Preprints.org

Article

Not peer-reviewed version

Evaluating a Multi-Modal Large Language Model for Ophthalmology Triage

Caius Goh , Jabez Ng , Au Wei Yung , Clarence See , Alva Lim , Fan Xiuyi , Kelvin Li *

Posted Date: 18 September 2025

doi: 10.20944/preprints202509.1349.v1

Keywords: ophthalmology triage; large-language models; vision-language models; artificial intelligence

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1349.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Evaluating a Multi-Modal Large Language Model for Ophthalmology Triage

Caius Goh [1], Jabez Ng [2], Au Wei Yung [1], Clarence See [1], Alva Lim [3], Fan Xiuyi [2] and Kelvin Li [1,*]

[1] Department of Ophthalmology, Tan Tock Seng Hospital, Singapore 308433, Singapore
[2] College of Computing and Data Science, Nanyang Technology University, Singapore 639798, Singapore
[3] Department of Ophthalmology and Visual Sciences, Khoo Teck Puat Hospital, Singapore 768828, Singapore
* Correspondence: kelvin.li.eye@gmail.com

Abstract

Background/Purpose: Ophthalmic triage is challenging for non-specialists due to limited training and rising global eye disease burden. This study evaluates a multimodal framework integrating clinical text and ophthalmic imaging with large language models (LLMs). Hallucination detection and chain-of-thought (CoT) reasoning were incorporated to improve diagnostic accuracy. **Methods**: A dataset of 41 ophthalmology cases from a Singapore restructured hospital was pre-processed with acronym expansion, sentence reconstruction, and hallucination detection. To address dataset size limitations, 100 synthetic cases were generated via one-shot GPT-4 prompting, validated by semantic checks and ophthalmologist review. Three diagnostic approaches were tested: Text-Only, Image-Assisted, and Image with CoT. Diagnostic performance was quantified using SNOMED-CT mapping and a dissimilarity score reflecting semantic distance between predicted and reference diagnoses. **Results**: The synthetic dataset included anterior segment (n=40), posterior segment (n=35), and extraocular (n=25) cases. The text-only approach yielded a mean dissimilarity of 6.353 +/- 1.685. Incorporation of image assistance reduced this to 5.234 +/- 1.305, while CoT prompting provided further gains when imaging cues were ambiguous. **Conclusions**: The multimodal pipeline improved diagnostic alignment in ophthalmology triage. Image inputs enhanced accuracy, and CoT reasoning reduced errors from ambiguous features, supporting its potential as an accurate tool for ophthalmology triage.

Keywords: ophthalmology triage; large-language models; vision-language models; artificial intelligence

1. Introduction

Efficient ophthalmology triage ensures timely care for high-risk patients while preventing unnecessary specialist referrals that burden ophthalmology services. In clinical practice, patients presenting with ocular complaints are often first evaluated by general or emergency physicians, whose accuracy in triage depends heavily on their training and experience. With ocular-related complaints representing a growing burden on emergency departments, proper triage of these cases is essential for proper healthcare resource allocation. A study conducted in the United States found that of 16.8 million eye-related emergency department visits occurring from 2010-2017, 44.8% of the cases were for non-emergent conditions [1]. Inaccurate or inconsistent referrals can delay urgent care, contribute to overcrowded clinics, and misallocate specialist resources [2,3]. These challenges are compounded by the rising prevalence of vision-threatening conditions such as diabetic retinopathy, cataract, and glaucoma [4].

Recent advances in large language models (LLMs) and vision-language models (VLMs) offer the potential to enhance triage decision-making. LLMs can process and interpret unstructured clinical text, while VLMs extend this capability to medical images, including anterior segment and fundus photographs [5,6]. When combined, these modalities enable a more holistic case assessment.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1349.v1

However, LLMs typically produce single-step outputs that can oversimplify complex reasoning, leading to diagnostic inaccuracies [7,8]. Chain-of-Thought (CoT) prompting addresses this limitation by breaking down decision-making into sequential steps, improving interpretability and diagnostic reasoning in medical tasks [9,10].

Existing research in AI-assisted ophthalmology triage is limited. Prior models have primarily used either text or image inputs, and few have examined structured multimodal reasoning. Studies evaluating GPT-4 and other generalist VLMs have reported variable performance, particularly for images requiring fine spatial interpretation, such as gonioscopy or visual field charts [11]. Moreover, issues such as hallucination, bias, and difficulty processing unstructured clinical notes remain significant barriers to safe deployment [12–15].

This study addresses these gaps by developing a multimodal diagnostic framework that integrates structured text processing, hallucination detection, and CoT-augmented reasoning. Diagnostic performance is evaluated using a novel SNOMED-CT Directed Acyclic Graph (DAG) metric, which accounts for semantic similarity between related ophthalmic diagnoses. We hypothesize that integrating image data and CoT reasoning will improve diagnostic accuracy compared to text-only prompts, and that the DAG metric will provide a clinically meaningful measure of AI performance.

2. Methods

2.1. Overview of Proposed Pipeline

Figure 1 depicts an end-to-end overview of the proposed methodology. It illustrates the sequential steps involved in processing unstructured ED notes and physical examination findings, detecting hallucinations, generating a validated synthetic dataset, and evaluating diagnosis predictions using multimodal LLMs and a DAG-based graph evaluation framework

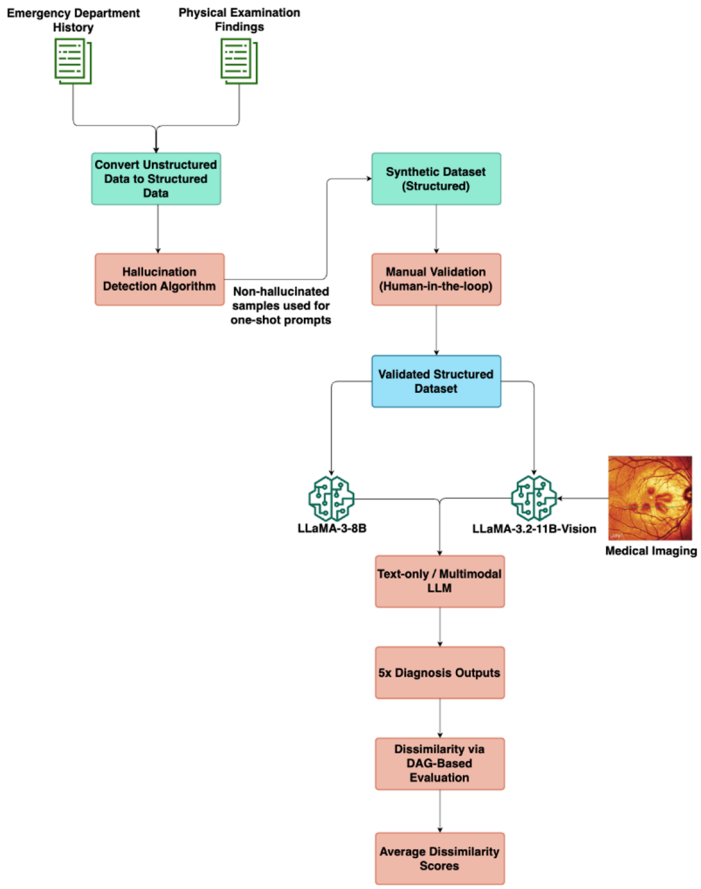

Figure 1. Proposed end-to-end processing pipeline.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1349.v1

2.2. Data Collection

This single-centre retrospective study involved patients who were referred to the Ophthalmology department of Tan Tock Seng Hospital, Singapore. The patients were randomly selected across a period from November 1st, 2022 to May 30th, 2023. Retrospective data collected for analysis included: the date of referral, history taken at the time of referral, the provisional diagnosis given at the time of referral, the time interval from referral to an Ophthalmology department visit, the history taken at the Ophthalmology department, and the diagnosis given at the Ophthalmology department. A full list of the collected data can be found in Supplemental Table S1.

This study was conducted in accordance with the Domain Specific Review Board (DSRB) of the National Healthcare Group (NHG), Singapore, which also approved this study protocol. Patient information was anonymized and de-identified before the analysis.

2.3. Clinical Dataset and Pre-Processing

A clinical dataset of ophthalmology triage information for 41 patients was obtained from Tan Tock Seng Hospital. All identifiable patient information was removed before analysis, and the resulting dataset was fully anonymised in accordance with the National Healthcare Group's data governance and the Singapore Personal Data Protection Act (PDPA), ensuring that it was untraceable to any individual patient. This anonymisation process was applied prior to pre-processing, thereby ensuring that the study did not infringe on patient confidentiality or ethical standards for secondary data use.

Raw medical notes may prove challenging for LLMs due to the heavy use of medical abbreviations, lack of standardised grammar, and ambiguity depending on the medical context. To ensure machine-readable LLM inputs whilst retaining the necessary medical information, a pre-processing pipeline was performed that reconstructed fragmented medical notes into grammatically coherent and structured sentences. A hallucination detection algorithm was designed to ensure that the original meaning of the notes was kept (Figure 2).

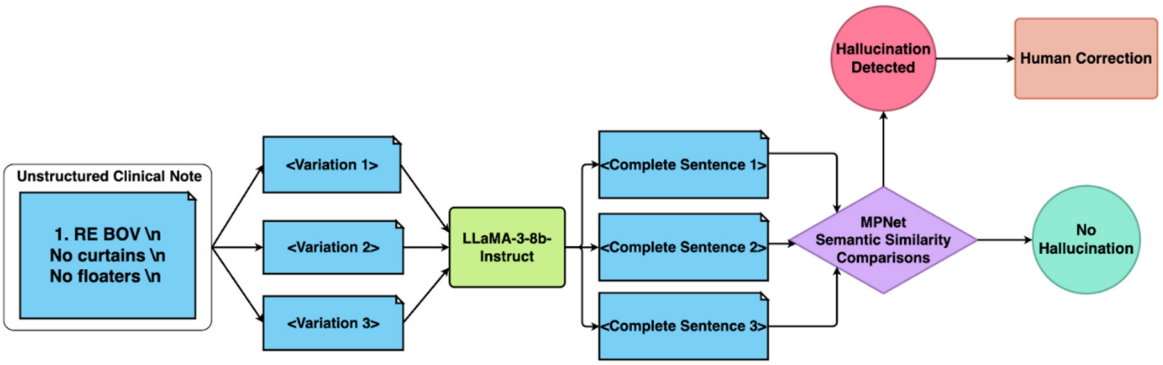

Figure 2. Hallucination Detection Pipeline Overview.

The first step of the pipeline was acronym expansion and text standardization. Domain-specific acronyms (e.g. "BOV" for blurring of vision) were replaced using a medical acronym dictionary (Supplemental Table S2). Following acronym expansion, multiple variations of the clinical note were generated by randomly shuffling sentence components while preserving the original meaning. The text was segmented using newline characters and re-combined. This approach introduces controlled randomness while maintaining the core clinical information.

The second step was sentence reconstruction: the LLaMA 3 8B Instruct model was employed to reconstruct fragmented text into complete, grammatically coherent sentences. To ensure fidelity, multiple reconstructions were generated for each case. Text generation was regulated based on the set of hyperparameters outlined in Table 1, to influence the diversity and stability of the output. The temperature parameter was set to 0.6 to introduce a moderate level of randomness while preventing

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1349.v1

excessive deviation from the expected phrasing. The top-p (nucleus value) of 0.9 restricts token selection to the most probable subset, reducing the likelihood of generating inconsistent or overly creative responses. The max token limit was set to 2048, ensuring that the generated text remains an appropriate length.

Table 1. LLaMA 3 8B Instruct Model Parameters.

Parameter	Value
Model Name	LLaMA 3 8B Instruct
Temperature	0.6
Max Tokens	2048
Top-p	0.9
Sampling Method	Nucleus Sampling (Top-p)

The final step in the pipeline was semantic similarity assessment. The reconstructions were then screened using the Masked and Permutated Pre-training Network (MPNet) embeddings with a cosine similarity threshold of 0.8 to filter out hallucinated or inconsistent outputs. In the case that all surpass the threshold of 0.8, the one with the highest overall similarity is selected. MPNet thus helps to prioritize consistent outputs while filtering out misleading or hallucinated responses, ensuring the processed text remains factually accurate and clinically safe for model input.

2.4. Synthetic Dataset Generation

The clinical dataset obtained from ophthalmologists at Tan Tock Seng Hospital proved insufficient for evaluation of the LLM-based diagnostic and triage systems, due to a small sample size of 41 cases, as well as poor depth of clinical documentation and completeness necessary for independent diagnostic reasoning.

To overcome this, a synthetic dataset of 100 ophthalmic triage cases was generated based on the clinical dataset using one-shot prompting with GPT-4 to create a more comprehensive and clinically realistic dataset that would mimic information obtainable from real-world cases more closely. The outputs of the earlier data pre-processing and hallucination detection pipeline were then fed into GPT-4 to generate this synthetic dataset. GPT-4 was selected for synthetic data generation due to its superior few-shot reasoning and clinical language capabilities. This approach produced high-quality cases that mirrored the linguistic and conceptual structure of authentic ophthalmic records. The diagnoses included in the dataset were selected to represent ophthalmological conditions commonly seen in the emergency and outpatient settings.

After selecting 100 final cases, GPT-4 was used to generate corresponding Emergency Department (ED) histories, physical examination findings, diagnostic reasoning, and urgency levels for each case. To enhance plausibility and diversity, a one-shot prompting approach was adopted. This involved presenting GPT-4 with the curated examples from TTSH's clinical notes which had passed through the data pre-processing and hallucination detection as detailed above. This strategy allowed for the synthesis of detailed and differentiated cases that reflect the complexity and nuance of actual ophthalmology presentations, while also addressing the limitations of the original dataset.

To assess the semantic alignment between the synthetic and real-world datasets, we computed sentence similarity scores using SBERT (MPNet cosine similarity). The results showed a moderate correlation, with scores ranging from 0.6 to 0.7. This evaluation confirms that the synthetic cases retained meaningful linguistic and conceptual resemblance to authentic ophthalmology records, while preserving variability and completeness necessary for robust model evaluation.

The final step in the synthetic data generation process involved expert validation by an ophthalmologist to ensure clinical accuracy, diagnostic validity, and real-world applicability. The expert conducted a comprehensive review of all generated cases, evaluating the consistency and plausibility of emergency department histories and physical examination findings. Diagnoses were revised as needed to align with expected clinical presentations, and textual descriptions were refined

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1349.v1

to enhance clarity, specificity, and realism. This validation process also incorporated ophthalmic imaging to enable multimodal evaluation. The expert assigned appropriate image modalities—such as fundus photographs, external eye images, and Humphrey visual field tests—according to each case's final diagnosis. This human-in-the-loop validation process enhanced the dataset's clinical fidelity and diagnostic robustness, establishing it as a reliable benchmark for evaluating both text-based and multimodal (text and image) ophthalmology triage models.

2.5. Diagnostic Framework

We evaluated three separate diagnostic approaches:

Firstly, the direct prompting (text-only, also known as non-image) approach serves as the baseline in this study. In this approach, the LLM generated a diagnosis based solely on textual inputs - specifically, the patient's Emergency Department (ED) History and Physical Examination findings. We utilized the LLaMA-3-8b-Instruct model from Meta, which has been fine-tuned for instruction-based reasoning tasks. The LLM received a structured prompt guiding it to produce the most likely diagnosis along with justification, minimizing ambiguity and facilitating systematic evaluation.

Secondly, the multi-modal input approach incorporates both textual inputs and visual data to evaluate whether image integration enhances clinical reasoning. The model is provided with three types of input:

- ED history and physical examination findings, which form the textual component of the diagnostic evaluation.
- A medical image associated with the patient's condition, corresponding to one of the ophthalmic imaging modalities under investigation.
- A description of the image type (e.g. fundus image, anterior segment photo) to provide context for interpretation.

Lastly, to overcome limitations observed with the direct multimodal prompting, we augmented the process with CoT reasoning. This seeks to improve LLM interpretability and diagnostic reasoning by decomposing complex tasks into sequential steps. CoT prompting was applied selectively to cases identified as visually ambiguous or diagnostically complex. These were defined as cases where initial image-assisted diagnosis showed high dissimilarity scores, in which an expert ophthalmologist identified subtle or overlapping visual features that could not clearly distinguish between differential diagnoses. This approach employed a two-step process: first, the model generates a preliminary diagnosis and justification based on the clinical text provided. Next, a subsequent prompt integrated the relevant imaging data along with modality-specific instructions (e.g. "review retinal vessels and optic disc changes" for fundus images) to refine the initial diagnosis. By decoupling text and image interpretation and requiring explicit reasoning across both, this approach fosters greater interpretability and aligns with real-world clinical workflows, where imaging is used to support or adjust an initial hypothesis.

2.6. Evaluation Metric

To assess the performance of our model, we quantified the dissimilarity between the predicted diagnosis and the actual final diagnosis. In experiments incorporating self- consistency, we computed the average dissimilarity across multiple predictions per case to account for variability in model outputs. A key challenge in this evaluation lies in the inherent difficulty of exact string matching due to variations in medical terminology (e.g. acute posterior vitreous detachment and posterior vitreous detachment would register as a mismatch despite their clinical relationship). Furthermore, conventional semantic similarity metrics may not be entirely reliable in the medical domain, as certain diagnoses with distinct lexical representations may be synonymous (e.g., "dry eyes" and "keratoconjunctivitis sicca").

To address this, we developed a Directed Acyclic Graph (DAG)-based dissimilarity metric using SNOMED-CT (Systematized Nomenclature of Medicine – Clinical Terms). While SNOMED-CT

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1349.v1

provides multiple API endpoints for querying medical concepts, our methodology specifically utilized two key endpoints to facilitate an accurate and structured evaluation of diagnostic dissimilarity. The first API endpoint was used to obtain the concept ID of a diagnosis via a search query. The second API endpoint retrieved the Concept IDs of all ancestral terms associated with the given concept ID. This allowed us to construct a DAG reflecting the hierarchical relationships among ophthalmic concepts. The diagnostic dissimilarity score was then defined as the sum of the shortest path distances from the predicted diagnosis and the reference diagnosis to their lowest common ancestor (LCA), with lower scores indicating a higher clinical alignment. This framework allows for an accurate and structured evaluation of model performance, ensuring that clinically relevant relationships between diagnoses are effectively captured.

A worked example of the DAG-based dissimilarity metric is shown in Table 2.

Table 2. Worked examples of diagnostic dissimilarity scores based on the DAG-based dissimilarity metric.

Reference Diagnosis	Predicted Diagnosis	LCA term	Distance (Ref → LCA)	Distance (Pred → LCA)	Total Dissimilarity
Rhegmatogenous Retinal Detachment	Retinal Detachment	Retinal Detachment	0	1	1
Non-proliferative Diabetic Retinopathy	Diabetic Retinopathy, Macular Edema	Diabetic Retinopathy	1	1	2
Bilateral Papilledema	Benign Intracranial Hypertension	Disorder of Intracranial Pressure	1	1	1

2.7. Statistical Analysis

Descriptive statistics (mean, median, variance) were computed for the dissimilarity scores across all experimental conditions. Subgroup analyses were performed to examine the effect of each imaging modality (fundus, anterior segment, external, etc.) and the additional benefit of the chain-of-thought approach in cases with ambiguous visual cues.

3. Results

The synthetic dataset of 100 ophthalmic triage cases was designed to closely represent the variety and depth of cases encountered in the emergency and outpatient settings. The distribution of cases is represented in Figure 3, and included anterior segment (n=40), posterior segment (n=35) and extraocular (n=25) diagnoses. Each case included standardised clinical history, physical examination findings, and ophthalmic imaging in one of several modalities: fundus photography, anterior segment with fluorescein staining, gonioscopy, Humphrey visual field testing, external ocular photographs, or 9-gaze photographs. This distribution enabled subgroup analyses by both diagnosis category and image modality.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1349.v1

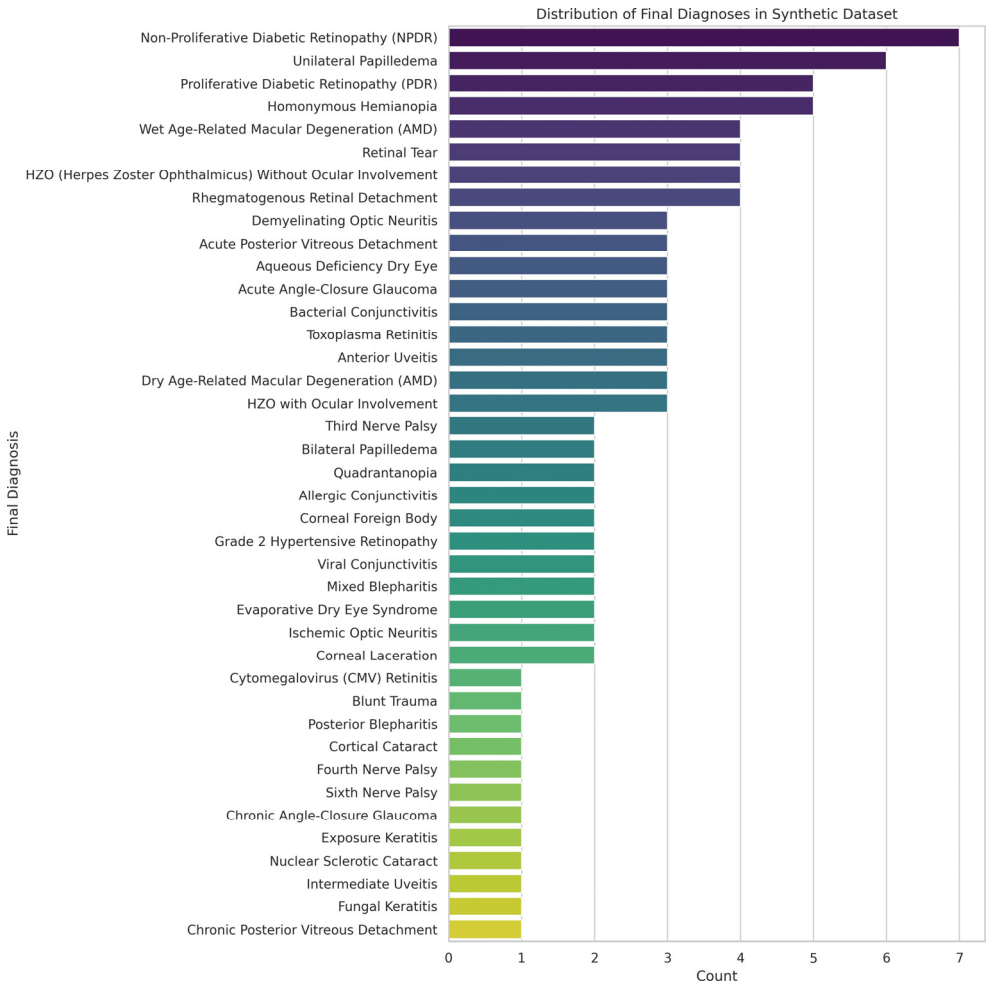

Figure 3. Hallucination Detection Pipeline Overview.

3.1. By Image Modality

Dissimilarity scores were grouped by image type and compared between image-assisted and non-image prompts to better understand how different ophthalmic image modalities influence diagnostic performance. As depicted in Figure 4, image-assisted prompts generally yielded lower mean dissimilarity scores across most modalities.

Notably, the largest decrease in mean dissimilarity score was observed in the Anterior Segment with Fluorescein modality, from 10.31 (non-image) to 4.99 (image). The largest dissimilarity scores were noted for the Humphrey Visual Field Result image modality, at 13.61 (non-image) and 12.51 (image). The only image modality where image assisted diagnoses yielded a higher dissimilarity score than text-only was External Photo, with a dissimilarity score of 3.55 (non-image) and 5.04 (image).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1349.v1

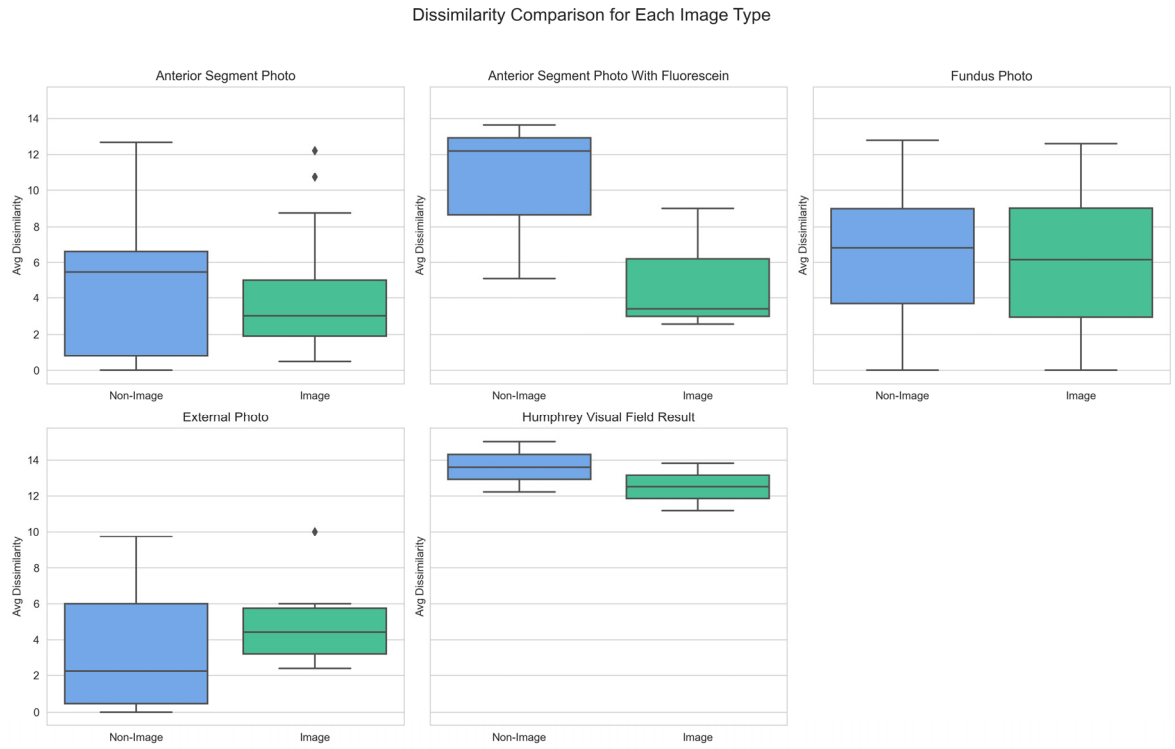

Figure 4. Box Plot of Dissimilarity Scores by Image Modality and Model Type (Non-image vs Image).

3.1. By Diagnosis Category

Dissimilarity scores were then grouped by diagnostic categories, to assess performance across different groups of ocular diagnoses. As seen in Figure 7, image-assisted prompts again produced lower dissimilarity scores across most categories.

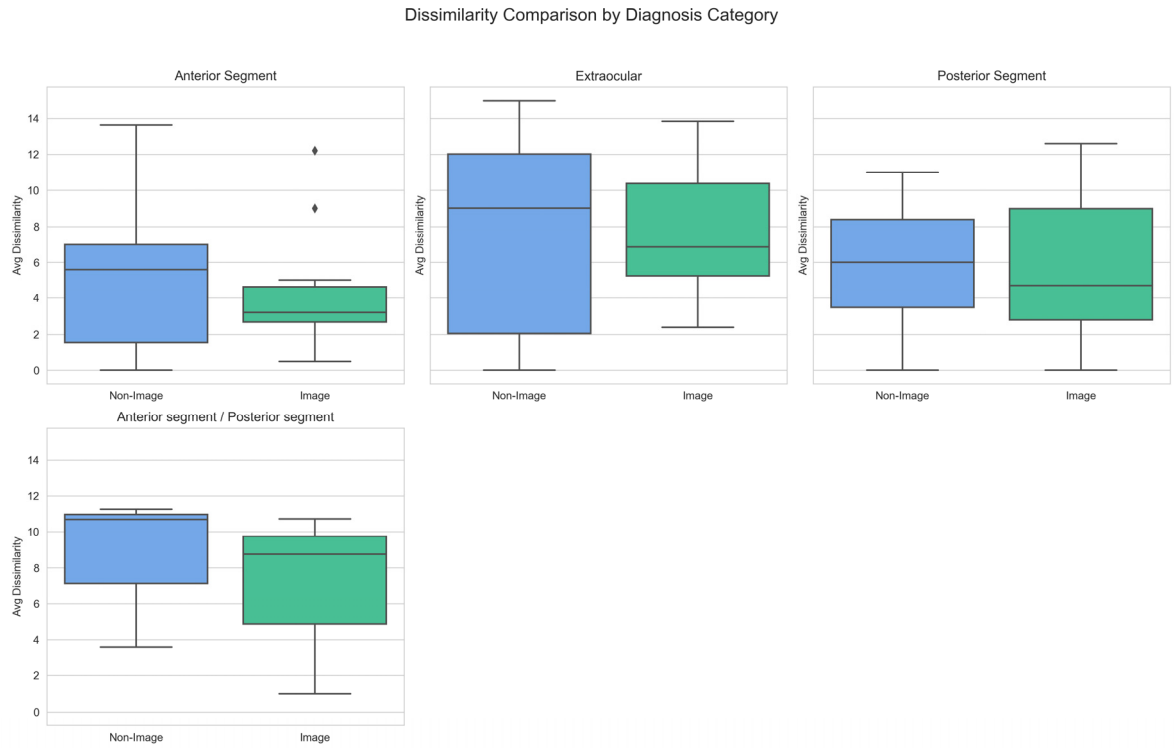

Figure 5. Box Plot of Dissimilarity Scores by Diagnosis Category and Model Type (Non-image vs Image).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1349.v1

The largest decrease in dissimilarity score with image-assisted diagnosis was noted in the Anterior Segment (5.75 vs 4.10) and Anterior Segment / Posterior Segment (8.51 vs 6.82) diagnosis group. The only category where image assisted diagnosis resulted in a higher dissimilarity score was the Extraocular (7.39 vs 7.56) diagnosis group.

In terms of specific diagnoses, visual data was noted to reduce diagnostic accuracy in specific cases. For anterior segment photos, dissimilarity scores were higher in Acute Angle Closure Glaucoma (+5.01), Cortical Cataract (+2.80) and Allergic Conjunctivitis (+2.22). For fundus photos, dissimilarity scores were notably higher in Arteritic Anterior Ischaemic Optic Neuropathy (+2.20). For gonioscopic imaging, the highest increase was noted for Chronic Angle Closure Glaucoma (+5.60). For External/Ocular Motility Photos, reduced accuracy was observed in Fourth Nerve Palsy (+8.12), Third Nerve Palsy (+5.00) and Surgical Third Nerve Palsy (+3.00).

3.3. Impact of Chain-of-Thought Reasoning

In cases characterized by ambiguous visual features, the CoT approach enhanced diagnostic precision. The Average Dissimilarity scores with and without CoT are depicted in Table 3.

Table 3. Average Dissimilarity Scores between Image-only and Image + Chain-of Thought (CoT) Approaches for Selected Diagnoses.

Final Diagnosis	Average Dissimilarity (Image)	Average Dissimilarity (Image + COT)
Acute Angle-Closure Glaucoma	5.01	0.00
Allergic Conjunctivitis	3.02	1.70
Arteritic Anterior Ischemic Optic Neuropathy	10.50	5.20
Chronic Angle-Closure Glaucoma	12.20	1.40
Cortical Cataract	3.40	0.00
Fourth Nerve Palsy	10.00	0.00
Sixth Nerve Palsy	6.00	3.00
Surgical Third Nerve Palsy	3.00	1.00
Third Nerve Palsy	5.00	3.40

Across the board, the inclusion of CoT prompts resulted in a consistent reduction in the dissimilarity scores. This was particularly notable in conditions such as Chronic Angle Closure Glaucoma (12.20 to 1.40) and Arteritic Anterior Ischaemic Optic Neuropathy (10.50 to 5.20). In addition, significant reductions in dissimilarity scores were seen for all diagnoses based on 9-gaze photographs.

4. Discussion

This pilot study provides three key insights: firstly, that LLMs can serve as a viable aid in ophthalmology triage; secondly, that incorporating multimodal inputs improves diagnostic accuracy; and lastly, that structured CoT prompting further enhances performance, particularly in visually ambiguous cases. These findings are important in addressing limitations in existing LLM use in ophthalmology triage, including inconsistent clinical image interpretation, and inaccurate diagnostic reasoning processes.

Multimodal inputs were found to improve diagnostic accuracy in LLM-assisted triage, as reflected by lower dissimilarity scores compared to text-only prompts. The average dissimilarity score dropped from 6.353 with text-only inputs to 5.235 when images were included, aligning with prior findings that multimodal systems can improve clinical decision-making when images are meaningfully integrated with clinical context[16,17].

However, the benefits of image-assisted diagnosis were not universal. Certain image modalities, such as external ocular photos and gonioscopic imaging, showed higher dissimilarity scores than text-only inputs. This phenomenon has been demonstrated in prior studies as well - Mikhail et al.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1349.v1

demonstrated that GPT-4's diagnostic accuracy was significantly lower when images were included, compared to text-only inputs [18]. A possible explanation is that LLMs trained only on general information may lack the depth of domain-specific knowledge necessary for a specialized task [19] - including the interpretation of complex ophthalmic images, such as complex or multi-panelled images. Other studies with similar findings give alternative explanations – such as the reliance of models on textual clues due to modality dominance rather than grounding in the visual evidence, leading to sidelining of the visual inputs [20].

4.1. Influence of Image Modality and Diagnosis Category on Dissimilarity

Dissimilarity scores varied across modalities, with anterior segment imaging demonstrating the greatest improvements. Specifically, fluorescein-stained images led to the most pronounced reduction in dissimilarity from 10.31 (text-only) to 4.99 (image-assisted), highlighting the utility of detailed, high-contrast imaging in surface-level pathologies. Conversely, image-assisted diagnosis of conditions requiring complex spatial interpretation, such as visual field defects, showed smaller improvements in diagnostic accuracy. This pattern is consistent with recent evaluations of VLMs, which reported that while they can leverage clear, structured imaging, they perform poorly on neuro-ophthalmic tasks such as optic disc swelling classification, even when domain-specific models or advanced prompting strategies are applied [21].

By diagnostic category, the greatest gains were observed in anterior and mixed anterior/posterior segment pathologies. On the other hand, diagnostic performance was diminished for extraocular conditions, such as cranial nerve palsies. A possible explanation is that clinical symptoms of extraocular conditions, such as diplopia or ptosis, are better conveyed through clinical history and examination findings rather than an ophthalmic image.

While our study demonstrates the value of a multimodal approach, it is important to acknowledge that the contribution of imaging may vary significantly across different ophthalmic sub-specialties. For instance, in neuro-ophthalmology, a sub-specialty with a heavy emphasis on patient history and neurological signs, a comprehensive analysis of clinical text alone can be highly effective. A study by Wang et al. found that a simple combination of patient history and chief complaint could predict a correct neuro-ophthalmology diagnosis with an accuracy of approximately 90% [22]. This suggests that for certain sub-specialties like neuro-ophthalmology, focusing on a robust text-only model may be more efficient and equally accurate, even reducing the risk of misinterpretation of ambiguous imaging prompts.

4.2. Chain of Thought Reasoning and Its Effect on Diagnostic Dissimilarity

The addition of CoT prompting consistently improved diagnostic accuracy across all tested conditions. Notably, in challenging cases such as chronic angle closure glaucoma and arteritic anterior ischaemic optic neuropathy, dissimilarity scores were reduced by more than 50% with COT prompting. These results corroborate earlier studies demonstrating that CoT enhances clinical reasoning by enforcing a stepwise, hypothesis-driven approach [8,23] capable of mimicking the diagnostic thought process utilized by specialists in a real-world clinical setting.

In the context of ophthalmology, CoT appears particularly effective in tasks involving complex diagnostic pathways or ambiguous visual findings, where simple visual-text alignment is insufficient. While no prior studies have assessed the use of CoT-augmented LLMs in the specific field of ophthalmology, other studies have verified the utility of CoT in improving the performance and consistency of LLMs in general diagnostic reasoning [9].

Our results further suggest that CoT may partially mitigate the limitations of VLMs in underperforming image modalities, such as external eye photographs or visual fields, by compensating for the lack of granular visual understanding with structured textual inference.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1349.v1

4.3. Significance of Study Findings

Our multi-modal diagnostic framework may have significant clinical implications for the future of ophthalmic triage and diagnostic decision support systems.

Firstly, the consistent improvement in diagnostic accuracy with addition of visual prompts suggests that multimodal inputs would achieve superior diagnostic alignment with real-world ophthalmic reasoning. This suggests that VLM-based systems may serve as reliable front-line triage tools in the emergency departments or primary care settings.

Secondly, the study identifies certain ophthalmic diagnoses and imaging modalities in which the addition of imaging prompts degrades diagnostic accuracy. This underscores the importance of selecting appropriate modalities when deploying AI models in the real-world setting. It also reinforces that the deployment of AI-assisted triage should be tailored to align with the diagnostic utility of specific image types, while highlighting potential areas for model refinement such as in diagnosis of complex motility disorders, or interpretation of gonioscopic images.

Thirdly, the integration of CoT prompting further improved diagnostic alignment especially in cases with ambiguous visual cues or overlapping diagnoses. CoT-augmented models may thus be of utility to support decision-making in complex referrals, or during after-hour evaluations where subspecialty expertise is limited.

Lastly, the novel, clinically informed diagnostic evaluation metric based on SNOMED-CT derived DAG accounts for semantic and clinical proximity between diagnoses. It better reflects the realities of diagnostic uncertainty and interrelatedness in ophthalmic conditions. With the ability to capture this nuance, the model evaluation becomes more aligned with real-world triage needs where identifying the general category and disease urgency outweighs the need to diagnose a potentially rare or esoteric condition.

4.4. Limitations

While the findings of this study are promising, the following limitations must be acknowledged. Firstly, the insufficiency of the clinical dataset obtained was evident in terms of sample size and quality of textual records, precluding its use in the experiments conducted. This prevented us from assessing the entire pipeline from pre-processing and hallucination detection to the performance of our model in terms of diagnostic triage. Secondly, the synthetic nature of the dataset may not capture all real-world variabilities. Due to the manual effort required for annotation and validation, the dataset was limited to 100 samples, which may constrain the generalizability of the findings. Lastly, although the SNOMED-CT is a comprehensive clinical ontology with robust information retrieval capabilities, it does not capture all diagnostic terms encountered in practice. As a result, a small number of cases required manual inspection and the application of rule-based mappings to ensure accurate alignment between predicted and reference diagnoses.

5. Conclusions

This study presents a robust, clinically grounded evaluation of a multimodal diagnostic framework in ophthalmology triage. By integrating real and synthetic datasets, image modalities, and diagnostic categories, we demonstrate that multimodal LLMs, when augmented with chain-of-thought prompting, can approach clinically meaningful diagnostic accuracy. The novel use of a SNOMED-CT-based DAG metric to evaluate diagnostic dissimilarity also offers a more nuanced and clinically relevant performance metric than conventional classification accuracy. These findings have significant implications for the future of AI-assisted triage systems in ophthalmology, particularly in resource-limited or emergency settings where rapid, accurate referrals are critical.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. Table S1: Ophthalmology triage clinical dataset, Table S2: Medical Acronym dictionary used in Hallucination Detection Pipeline.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1349.v1

Author Contributions: Conceptualization, K.L.; investigation, C.G. and J.N.; methodology, J.N., F.X. and K.L.; resources, A.W.Y., C.S. and A.L.; software, J.N. and F.X.; supervision, K.L.; visualization: C.G. and J.N.; writing – original draft, C.G. and J.N.; writing – review and editing, C.S., A.L., F.X. and K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki, and approved by the Domain Specific Review Board of the National Healthcare Group (reference number 2023/00348, 10 October 2023).

Informed Consent Statement: Patient consent was waived by the Domain Specific Review Board of the National Healthcare Group due to complete anonymization and de-identification of patient information prior to analysis.

Data Availability Statement: The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mir, T.A.; Mehta, S.; Qiang, K.; Adelman, R.A.; Del Priore, L.V.; Chow, J. Association of the Affordable Care Act with Eye-Related Emergency Department Utilization in the United States. *Ophthalmology* **2022**, *129*, 1412–1420, doi:10.1016/j.ophtha.2022.06.038.
2. Khou, V.; Ly, A.; Moore, L.; Markoulli, M.; Kalloniatis, M.; Yapp, M.; Hennessy, M.; Zangerl, B. Review of Referrals Reveal the Impact of Referral Content on the Triage and Management of Ophthalmology Wait Lists. *BMJ Open* **2021**, *11*, e047246, doi:10.1136/bmjopen-2020-047246.
3. Grossmann, F.F.; Zumbrunn, T.; Ciprian, S.; Stephan, F.-P.; Woy, N.; Bingisser, R.; Nickel, C.H. Undertriage in Older Emergency Department Patients – Tilting against Windmills? *PLoS ONE* **2014**, *9*, e106203, doi:10.1371/journal.pone.0106203.
4. Yin, J.; Jiang, B.; Zhao, T.; Guo, X.; Tan, Y.; Wang, Y. Trends in the Global Burden of Vision Loss among the Older Adults from 1990 to 2019. *Front. Public Health* **2024**, *12*, 1324141, doi:10.3389/fpubh.2024.1324141.
5. Cascella, M.; Semeraro, F.; Montomoli, J.; Bellini, V.; Piazza, O.; Bignami, E. The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives. *J. Med. Syst.* **2024**, *48*, 22, doi:10.1007/s10916-024-02045-3.
6. Nazi, Z.A.; Peng, W. Large Language Models in Healthcare and Medical Domain: A Review. *Informatics* **2024**, *11*, 57, doi:10.3390/informatics11030057.
7. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models Are Zero-Shot Reasoners 2022.
8. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models 2022.
9. Wu, C.-K.; Chen, W.-L.; Chen, H.-H. Large Language Models Perform Diagnostic Reasoning 2023.
10. Liu, J.; Wang, Y.; Du, J.; Zhou, J.T.; Liu, Z. MedCoT: Medical Chain of Thought via Hierarchical Expert. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Miami, Florida, USA, 2024; pp. 17371–17389.
11. Sittig, D.F.; Singh, H. Recommendations to Ensure Safety of AI in Real-World Clinical Care. *JAMA* **2025**, *333*, 457, doi:10.1001/jama.2024.24598.
12. Xu, H.; Stetson, P.D.; Friedman, C. A Study of Abbreviations in Clinical Notes. *AMIA Annu. Symp. Proc. AMIA Symp.* **2007**, *2007*, 821–825.
13. Moon, S.; McInnes, B.; Melton, G.B. Challenges and Practical Approaches with Word Sense Disambiguation of Acronyms and Abbreviations in the Clinical Domain. *Healthc. Inform. Res.* **2015**, *21*, 35, doi:10.4258/hir.2015.21.1.35.
14. Koga, S.; Du, W. Challenges of Integrating Chatbot Use in Ophthalmology Diagnostics. *JAMA Ophthalmol.* **2024**, *142*, 883, doi:10.1001/jamaophthalmol.2024.2303.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1349.v1

15. Mihalache, A.; Huang, R.S.; Popovic, M.M.; Patil, N.S.; Pandya, B.U.; Shor, R.; Pereira, A.; Kwok, J.M.; Yan, P.; Wong, D.T.; et al. Accuracy of an Artificial Intelligence Chatbot's Interpretation of Clinical Ophthalmic Images. *JAMA Ophthalmol.* **2024**, *142*, 321, doi:10.1001/jamaophthalmol.2024.0017.
16. Chen, J.; Wu, X.; Li, M.; Liu, L.; Zhong, L.; Xiao, J.; Lou, B.; Zhong, X.; Chen, Y.; Huang, W.; et al. EE-Explorer: A Multimodal Artificial Intelligence System for Eye Emergency Triage and Primary Diagnosis. *Am. J. Ophthalmol.* **2023**, *252*, 253–264, doi:10.1016/j.ajo.2023.04.007.
17. Tomita, K.; Nishida, T.; Kitaguchi, Y.; Kitazawa, K.; Miyake, M. Image Recognition Performance of GPT-4V(Ision) and GPT-4o in Ophthalmology: Use of Images in Clinical Questions. *Clin. Ophthalmol.* **2025**, *Volume 19*, 1557–1564, doi:10.2147/OPTH.S494480.
18. Mikhail, D.; Milad, D.; Antaki, F.; Milad, J.; Farah, A.; Khairy, T.; El-Khoury, J.; Bachour, K.; Szigiato, A.-A.; Nayman, T.; et al. Multimodal Performance of GPT-4 in Complex Ophthalmology Cases. *J. Pers. Med.* **2025**, *15*, 160, doi:10.3390/jpm15040160.
19. Balaskas, G.; Papadopoulos, H.; Pappa, D.; Loisel, Q.; Chastin, S. A Framework for Domain-Specific Dataset Creation and Adaptation of Large Language Models. *Computers* **2025**, *14*, 172, doi:10.3390/computers14050172.
20. Buckley, T.; Diao, J.A.; Rajpurkar, P.; Rodman, A.; Manrai, A.K. Multimodal Foundation Models Exploit Text to Make Medical Image Predictions 2023.
21. Li, K.Z.; Nguyen, T.T.; Moss, H.E. Performance of Vision Language Models for Optic Disc Swelling Identification on Fundus Photographs. *Front. Digit. Health* **2025**, *7*, 1660887, doi:10.3389/fdgth.2025.1660887.
22. Wang, M.Y.; Asanad, S.; Asanad, K.; Karanjia, R.; Sadun, A.A. Value of Medical History in Ophthalmology: A Study of Diagnostic Accuracy. *J. Curr. Ophthalmol.* **2018**, *30*, 359–364, doi:10.1016/j.joco.2018.09.001.
23. WU, J.; Wu, X.; Yang, J. Guiding Clinical Reasoning with Large Language Models via Knowledge Seeds 2024.
24. Author 1, A.B. (University, City, State, Country); Author 2, C. (Institute, City, State, Country). Personal communication, 2012.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.