DOI:10.14744/bej.2025.06332
Beyoglu Eye J 2025

Original Article

A Novel Multimodal Large Language Model for Interpreting Image-Based Ophthalmology Case Questions: Comparative Analysis of Multiple-Choice and Open-Ended Response

Pelin Kiyat, Hazan Gul Kahraman

Department of Ophthalmology, İzmir Democracy University, Buca Seyfi Demirsoy Training and Research Hospital, İzmir, Türkiye

Abstract

Objectives: The objective of the study is to evaluate the performance of Claude 3.5 Sonnet, a novel multimodal large language model, in interpreting image-based ophthalmology case questions.

Methods: A total of 174 image-based ophthalmology questions from a comprehensive ophthalmology education platform were analyzed by Claude 3.5 Sonnet. Each question was presented in both multiple-choice and open-ended formats. Questions were categorized into six subspecialties: Retina and uveitis; external eye and cornea; orbit and oculoplastics; neuroophthalmology; glaucoma and cataract; and strabismus, pediatric ophthalmology, and genetics. Performance was evaluated by two board-certified ophthalmologists.

Results: Claude 3.5 Sonnet demonstrated an overall accuracy rate of 89.65% in multiple-choice questions and a comparable 87.93% in open-ended questions, with no statistically significant difference between formats (p=0.72). Performance showed slight variations among subspecialties, with the highest accuracy in external eye and cornea cases (95.65% in both formats) and lower accuracy in strabismus, pediatric ophthalmology, and genetics (87.50% in multiple-choice and 84.38% in open-ended).

Conclusion: Claude 3.5 Sonnet showed strong capabilities in interpreting image-based ophthalmology questions across all subspecialties, with consistent performance between different question formats. These findings suggest potential applications in ophthalmology education and board examination preparation; however, validation of its utility in real-world clinical scenarios needs further evaluation.

Keywords: Artificial intelligence, Claude 3.5 Sonnet, ophthalmology board examinations

Introduction

Artificial intelligence (AI) has begun to play a pivotal role in the field of medicine, with remarkable improvements recently, especially the development of large language models (LLMs) (1,2). LLMs are defined as a type of generative AI that uses conversation-based technology and allows users to receive contextually appropriate textual responses to their questions (3). A recent innovation in LLM technology is the addition of image interpretation capabilities. These multimodal LLMs, also referred to as vision-language models (VLMs), have the potential to lead a new era in medicine by processing and interpreting both visual and textual con-

How to cite this article: Kiyat P, Kahraman HG. A Novel Multimodal Large Language Model for Interpreting Image-Based Ophthalmology Case Questions: Comparative Analysis of Multiple-Choice and Open-Ended Response. Beyoglu Eye J 2025.

Address for correspondence: Pelin Kiyat, MD. Department of Ophthalmology, İzmir Democracy University, Buca Seyfi Demirsoy Training and Research Hospital, İzmir, Türkiye
Phone: +90 536 256 11 12 **E-mail:** pelinkiyat@hotmail.com

Submitted Date: October 29, 2024 **Revised Date:** May 28, 2025 **Accepted Date:** June 23, 2025 **Available Online Date:** October 08, 2025

Beyoglu Eye Training and Research Hospital - Available online at www.beyoglueye.com

tent (4). Claude 3.5 Sonnet (Anthropic, California, United States), a multimodal LLM released in early 2024, has the capability to analyze both textual and image data inputs (5).

In recent years, internationally recognized qualifications such as the European Board of Ophthalmology and the Fellowship of the Royal College of Ophthalmologists (FRCOphth) examinations have gained great popularity among young ophthalmologists, particularly ophthalmology residents, in our country. In preparation for these examinations, candidates frequently use question banks, as these resources closely resemble the format and content of the actual examinations. "Cybersight" is a comprehensive online training and mentorship platform for eye health professionals worldwide, with a particular focus on regions where access to learning resources is limited. Cybersight aims to improve the knowledge, skills, and expertise of eye care professionals globally. The platform offers a robust question bank that includes case-based scenarios with high-quality ophthalmic images across various subspecialties. This resource serves as an effective tool for ophthalmologists preparing for board examinations, providing them with opportunities to enhance their diagnostic and management skills through practical case scenarios (6). While previous studies have investigated the performance of LLMs in text-based ophthalmology board examination practice questions (2,7,8), no study up to date has evaluated the image-based case questions. Given that ophthalmology is a subspecialty heavily reliant on multimodal imaging and visual data interpretation, multimodal LLMs capable of image analysis are gaining increasing significance.

The present study aims to evaluate the performance of the novel multimodal LLM, Claude 3.5 Sonnet, in interpreting image-based ophthalmology case questions. The study utilizes case-based scenarios sourced from the "Cybersight" educational platform, which provides comprehensive coverage across ophthalmology subspecialties.

Methods

Claude 3.5 Sonnet (Anthropic, California, United States), a multimodal LLM released on June 21, 2024, was used to evaluate its performance on image-based case questions. The study utilized visual case questions from "Cybersight," a comprehensive online training and mentorship platform for eye care professionals. A total of 174 image-based questions were selected for this study from the Cybersight question bank.

The questions were categorized into six subspecialties: "Retina and Uveitis" (n=30); "External Eye and Cornea" (n=23); "Orbit and Oculoplastics" (n=28); "Neuroophthalmology" (n=31); "Glaucoma and Cataract" (n=30); and "Strabismus, Pediatric Ophthalmology, and Genetics" (n=32).

While the original questions in Cybersight were presented in multiple-choice format, we conducted the study by presenting each identical question in two different formats: (1) Presenting the complete question with multiple-choice options as originally designed and (2) presenting only the case scenario and images without the answer choices to assess whether Claude 3.5 Sonnet could generate correct open-ended responses. This approach allowed for direct comparison of the model's performance on the same clinical scenarios in both multiple-choice and open-ended formats.

To standardize the input process, all questions were formatted using Microsoft Word, following the methodology described by Gilson et al. (9) For each question, the visual stem and relevant text were combined into a single paragraph. In multiple-choice questions, answer choices were placed on separate lines, with two empty lines inserted between the question stem and the choices. For open-ended evaluation, the same case descriptions and images were presented without the multiple-choice options. The images used in the study were directly obtained from the Cybersight question bank without any modifications. These images represented a comprehensive range of ophthalmological imaging modalities commonly used in clinical practice, including anterior segment photographs, slit-lamp images, fundus photographs, optical coherence tomography (OCT) scans, orbital imaging, and other diagnostic images typically used in clinical practice. Image quality varied but was consistently of diagnostic standard, with sufficient resolution and clarity to allow for identification of key pathological features. The diversity of imaging techniques across the different subspecialties provided an opportunity to evaluate Claude 3.5 Sonnet's performance across the full spectrum of visual data encountered in ophthalmology practice.

A new account was created specifically for this study to eliminate potential bias from previous conversations. The conversation history was cleared, and the chatbot was refreshed before each new question to prevent carryover effects. All question inputs were performed by a single researcher (P.K.) to ensure consistency.

Researchers manually reviewed all answers to evaluate Claude 3.5 Sonnet's performance. The answers provided by Claude 3.5 Sonnet were independently evaluated by two board-certified ophthalmologists. The evaluation was conducted by comparing Claude's responses against the validated answers and explanations provided in our reference source material. Each evaluator assessed the accuracy and clinical appropriateness of the model's responses utilizing the official answer key. Responses were recorded as correct or incorrect based on the official solutions provided by the Cybersight platform. This dual-review process ensured a consistent and objective assessment of the model's performance across all subspecialty domains. The percentage of correct answers was calculated overall and for each subspecialty. Re-

sponses were scored as correct only if they demonstrated accurate identification of the pathology, correct diagnosis, and appropriate management consistent with the reference answers provided by the question bank.

Official permission was obtained from the Cybersight platform to use their questions for this research purpose. As this study did not involve human participants, institutional review board approval was not required.

The primary outcome measure was Claude 3.5 Sonnet's performance in providing correct responses to image-based ophthalmology practice questions. Secondary outcomes included comparisons of performance across the six ophthalmology subspecialties.

Statistical Analysis

IBM the Statistical Package for the Social Sciences version 25 (SPSS Inc., Chicago, IL, USA) was used for statistical purposes. Categorical variables were expressed as frequencies and percentages, and numerical variables were expressed as means and standard deviations. Researchers recorded the answers as correct or incorrect and the percentage of correct answers was calculated overall and for each subspecialty. Kolmogorov–Smirnov tests were used to determine whether the data were normally distributed. Independent t-test was performed to determine the differences in the normality of the distribution or Mann–Whitney U test was performed to determine differences in non-normal distribution. A P-value under 0.05 was considered statistically significant.

Results

The performance of Claude 3.5 Sonnet was evaluated in both multiple-choice and open-ended image-based questions (n=174), with further analysis by subspecialty.

For multiple-choice image-based questions, Claude 3.5 Sonnet demonstrated an 89.65% accuracy rate based on the images. In an open-ended format using the same questions, the model achieved a slightly lower but comparable 87.93% accuracy rate.

The model's performance across different subspecialties is detailed in Table 1, showing both multiple-choice and open-ended results. In both formats, "external eye and cornea" showed the highest accuracy (95.65% in both formats). The lowest performance was observed in "Strabismus and pediatric ophthalmology and genetics" (87.50% in multiple-choice and 84.38% in open-ended).

The difference in performance between multiple-choice and open-ended formats was not statistically significant overall (p=0.72) or within any individual subspecialty (all p>0.05), suggesting that Claude 3.5 Sonnet's diagnostic capabilities remain consistent regardless of question format.

Discussion

The present study evaluated the performance of Claude 3.5 Sonnet, a multimodal LLM, in interpreting image-based ophthalmology practice questions in various subspecialties. We compared its performance in both multiple-choice and open-ended question formats using identical clinical scenarios.

Claude 3.5 Sonnet demonstrated strong performance in interpreting ophthalmic images, with an overall accuracy of 89.65% in multiple-choice format and a comparable 87.93% in open-ended format.

To the best of our knowledge, this study represents the first comprehensive evaluation of Claude 3.5 Sonnet's performance specifically in ophthalmology-related images across all major subspecialties. The model's performance in ophthalmology significantly exceeds its previously reported capabilities in other medical imaging domains. In a recent study by Kurokawa et al., (10) Claude 3.5 Sonnet successfully diagnosed only 30.1% of radiology case questions with key images. In addition, another study reported that Claude 3.5 Sonnet achieved a 59% success rate in diagnosing breast ultrasound images (11).

Although previous studies have evaluated the performance of LLMs on text-based ophthalmology board practice questions, up to date, no study has specifically evaluated

Table 1. Comparison of Claude 3.5 Sonnets' performance between multiple-choice and open-ended formats across ophthalmology subspecialties

Subspecialties	Number of questions	Multiple-choice format Correct/Total (%)	Open-ended format Correct/Total (%)	p
Retina and uveitis	30	26/30 (86.67)	27/30 (90.00)	0.69
External eye and cornea	23	22/23 (95.65)	22/23 (95.65)	1.00
Orbit and oculoplastics	28	25/28 (89.29)	24/28 (85.71)	0.71
Neuroophthalmology	31	28/31 (90.32)	27/31 (87.10)	0.68
Glaucoma and cataract	30	27/30 (90.00)	26/30 (86.67)	0.72
Strabismus and Ped. Oph. and genetics	32	28/32 (87.50)	27/32(84.38)	0.73
Overall	174	156/174 (89.65)	153/174 (87.93)	0.72

LLMs' performance on ophthalmology-related image-based case questions. Previous studies assessing text-based ophthalmology board practice questions have reported that ChatGPT, a popular LLM, achieved success rates ranging from 60% to 80% in various practice question sources (2,7,8). Recently, attempts have been made to evaluate LLMs in interpreting ophthalmological images. A study by Mihalache et al. (12) evaluated LLMs' ability to interpret OCT images. In this study, 448 OCT images were analyzed and their model demonstrated a 65% success rate in correct detection. In another study by Antaki et al., (13) the diagnostic capabilities of the LLMs-Gemini Pro model in interpreting OCT images were evaluated. The research included 50 patients with various retinal pathologies. In that study, the LLMs-Gemini Pro model showed a correct diagnosis rate of 34%.

Claude 3.5 Sonnet showed remarkably consistent performance between multiple-choice (89.65% accuracy) and open-ended questions (87.93% accuracy). This consistency in performance between different question formats is worth analyzing, given the structural differences between these types of questions. Multiple-choice questions, with their predefined options, typically align closely with the pattern recognition and classification algorithms intrinsic to many AI models (14). Open-ended questions, on the other hand, necessitate a more complex set of cognitive processes. The model must not only recognize and classify the pathology present in the image but also generate a coherent, relevant response without the guidance of predefined options. This involves a higher level of language understanding and generation capabilities, requiring the model to respond in a flexible manner, drawing from its training across medical knowledge domains. The consistently strong performance across both question types with no statistically significant difference highlights Claude 3.5 Sonnet's versatility in medical image interpretation regardless of the response format required. This suggests that the model possesses not only strong pattern recognition capabilities for identifying ophthalmic pathologies but also robust medical reasoning abilities that allow it to independently formulate accurate diagnostic and management recommendations when no options are provided. This capability shows a remarkable improvement in AI-related medical image interpretation, potentially paving the way for more comprehensive clinical decision support. However, it is imperative to emphasize that the images used in this study were sourced from board examination preparation materials which represent a standardized set of clinical scenarios and they may not fully capture the complexity of real-world clinical presentations.

The successful performance of Claude 3.5 Sonnet in cornea and external eye cases (95.65% accuracy in both open-ended and multiple-choice questions) compared to other subspecialties is an important finding. Several factors may contribute to this higher performance. Corneal and external eye conditions often present with more visually distinct features which may align better with the pattern recognition capabilities of AI models. In addition, clear views typically offered by external eye photographs and slit-lamp photography might enable more accurate interpretation. The relatively lower performance in "Strabismus, Pediatric Ophthalmology, and Genetics" (84.38% in open-ended and 87.50% in multiple-choice) may be attributed to several factors. First, this subspecialty often involves complex alignment issues that require a three-dimensional understanding from two-dimensional images. Second, pediatric ophthalmology cases frequently require integration of age-specific considerations and developmental factors that may not be as prominently featured in the training data. Third, genetic conditions in ophthalmology often present with subtle clinical manifestations that may be challenging to distinguish from static images alone.

In contrast to this current study, Minalache et al.'s study (12) evaluated both image-based and non-image-based case scenarios in ophthalmology and their LLMs showed the highest performance in the retina category (77% correct responses) and the lowest in neuro-ophthalmology (58% correct responses). Our findings differ significantly, with external eye and cornea showing the highest performance (95.65% accuracy in both open-ended and multiple-choice questions) and Strabismus and Pediatric Ophthalmology and Genetics showing the lowest (84.38% in open-ended and 87.50% in multiple-choice). Importantly, our study demonstrates substantially higher accuracy across all subspecialties. In addition, their research did not evaluate image-based questions related to cornea and external eye diseases or orbital-oculoplastic pathologies, which were included in our comprehensive analysis of six major ophthalmology subspecialties.

The performance of Claude 3.5 Sonnet suggests potential applications in both ophthalmology education and clinical practice. The model's high accuracy in board-style questions suggests its potential use in examination preparation, allowing students and residents to practice image interpretation and receive immediate feedback. Moreover, the model's ability to handle both multiple-choice and open-ended questions with similar accuracy could support a variety of learning styles and formats. However, it is crucial to emphasize that AI should be only complementary, not a replacement for traditional clinical education methods. In this current study, the aim was also to emphasize the potential of LLMs, and Claude 3.5 Sonnet's ability to efficiently analyze high volumes of images in busy clinical departments might offer an advantage in image-intensive

subspecialties like ophthalmology and it might serve as a valuable diagnostic tool. However, this research also highlights the need for cautious implementation to reduce the risk of over-reliance on it.

The current study has several limitations. First, the questions were derived from ophthalmology examination practice materials, which may not fully represent the complexity of real-world clinical scenarios. Second, while we achieved a more balanced distribution of questions across subspecialties, there were still slight variations in sample sizes between subspecialties that may have influenced performance comparisons. Third, our evaluation focused on a single multimodal LLM when the number of LLMs capable of processing medical images at this level was quite limited. Fourth, we did not include a comparative analysis with human ophthalmologists at different training levels, which would have provided valuable context for interpreting the model's performance relative to human experts.

Future research should include comparative analyses with human experts at various training levels (residents, fellows, and attending physicians) to provide valuable context about the model's relative capabilities across different ophthalmological issues. In addition, head-to-head comparisons between multiple LLMs with different architectures would help understand their relative strengths and limitations in ophthalmological image interpretation. Further work should explore how these models perform with more complex, ambiguous cases or rare conditions that might not be well-represented in standard question banks. Investigating how these models might be optimized specifically for ophthalmological applications through fine-tuning or specialized training could potentially enhance their performance in this domain.

Conclusion

These results demonstrate that Claude 3.5 Sonnet shows strong performance in interpreting ophthalmic images across all major ophthalmology subspecialties, with comparable accuracy in both multiple-choice (89.65%) and open-ended question formats (87.93%).

The model performed most effectively in the cornea and external eye subspecialty, while showing slightly lower but still impressive accuracy in strabismus, pediatric ophthalmology, and genetics cases. The consistent performance across different question formats highlights Claude 3.5 Sonnet's versatility in medical image interpretation and reasoning. These findings suggest potential applications in ophthalmology education, board examination preparation, and as a complementary tool in clinical settings. However, further research is imperative to validate the model's utility in real-world clinical scenarios and to compare its performance with that of ophthalmologists at various training levels.

Disclosures

Acknowledgements: This study was presented as an oral presentation at the SOE 2025 Congress, held on June 7-9 2025

Ethics Committee Approval: As this study did not involve human participants, institutional review board approval was not required.

Conflict of Interest: None declared.

Funding: The authors declare that this study has received no financial support.

Use of AI for Writing Assistance: Not declared.

Author Contributions: Concept – P.K., H.G.K.; Design – P.K.; Supervision – P.K.; Resource – H.G.K.; Materials – P.K., H.G.K.; Data Collection and/or Processing – P.K.; Analysis and/or Interpretation – P.K.; Literature Search – P.K., H.G.K.; Writing – P.K.; Critical Reviews – P.K., H.G.K.

Peer-review: Externally peer-reviewed.

References

1. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019;103:167–75. [CrossRef]
2. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol 2023;141:589–97. [CrossRef]
3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198. [CrossRef]
4. Oura T, Tatekawa H, Horiuchi D, Matsushita S, Takita H, Atsukawa N, et al. Diagnostic accuracy of vision-language models on Japanese diagnostic radiology, nuclear medicine, and interventional radiology specialty board examinations. Jpn J Radiol 2024;42:1392–8. [CrossRef]
5. Anthropic AI. The Claude 3 model family: Opus, Sonnet, Haiku. Claude-3 Model Card. 2024.
6. Cybersight. Available from: https://cybersight.org/. Accessed Sep 1, 2025.
7. Haddad F, Saade JS. Performance of ChatGPT on ophthalmology-related questions across various examination levels: Observational study. JMIR Med Educ 2024;10:e50842. [CrossRef]
8. Cai LZ, Shaheen A, Jin A, Fukui R, Yi JS, Yannuzzi N, et al. Performance of generative large language models on ophthalmology board-style questions. Am J Ophthalmol 2023;254:141–9. [CrossRef]
9. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312. Erratum in: JMIR Med Educ 2024;10:e57594. [CrossRef]
10. Kurokawa R, Ohizumi Y, Kanzawa J, Kurokawa M, Sonoda Y, Nakamura Y, et al. Diagnostic performances of Claude 3 Opus

and Claude 3.5 Sonnet from patient history and key images in Radiology's "Diagnosis Please" cases. Jpn J Radiol 2024;42:1399–402. [CrossRef]

11. Güneş YC, Cesur T, Çamur E, Günbey Karabekmez L. Evaluating text and visual diagnostic capabilities of large language models on questions related to the Breast Imaging Reporting and Data System Atlas 5th edition. Diagn Interv Radiol 2025;31:111–29. [CrossRef]
12. Mihalache A, Huang RS, Popovic MM, Patil NS, Pandya BU, Shor R, et al. Accuracy of an Artificial Intelligence chatbot's interpretation of clinical ophthalmic images. JAMA Ophthalmol 2024;142:321–6. [CrossRef]
13. Antaki F, Chopra R, Keane PA. Vision-language models for feature detection of macular diseases on optical coherence tomography. JAMA Ophthalmol 2024;142:573–6. [CrossRef]
14. Jiao C, Edupuganti NR, Patel PA, Bui T, Sheth V. Evaluating the artificial intelligence performance growth in ophthalmic knowledge. Cureus 2023;15:e45700. [CrossRef]