Contenst lists available at OneHealth Plus

OneHealth Plus Journal

Journal homepage: www.onehealthplus.com.tr

ChatGPT-5.2 for single B-scan macular OCT classification and triage: a comparative safety-focused study

Hakan Veli SAVAŞ[1]*, Mehmet Faruk BAŞ[2]

1 Ophthalmology, Elazığ Medical Hospital, Elazığ/TÜRKİYE
2Ophthalmology, Faculty of Medicine, Fırat University, Elazığ/TÜRKİYE

R E S E A R C H A R T I C L E

Received 22 December 2025;
Received in revised form 04 March 2026;
Accepted 09 March 2026

ORCID:
H.V. SAVAŞ: 0000-0002-3281-9892
M.F. BAŞ: 0009-0004-0993-0252

*Correspondence: H.V. SAVAŞ
Address: Cumhuriye Mah. Mustafa Doğak sk. Yağmur Apt. No:2 Kat:4 D:13 Merkez/Elazığ
Phone: +90 545 6919004
Mobil Phone: +90 545 6919004
e-mail: hakanvelisavas@hotmail.com

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics Approval
This study was conducted using anonymized, publicly available optical coherence tomography (OCT) images obtained from open-access datasets (Kaggle). As no human participants were directly involved and no identifiable personal data were used, ethics committee approval was not required in accordance with institutional and national research guidelines; therefore, no committee/university/institution name or approval letter number applies.

Conflict of Interest
The authors declared that there is no conflict of interest.

Author contribution
Idea, concept and design: HVS, MFB
Data collection and analysis: HVS, MFB
Drafting of the manuscript: HVS, MFB
Critical review: HVS, MFB

Data Availability
The datasets analyzed during the current study are publicly available from open-access sources, including the Kaggle repository (https://www.kaggle.com). No new data were generated for this study. Further details regarding the data sources are available from the corresponding author upon reasonable request.

Acknowledgements
The authors would like to thank the contributors of the publicly available OCT dataset—particularly Paul Timothy Mooney—for enabling open scientific research through the development and dissemination of open-access retinal imaging resources that supported this study.

A B S T R A C T

Purpose: To evaluate whether a general-purpose multimodal large language model (ChatGPT-5.2) can accurately classify single B-scan macular optical coherence tomography images into four diagnostic categories and provide appropriate clinical triage decisions.

Materials and Methods: This comparative image-interpretation study used 360 single B-scan macular optical coherence tomography images obtained from an open-access dataset, balanced across four categories: normal macula, choroidal neovascular membrane, diabetic macular edema, and drusen (90 images per class). Each image was independently assessed by ChatGPT-5.2 and by an evaluating ophthalmologist blinded to the reference labels. Diagnostic accuracy and Cohen's κ statistics were calculated. Paired comparisons were performed using an exact McNemar test. Triage recommendations generated by both raters were independently reviewed by a retina specialist and classified as appropriate or inappropriate.

Results: The ophthalmologist achieved a diagnostic accuracy of 98.1% (353/360), while ChatGPT-5.2 achieved 76.9% (277/360). Cohen's κ values were 0.974 for the ophthalmologist and 0.693 for ChatGPT-5.2. The model demonstrated high sensitivity for diabetic macular edema (98.9%) and normal macula (90.0%), but lower sensitivity for choroidal neovascular membrane (58.9%) and drusen (60.0%). Common misclassifications included drusen labeled as normal macula (40.0%) and choroidal neovascular membrane labeled as diabetic macular edema (23.3%). Triage appropriateness rates were 98.1% for the ophthalmologist and 77.2% for ChatGPT-5.2.

Conclusion

ChatGPT-5.2 demonstrated moderate performance for single B-scan macular optical coherence tomography diagnosis and triage, with substantially lower accuracy and triage appropriateness than clinician assessment and marked category-dependent variability. The reduced performance in choroidal neovascular membrane and drusen cases suggests that, in safety-critical triage settings, general-purpose multimodal models should be used as clinician-supervised decision-support tools rather than as stand-alone interpreters.

Keywords: optical coherence tomography, teleophthalmology, macular diseases, clinical triage, multimodal language models, artificial intelligence, patient safety

INTRODUCTION

Optical coherence tomography (OCT) is a fundamental imaging modality in the diagnosis, treatment planning, and follow-up of posterior segment diseases, as it enables layer-by-layer visualization of the retinal microanatomy. In macular pathologies that are particularly critical with respect to visual loss—such as CNV and DME—the identification of findings including intraretinal and subretinal fluid, hyperreflective material, disruption of retinal layers, and irregularities of the retinal pigment epithelium (RPE) directly influences clinical decision-making (Ahn, 2025; Murthy et al., 2016).

However, the clinical value of OCT depends not only on image acquisition but also on accurate image interpretation and the

timely referral of patients to the appropriate level of care. In rural and underserved regions where access to retina specialists is limited, delayed evaluations due to appointment overload or geographic barriers may increase the risk of preventable visual loss and lead to inefficiencies in referral pathways (Than et al., 2023; Walsh et al., 2024).

Teleophthalmology aims to provide a practical solution to these access challenges. In retinal diseases, studies have shown that teleophthalmology applications can enhance screening and follow-up capacity; however, their effectiveness depends on factors such as technical infrastructure, workflow integration, and standardization of image evaluation(Khan et al., 2024; Walsh et al., 2024). Following the widespread adoption of teleophthalmology services in the post–COVID-19 era, systematic reviews summarizing implementations across different countries have demonstrated that tele-services can improve access to care, while highlighting the persistent need for standardized approaches to clinical decision-making and triage (Walsh et al., 2024).

A systematic review focusing on low-resource settings similarly acknowledges the potential benefits of teleophthalmology but emphasizes the importance of sustainability and reliable evaluation processes in real-world practice (Khan et al., 2024). Therefore, the most critical component of teleophthalmology is not merely the remote transmission of images, but the generation of image-based triage decisions that are safe, traceable, and reproducible.

Artificial intelligence–based systems have the potential to reduce clinical workload and improve referral management at this juncture. A comprehensive review outlining the overall landscape of deep learning–based approaches in ophthalmology discusses not only the strengths of image-based artificial intelligence but also the challenges associated with clinical integration (Ting et al., 2019). In retinal diseases, a clinically applicable deep learning approach capable of generating combined "diagnosis + referral" outputs has demonstrated that OCT data can be used to achieve referral performance comparable to that of expert clinicians (De Fauw et al., 2018). In addition, the seminal work by Kermany and colleagues showed that image-based deep learning can classify pathologies associated with macular degeneration and diabetic retinopathy, laying the foundation for subsequent OCT-based studies (Kermany et al., 2018).Importantly, diagnostic and referral standards must remain aligned with established clinical guidelines. The current American Academy of Ophthalmology (AAO) Preferred Practice Pattern (PPP) documents for age-related macular degeneration and diabetic retinopathy emphasize the critical importance of timely clinical evaluation and appropriate follow-up strategies, particularly in the management of CNV and DME (American Academy of Ophthalmology, 2024).

In recent years, multimodal large language models (MLLMs), particularly generalist systems, have opened a new avenue of opportunity. Unlike conventional classifiers, these models can generate not only categorical outputs but also a brief rationale (morphological explanation) and a clinical routing recommendation, making them conceptually well aligned with teleophthalmology triage scenarios (Chen et al., 2023; Ma et al., 2025). However, the reliability of MLLMs in medical imaging remains an active area of investigation. A recent review synthesizing the current landscape of multimodal large language models in medical imaging highlights not only their potential clinical applications but also emphasizes the critical need for robust validation, safety frameworks, and standardized evaluation protocols (Nam et al., 2025). Studies that have systematically assessed GPT-4 in image-based tasks further demonstrate that, while the model can perform strongly in certain domains, its performance varies substantially by task type, with notable limitations reported in areas such as visual localization and grounding (Liu et al., 2024). Recent advances in large language models have prompted their evaluation in medical education and exam-style assessment settings. In ophthalmology, one study benchmarked multiple large language models on a Turkish chief-assistant examination question set and reported substantial variability in accuracy across models, highlighting both the promise of these systems and the need for safety-oriented evaluation before clinical translation (Canleblebici et al., 2024).

Within ophthalmology, the published evidence presents a heterogeneous picture. A benchmarking study reported that a GPT-4–based multimodal approach is not yet sufficient for clinical decision-making in ocular multimodal image interpretation and requires careful validation (Xu et al., 2024). In contrast, other studies have shown that incorporating visual information into clinical questions can improve model performance (Tomita et al., 2025). Taken together, these divergent findings suggest that, while the visual pattern

recognition capabilities of generalist MLLMs are promising, their safe deployment in high-impact scenarios such as teleophthalmology triage depends on critical safety layers. These include realistic prompt design, standardized output formats, and independent expert validation to ensure reliability and clinical appropriateness (Liu et al., 2024; Nam et al., 2025)

This study aims to comparatively evaluate the performance of a generalist multimodal large language model, ChatGPT-5.2, in generating four-class diagnoses (normal macula, CNV, DME, and drusen) and clinically meaningful triage decisions from single macular OCT B-scan images, benchmarked against an evaluating ophthalmologist and an independent retina specialist who was not involved in the study.

Reviews summarizing existing tools and applications of teleophthalmology in the retinal field emphasize the central role of triage within clinical workflows (Than et al., 2023). In addition, studies demonstrating the feasibility of teleophthalmology triage models for acute ophthalmic complaints highlight the practical value of accurate remote decision-making (Meshkin et al., 2022; Townsend et al., 2025). In parallel, a systematic review of remote and home-based OCT technologies reports that OCT is becoming an increasingly feasible component of teleophthalmology workflows (Dolar-Szczasny et al., 2024). Within this context, the present study evaluates ChatGPT-5.2's capacity to generate combined "diagnosis + brief explanation + triage" outputs using a safety-oriented assessment framework (Xu et al., 2024).

MATERIALS and METHODS

This study is a descriptive and comparative image-interpretation study that evaluates the performance of a general-purpose multimodal large language model (ChatGPT-5.2) in generating four-class diagnoses and clinical triage decisions from single macular OCT B-scan images, in comparison with assessments by an evaluating ophthalmologist and an independent retina specialist.

Data Source and Sample Construction

Ethics statement: This study used anonymized, publicly available OCT images obtained from an open-access dataset. No human participants were recruited, and no identifiable personal data were accessed. Therefore, in accordance with institutional and national guidance for secondary analysis of anonymized public data, the study was considered exempt from IRB/ethics committee review, and informed consent was not required.

OCT images were downloaded from a publicly available open-access repository (Kaggle: 'Retinal OCT Images [Optical Coherence Tomography]') derived from the retinal OCT dataset originally published by Kermany et al. (2018).A total of 360 images were analyzed, with 90 images selected for each diagnostic category. The public dataset provides individual 2D macular OCT B-scan images labeled by diagnostic category rather than volumetric OCT cubes. Images were analyzed as provided by the dataset, and no disease-specific curation (e.g., selecting the "most representative" slice) was applied across groups. All images were fovea-centered macular B-scans, as verified during the selection process. The reference ("ground truth") diagnostic labels were fixed prior to the study, and both the model and the evaluating ophthalmologist were masked to these labels during image assessment.

All images were evaluated using ChatGPT-5.2. Each OCT image was uploaded individually to the model, and a standardized prompt was applied before each evaluation. This prompt required the model to assign the image to one of the four diagnostic categories (normal, CNV, DME, or drusen), generate a brief morphological explanation consisting of 3–5 sentences, and classify the case into one of four triage levels (1: urgent referral, 2: short-term evaluation, 3: follow-up every three months, 4: routine follow-up). All outputs were recorded in a standardized three-line format for each case (Diagnosis/ Explanation/Triage).

To minimize context carryover ("in-context learning") and ensure independence between cases, each OCT image was evaluated in a fresh, isolated chat/session with no prior images or messages present. No iterative prompting, feedback, or follow -up questions were used, and the standardized prompt was applied once per image. The exact prompt text and the required three-line output format are provided in Appendix A.

The same images were independently classified by an evaluating ophthalmologist who was blinded to the reference diagnostic labels. No predefined reference label was assigned for triage decisions; instead, triage outputs from both the ophthalmologist and ChatGPT-5.2 were reviewed by an independent retina specialist and categorized as appropriate or inappropriate. The retina specialist also assessed the morphological explanations

generated by ChatGPT-5.2 and rated them as concordant, partially concordant, or discordant.

The study reported: (i) the agreement between model- and clinician-generated diagnoses and the reference labels (diagnostic accuracy), (ii) the agreement of triage decisions with the independent retina specialist's assessment (triage appropriateness), and (iii) patterns of diagnostic confusion among categories generated by ChatGPT-5.2.

Statistical Analysis

Overall performance was assessed using accuracy and class-specific recall. To account for agreement beyond chance, Cohen's κ was calculated, and 95% confidence intervals were estimated for proportion-based metrics. Because the evaluating ophthalmologist and ChatGPT-5.2 assessed the same set of images, a paired analytical approach was adopted for between-method comparisons, with the exact McNemar test applied where appropriate. All statistical analyses were performed using Python version 3.11.2.

RESULTS

A total of 360 macular OCT B-scan images were included in the study. The dataset was balanced across four predefined diagnostic categories: DME, CNV, drusen, and normal macula (n = 90 for each category). All images were evaluated in a blinded manner by the evaluating ophthalmologist and independently by ChatGPT-5.2. The appropriateness of triage decisions was assessed by an independent retina specialist who was not involved in the study.

Table 1. Diagnostic performance of the evaluating ophthalmologist and ChatGPT-5.2 for four-class classification of macular OCT images.

Metric	Evaluating Ophthalmologist	ChatGPT-5.2
Overall accuracy	353/360 (98.1%)	277/360 (76.9%)
95% CI (accuracy)	96.0–99.1	72.3–81.0
Cohen's κ	0.974	0.693
Recall – DME	88/90 (97.8%)	89/90 (98.9%)
Recall – CNV	87/90 (96.7%)	53/90 (58.9%)
Recall – Drusen	89/90 (98.9%)	54/90 (60.0%)
Recall – Normal	89/90 (98.9%)	81/90 (90.0%)

DME, diabetic macular edema; CNV, choroidal neovascularization; CI, confidence interval.

Diagnostic Performance: Evaluating Ophthalmologist vs. ChatGPT-5.2

The evaluating ophthalmologist correctly classified 353 of 360 cases, yielding an overall accuracy of 98.1%. In contrast, ChatGPT-5.2 correctly classified 277 of 360 cases, corresponding to an accuracy of 76.9%. The Cohen's κ coefficient was 0.974 for the evaluating ophthalmologist and 0.693 for ChatGPT-5.2.

In paired comparisons, the diagnostic accuracy of the evaluating ophthalmologist was significantly higher than that of ChatGPT-5.2 ($p < 0.001$). There were 81 cases in which the evaluating ophthalmologist was correct while ChatGPT-5.2 was incorrect, and 5 cases in which ChatGPT-5.2 was correct while the evaluating ophthalmologist was incorrect. Diagnostic performance metrics are summarized in Table 1.

Triage Appropriateness: Comparison with Retina Specialist Review

The triage decisions of the evaluating ophthalmologist were deemed appropriate in 353 of 360 cases (98.1%), whereas the triage decisions generated by ChatGPT-5.2 were considered

Table 2. Triage appropriateness based on independent retina specialist review. The between-method difference in paired appropriateness was statistically significant ($p<0.001$).

Metric	Evaluating Ophthalmologist	ChatGPT-5.2
Overall appropriateness (retina specialist)	353/360 (98.1%)	278/360 (77.2%)
95% CI (appropriateness)	96.0–99.1	72.6–81.3
Appropriateness – DME	88/90 (97.8%)	87/90 (96.7%)
Appropriateness – CNV	87/90 (96.7%)	57/90 (63.3%)
Appropriateness – Drusen	89/90 (98.9%)	53/90 (58.9%)
Appropriateness – Normal	89/90 (98.9%)	81/90 (90.0%)

DME, diabetic macular edema; CNV, choroidal neovascularization; CI, confidence interval.

appropriate in 278 of 360 cases (77.2%). Paired analysis demonstrated that the triage appropriateness of the evaluating ophthalmologist was significantly higher than that of ChatGPT-5.2 ($p < 0.001$).

When stratified by diagnostic category, ChatGPT-5.2 showed high triage appropriateness in cases of DME (96.7%) and normal macula (90.0%), but substantially lower appropriateness in CNV (63.3%) and drusen (58.9%) cases. Triage appropriateness results are presented in Table 2.

Class-Specific Diagnostic Performance and Diagnostic Confusions

On a class-specific basis, ChatGPT-5.2 demonstrated high sensitivity in DME cases (98.9%, 89/90) and normal macula cases (90.0%, 81/90), whereas a marked decline in performance was observed for CNV (58.9%, 53/90) and drusen (60.0%, 54/90) cases (Table 1).

The most frequent error made by ChatGPT-5.2 was the misclassification of drusen as normal macula (36/90; 40.0%). In CNV cases, the most common misclassifications involved confusion with DME (21/90; 23.3%) and drusen (15/90; 16.7%). The most frequent diagnostic confusions are summarized in Table 3.

Table 3. Most frequent diagnostic confusions made by ChatGPT-5.2. Values are presented as n and percentage within the true class.

True diagnosis	ChatGPT-5.2 prediction	n	% within true class
Drusen	Normal	36	40.0%
CNV	DME	21	23.3%
CNV	Drusen	15	16.7%
Normal	Drusen	9	10.0%
DME	Drusen	1	1.1%
CNV	Normal	1	1.1%

DME, diabetic macular edema; CNV, choroidal neovascularization

DISCUSSION

In this descriptive, comparative image-interpretation study, we assessed whether a general-purpose multimodal large language model (ChatGPT-5.2) can generate a macular OCT diagnosis and an associated triage recommendation from a single macular B-scan, and we benchmarked its outputs against an evaluating ophthalmologist; triage appropriateness was adjudicated by an independent retina specialist. To our knowledge, this is the first safety-oriented, systematic evaluation of ChatGPT-5.2 for OCT-based macular diagnosis and triage within this single B-scan framework.

Three messages stand out. First, ChatGPT-5.2 demonstrated moderate overall capability, but its reliability did not match clinician performance when used as a stand-alone interpreter in this setting; in particular, triage appropriateness was substantially lower than that of the evaluating ophthalmologist (overall 77.2% vs 98.1% by retina-specialist review). Second, performance was strongly class-dependent: it remained relatively strong in DME and normal scans but weakened notably in CNV and drusen. Third, the error profile was systematic rather than random: drusen most often shifted toward "normal," and CNV was most frequently confused with DME (and, less commonly, drusen).

These findings matter because misrouting CNV is not a benign "pattern similarity" mistake: in neovascular AMD, early detection and prompt treatment improve visual outcomes, and triage systems should be designed to minimize high-impact misses, not merely optimize average accuracy (American Academy of Ophthalmology, 2024). From a tele-triage safety perspective, therefore, our results support focusing on worst-case risks in the categories where delayed referral can be vision-threatening.

Why CNV and drusen are more challenging from a single B-scan

A plausible explanation is that our design intentionally restricts input to a single B-scan, removing the volumetric context clinicians routinely use, such as cross-slice confirmation of subtle retinal pigment epithelium and pigment epithelial detachment configurations (RPE/PED) and assessment of lesion extent. In borderline disease, CNV and drusen often depend on subtle layer-level cues and contextual corroboration (adjacent slices, fluid distribution patterns, and clinical correlation) that may be underdetermined in one section. In addition, ChatGPT-5.2 was used in a zero-shot, generalist configuration with a standardized prompt rather than being trained specifically on OCT or optimized for retina micro-morphology. Finally, requiring a definitive label with a brief rationale may inadvertently encourage "explanatory completion" when visual grounding is uncertain—an inherent risk in generative multimodal systems—potentially compounding errors in subtle categories.

Positioning within OCT-AI and referral-recommendation literature

Our results are consistent with a broader pattern in CNV-AI: specialized models trained for retinal OCT tasks, often at scale and sometimes leveraging richer context, can achieve clinician-level performance for diagnosis and referral, whereas generalist, zero-shot systems may show more variable class-wise reliability. The work by Kermany et al. demonstrated strong performance for classifying major retinal pathologies from OCT images under a dedicated deep learning training paradigm(Kermany et al., 2018). De Fauw et al. extended this concept toward clinical pathways by showing that a deep learning system trained on large sets of 3D OCT scans can generate referral recommendations reaching or exceeding expert performance for sight-threatening retinal diseases (De Fauw et al., 2018).

The contrast is instructive: the limitations we observed do not imply that OCT is "unsuitable for AI," but rather suggest a mismatch between (i) safety-critical triage based on fine retinal morphology and (ii) a generalist multimodal model asked to infer diagnosis and urgency from a single B-scan without OCT -specific training or volumetric context. Importantly, the clinician benchmark in our study underscores this point: the evaluating ophthalmologist's triage decisions were judged appropriate in nearly all cases, reinforcing that—at present—generalist models may be better viewed as adjunctive tools rather than replacements for expert interpretation in high-stakes categories.

Tele-ophthalmology relevance and why this question is becoming more urgent

Tele-ophthalmology is increasingly moving beyond remote fundus imaging toward workflows that include OCT acquired outside traditional clinics. Recent literature on home/remote OCT in tele-ophthalmology highlights growing feasibility and clinical interest, while also emphasizing that rigorous validation of safety, accuracy, and clinical utility is still needed as these technologies mature (Dolar-Szczasny et al., 2024). In this context, our work addresses a practical, safety-framed question: as OCT becomes more accessible in remote or home -based settings, what happens when interpretation and triage are delegated—partly or wholly—to a generalist multimodal model, and in which disease categories do the risks become unacceptable without human oversight?

Implications for clinical translation and governance

Our findings support a cautious, decision-support positioning: a generalist multimodal model may provide workflow augmentation for lower-risk patterns, but categories with higher clinical cost of delay (notably CNV and some drusen scenarios) should default toward conservative routing and human review. This aligns with emerging governance principles for large multimodal models in health, which emphasize clear definition of intended use, risk management, validation, and accountability (WHO Releases AI Ethics and Governance Guidance for Large Multi-Modal Models). It also resonates with the FDA's framing of clinical decision support, which stresses that systems should be designed so clinicians can independently review the basis for recommendations rather than rely primarily on the software output (U.S Food and drug Administration, n.d.). Practically, this highlights an evaluation priority beyond top-line accuracy: future studies should explicitly test whether model outputs remain auditable and safety-aligned when uncertainty is high, and whether the model's stated morphological rationale genuinely corresponds to the image features clinicians would use.

Limitations and future directions

Several limitations warrant emphasis. First, images were sampled from an open-access Kaggle dataset with a balanced design (n = 360; 90 per class), which supports controlled comparison but may not represent real-world prevalence, acquisition variability, comorbidity, or mixed pathology. Second, although all images were fovea-centered, we intentionally used a single B-scan per case; this isolates the question of single-slice interpretability but does not reflect routine clinical OCT interpretation, where volumetric scans and multimodal correlation (e.g., OCTA and clinical history) are often decisive, and it increases the risk of missing pathology outside the selected slice. Third, triage appropriateness was determined by a single independent retina specialist; broader adjudication panels could improve generalizability, and prospective workflow studies could better capture downstream outcomes such as delayed care, unnecessary referrals, and patient-centered harms/benefits. Finally, we evaluated one generalist model in a zero-shot configuration under one standardized prompting approach; comparative evaluations across multiple models, prompt strategies, and uncertainty-aware output policies are needed.

CONCLUSION

This study shows that ChatGPT-5.2 can generate macular OCT diagnoses and triage recommendations from a single B-scan with meaningful potential, but with performance that varies substantially by disease category. The relative weaknesses observed in CNV and drusen—classes where delayed or inappropriate routing can carry high clinical cost—support a safety-centered interpretation rather than reliance on average performance. Overall, our findings suggest that generalist multimodal models are currently better positioned as clinician-supervised decision-support tools than as independent triage agents. As remote and home-based OCT becomes more common, larger prospective validations embedded within real-world workflows will be essential to define safe and effective clinical integration.

REFERENCES

Age-Related Macular Degeneration PPP 2024 - American Academy of Ophthalmology. (n.d.). Retrieved 17 December 2025, from https://www.aao.org/education/preferred-practice-pattern/age-related-macular-degeneration-ppp?utm_source=chatgpt.com

Ahn, S. J. (2025). Retinal Thickness Analysis Using Optical Coherence Tomography: Diagnostic and Monitoring Applications in Retinal Diseases. *Diagnostics*, *15*(7). https://doi.org/10.3390/DIAGNOSTICS15070833

CANLEBLEBİCİ, M., DAL, A., & ERDAĞ, M. (2024). Evaluation of the Performance of Large Language Models (ChatGPT-3.5, ChatGPT-4, Bing and Bard) in Turkish Ophthalmology Chief-Assistant Exams: A Comparative Study. *Türkiye Klinikleri Journal of Ophthalmology*, *33*(3), 163. https://doi.org/10.5336/ophthal.2024-102632

Chen, J., Wu, X., Li, M., Liu, L., Zhong, L., Xiao, J., Lou, B., Zhong, X., Chen, Y., Huang, W., Meng, X., Gui, Y., Chen, M., Wang, D., Dongye, M., Zhang, X., Cheung, C. Y., Lai, I. F., Yan, H., … Lin, H. (2023). EE-Explorer: A Multimodal Artificial Intelligence System for Eye Emergency Triage and Primary Diagnosis. *American Journal of Ophthalmology*, *252*, 253–264. https://doi.org/10.1016/j.ajo.2023.04.007

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., … Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine 2018 24:9*, *24*(9), 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

Dolar-Szczasny, J., Drab, A., & Rejdak, R. (2024). Home-monitoring/remote optical coherence tomography in teleophthalmology in patients with eye disorders-a systematic review. *Frontiers in Medicine*, *11*. https://doi.org/10.3389/FMED.2024.1442758

Fda. (n.d.). *Contains Nonbinding Recommendations Clinical Decision Support Software Guidance for Industry and Food and Drug Administration Staff*. Retrieved 21 December 2025, from https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., … Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, *172*(5), 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

Khan, I. A., Bashar, Md. A., Tripathi, A., & Priyanka, N. (2024). The Benefits and Challenges of Implementing Teleophthalmology in Low-Resource Settings: A Systematic Review. *Cureus*, *16*(9), e70565. https://doi.org/10.7759/CUREUS.70565

Liu, Y., Li, Y., Wang, Z., Liang, X., Liu, L., Wang, L., Cui, L., Tu, Z., Wang, L., & Zhou, L. (2024). A systematic evaluation of GPT-4V's multimodal capability for chest X-ray image analysis. *Meta-Radiology*, *2*(4), 100099. https://doi.org/10.1016/J.METRAD.2024.100099

Ma, R., Cheng, Q., Yao, J., Peng, Z., Yan, M., Lu, J., Liao, J., Tian, L., Shu, W., Zhang, Y., Wang, J., Jiang, P., Xia, W., Li, X., Gan, L., Zhao, Y., Zhu, J., Qin, B., Jiang, Q., … Zhao, C. (2025). Multimodal machine learning enables AI chatbot to diagnose ophthalmic diseases and provide high-quality medical responses. *Npj Digital Medicine 2025 8:1*, *8*(1), 64-. https://doi.org/10.1038/s41746-025-01461-0

Meshkin, R. S., Armstrong, G. W., Hall, N. E., Rossin, E. J., Hymowitz, M. B., & Lorch, A. C. (2022). Effectiveness of a telemedicine program for triage and diagnosis of emergent ophthalmic conditions. *Eye 2022 37:2*, *37*(2), 325–331. https://doi.org/10.1038/s41433-022-01940-8

Murthy, R. K., Haji, S., Sambhav, K., Grover, S., & Chalam, K. V. (2016). Clinical applications of spectral domain optical coherence tomography in retinal diseases. *Biomedical Journal*, *39*(2), 107–120. https://doi.org/10.1016/J.BJ.2016.04.003

Nam, Y., Kim, D. Y., Kyung, S., Seo, J., Song, J. M., Kwon, J., Kim, J., Jo, W., Park, H., Sung, J., Park, S., Kwon, H., Kwon, T., Kim, K., & Kim, N. (2025). Multimodal Large Language

Models in Medical Imaging: Current State and Future Directions. *Korean Journal of Radiology*, *26*(10), 900–923. https://doi.org/10.3348/KJR.2025.0599

Than, J., Sim, P. Y., Muttuvelu, D., Ferraz, D., Koh, V., Kang, S., & Huemer, J. (2023). Teleophthalmology and retina: a review of current tools, pathways and services. *International Journal of Retina and Vitreous*, *9*(1), 76. https://doi.org/10.1186/S40942-023-00502-8

Ting, D. S. W., Pasquale, L. R., Peng, L., Campbell, J. P., Lee, A. Y., Raman, R., Tan, G. S. W., Schmetterer, L., Keane, P. A., & Wong, T. Y. (2019). Artificial intelligence and deep learning in ophthalmology. *The British Journal of Ophthalmology*, *103*(2), 167–175. https://doi.org/10.1136/BJOPHTHALMOL-2018-313173

Tomita, K., Nishida, T., Kitaguchi, Y., Kitazawa, K., & Miyake, M. (2025a). Image Recognition Performance of GPT-4V(ision) and GPT-4o in Ophthalmology: Use of Images in Clinical Questions. *Clinical Ophthalmology (Auckland, N.Z.)*, *19*, 1557–1564. https://doi.org/10.2147/OPTH.S494480

Townsend, N. A., Shah, S., Reyes, J., Townsend, J. H., Bozung, A., Ricur, G., & Aboumourad, R. J. (2025). Tele-ophthalmology as an effective triaging tool for acute ophthalmic concerns. *Frontiers in Ophthalmology*, *4*, 1511378. https://doi.org/10.3389/FOPHT.2024.1511378

Walsh, L., Hong, C. Y., Chalakkal, R., Hong, S. C., O'Keeffe, B., & Ogbuehi, K. (2024). A Systematic Review of Teleophthalmology Services Post-COVID-19 Pandemic in New Zealand, the United Kingdom, Australia, the United States of America, and Canada. *Telemedicine Journal and E-Health : The Official Journal of the American Telemedicine Association*, *30*(12), 2795–2804. https://doi.org/10.1089/TMJ.2024.0258

WHO releases AI ethics and governance guidance for large multi-modal models. (n.d.). Retrieved 21 December 2025, from https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models?utm_source=chatgpt.com

Xu, P., Chen, X., Zhao, Z., & Shi, D. (2024a). Unveiling the clinical incapabilities: a benchmarking study of GPT-4V(ision) for ophthalmic multimodal image analysis. *The British Journal of Ophthalmology*, *108*(10), 1384–1389. https://doi.org/10.1136/BJO-2023-325054