasound images. Endocrine. 2025;87(3):1041-9. PMID: 39394537; PMCID: PMC11845565.
8. Kumar P. Large language models (LLMs): survey, technical frameworks, and future challenges. Artifcial Intelligence Review. 2024;57(10):260. https://link.springer.com/article/10.1007/s10462-024-10888-y
9. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. Digit Health. 2023;9:20552076231186520. PMID: 37426593; PMCID: PMC10328041.
10. Longwell JB, Hirsch I, Binder F, Gonzalez Conchas GA, Mau D, Jang R, et al. Performance of large language models on medical oncology examination questions. JAMA Netw Open. 2024;7(6):e2417641. PMID: 38888919; PMCID: PMC11185976.
11. Benet D, Pellicer-Valero OJ. Artificial intelligence: the unstoppable revolution in ophthalmology. Surv Ophthalmol. 2022;67(1):252-70. PMID: 33741420; PMCID: PMC9757817.
12. Yaghy A, Yaghy M, Shields JA, Shields CL. Large language models in ophthalmology: potential and pitfalls. Semin Ophthalmol. 2024;39(4):289-93. PMID: 38179986.
13. Bicknell BT, Butler D, Whalen S, Ricks J, Dixon CJ, Clark AB, et al. ChatGPT-4 omni performance in USMLE disciplines and clinical skills: comparative analysis. JMIR Med Educ. 2024;10:e63430. PMID: 39504445; PMCID: PMC11611793.
14. Sonoda Y, Kurokawa R, Nakamura Y, Kanzawa J, Kurokawa M, Ohizumi Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. Japanese Journal of Radiology. 2024;42(11):1231-5. https://link.springer.com/article/10.1007/s11604-024-01619-y
15. Fujimoto M, Kuroda H, Katayama T, Yamaguchi A, Katagiri N, Kagawa K, et al. Evaluating large language models in dental anesthesiology: a comparative analysis of ChatGPT-4, Claude 3 Opus, and Gemini 1.0 on the Japanese Dental Society of anesthesiology board certification exam. Cureus. 2024;16(9):e70302. PMID: 39469383; PMCID: PMC11513210.
16. Kim W, Kim BC, Yeom HG. Performance of large language models on the Korean dental licensing examination: a comparative study. Int Dent J. 2025;75(1):176-84. PMID: 39370338; PMCID: PMC11806296.
17. Schmidl B, Hütten T, Pigorsch S, Stögbauer F, Hoch CC, Hussain T, et al. Assessing the use of the novel tool Claude 3 in comparison to ChatGPT 4.0 as an artificial intelligence tool in the diagnosis and therapy of primary head and neck cancer cases. Eur Arch Otorhinolaryngol. 2024;281(11):6099-109. PMID: 39112556; PMCID: PMC11512878.
18. Liu X, Duan C, Kim MK, Zhang L, Jee E, Maharjan B, et al. Claude 3 Opus and ChatGPT With GPT-4 in dermoscopic image analysis for melanoma diagnosis: comparative performance analysis. JMIR Med Inform. 2024;12:e59273. PMID: 39106482; PMCID: PMC11336503.
19. Sensoy E, Citirik M. A comparative study on the knowledge levels of artificial intelligence programs in diagnosing ophthalmic pathologies and intraocular tumors evaluated their superiority and potential utility. Int Ophthalmol. 2023;43(12):4905-9. PMID: 37880412.
20. Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A, et al. GPT-4o system card. arXiv Preprint. 2024. https://arxiv.org/abs/2410.21276
21. Grote T, Berens P. A paradigm shift?-On the ethics of medical large language models. Bioethics. 202438(5):383-90. PMID: 38523587.
22. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). NPJ Digit Med. 2024;7(1):183. https://www.nature.com/articles/s41746-024-01157-x
23. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology-a recent scoping review. Diagn Pathol. 2024;19(1):43. PMID: 38414074; PMCID: PMC10898121.
24. Demir S. Investigating the role of large language models on questions about refractive surgery. Int J Med Inform. 2025;195:105787. PMID: 39787660.
25. Lin Z. How to write effective prompts for large language models. Nat Hum Behav. 2024;8(4):611-5. PMID: 38438650.