Ophthalmol Ther (2025) 14:3007–3020
https://doi.org/10.1007/s40123-025-01248-2

ORIGINAL RESEARCH

Leveraging ChatGPT for Report Error Audit: An Accuracy-Driven and Cost-Efficient Solution for Ophthalmic Imaging Reports

Yufeng Xu · Daohuan Kang · Danli Shi · Yih Chung Tham · Andrzej Grzybowski · Kai Jin

Received: August 4, 2025 / Accepted: September 8, 2025 / Published online: September 30, 2025

ABSTRACT

Introduction: Accurate ophthalmic imaging reports, including fundus fluorescein angiography (FFA) and ocular B-scan ultrasound, are essential for effective clinical decision-making. The current process, involving drafting by residents followed by review by ophthalmic technicians and ophthalmologists, is time-consuming and prone to errors. This study evaluates the effectiveness of ChatGPT-4o in auditing errors in FFA and ocular B-scan reports and assesses its potential to reduce time and costs within the reporting workflow.

Methods: Preliminary 100 FFA and 80 ocular B-scan reports drafted by residents were analyzed using GPT-4o to identify the errors in identifying left or right eye and incorrect anatomical descriptions. The accuracy of GPT-4o was compared to retinal specialists, general

Yufeng Xu and Daohuan Kang contributed equally to this work and share the first authorship.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40123-025-01248-2.

Y. Xu · K. Jin (✉)
Eye Center of Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China
e-mail: jinkai@zju.edu.cn

Y. Xu · K. Jin
Zhejiang Provincial Key Laboratory of Ophthalmology, Zhejiang Provincial Clinical Research Center for Eye Diseases, Zhejiang Provincial Engineering Institute on Eye Diseases, Hangzhou, China

D. Kang
Department of Ophthalmology, Children's Hospital, Zhejiang University School of Medicine, National Clinical Research Center for Child Health, Hangzhou, China

D. Shi
School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong

D. Shi
Research Centre for SHARP Vision (RCSV), The Hong Kong Polytechnic University, Kowloon, Hong Kong

Y. C. Tham
Centre for Innovation and Precision Eye Health, National University of Singapore, Singapore, Singapore

Y. C. Tham
Department of Ophthalmology, National University of Singapore, Singapore, Singapore

Y. C. Tham
Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore

A. Grzybowski
Institute for Research in Ophthalmology, Foundation for Ophthalmology Development, Poznan, Poland

A. Grzybowski
Department of Ophthalmology, University of Warmia and Mazury, Olsztyn, Poland

ophthalmologists, and ophthalmic technicians. Additionally, a cost-effective analysis was conducted to estimate time and cost savings from integrating GPT-4o into the reporting process. A pilot real-world validation with 20 erroneous reports was also performed between GPT-4o and human reviewers.

Results: GPT-4o demonstrated a detection rate of 79.0% (158 of 200; 95% CI 73.0–85.0) across all examinations, which was comparable to the average detection performance of general ophthalmologists (78.0% [155 of 200; 95% CI 72.0–83.0]; $P \geq 0.09$). Integration of GPT-4o reduced the average report review time by 86%, completing 180 ophthalmic reports in approximately 0.27 h compared to 2.17–3.19 h by human ophthalmologists. Additionally, compared to human reviewers, GPT-4o lowered the cost from $0.21 to $0.03 per report (savings of $0.18). In the real-world evaluation, GPT-4o detected 18 of 20 errors with no false positives, compared to 95–100% by human reviewers.

Conclusions: GPT-4o effectively enhances the accuracy of ophthalmic imaging reports by identifying and correcting common errors. Its implementation can potentially alleviate the workload of ophthalmologists, streamline the reporting process, and reduce associated costs, thereby improving overall clinical workflow and patient outcomes.

Keywords: Ophthalmology; Imaging reports; Error audit; ChatGPT; Cost-effective

Key Summary Points

Why carry out this study?

Accurate ophthalmic imaging reports, such as fundus fluorescein angiography (FFA) and ocular B-scan ultrasound, are crucial for effective clinical decision-making. The current process, involving report drafting by residents followed by reviews from ophthalmic technicians and specialists, is time-consuming and prone to errors, potentially delaying patient care.

This study evaluates the effectiveness of GPT-4o in auditing errors in FFA and ocular B-scan reports, with the goal of reducing time and costs in the reporting workflow.

What was learned from the study?

GPT-4o significantly improved the accuracy of ophthalmic imaging reports by detecting common errors such as laterality confusions and descriptor misregistrations. The integration of GPT-4o reduced the time required for report review and significantly lowered associated costs.

The implementation of GPT-4o can reduce the workload for ophthalmologists, streamline the reporting process, enhance diagnostic accuracy, and lead to significant time and cost savings. Despite the promising results, challenges remain, such as ensuring integration into real-world clinical workflows and maintaining the transparency and reliability of AI models for broader clinical adoption.

INTRODUCTION

Ophthalmic multimodal imaging examinations are crucial for developing effective treatment plans for various eye diseases [1]. Accurate and consistent ophthalmic imaging reports, including fundus fluorescein angiography (FFA) and ocular B-scan ultrasound, are essential for effective clinical decision-making and patient management [2, 3]. At some institutions, these reports are initially drafted by resident doctors and then reviewed by board-certified ophthalmologists [4]. While this review process enhances accuracy, it is time-consuming and resource-intensive [5]. Additionally, increasing workloads, high-pressure environments, and limitations of current speech recognition technologies contribute to frequent reporting errors, such as errors in identifying left or right eye and incorrect anatomical descriptions [6]. Although robust prevalence estimates for ophthalmic imaging reports are currently lacking, evidence from radiology indicates that real-time

day-to-day reporting errors average approximately 3–5%, while retrospective reviews identify errors in roughly 30% of examinations [7, 8]. These errors can lead to significant clinical repercussions if not corrected. Currently, advanced proofreading tools to detect these specific errors are not widely available.

Advanced artificial intelligence models, particularly ChatGPT developed by OpenAI, offer promising solutions to these challenges [9–11]. ChatGPT has demonstrated potential in various medical applications, including transforming free-text reports into structured formats, generating impression sections, and supporting diagnostic processes through differential diagnosis generation [12–17]. In ophthalmology, ChatGPT could revolutionize report accuracy by automatically identifying and correcting common errors in FFA and ocular B-scan reports. This would reduce the burden on supervising ophthalmologists and serve as an educational tool for resident doctors [18, 19].

This study aimed to evaluate ChatGPT 4o's effectiveness in auditing prevalent errors in FFA and ocular B-scan reports. It will assess ChatGPT's accuracy in identifying laterality confusions and descriptor misregistrations, compare its performance with different levels of ophthalmologists, and estimate potential time and cost savings. By leveraging ChatGPT's advanced language processing capabilities, this research seeks to enhance the accuracy and efficiency of ophthalmic reporting, ultimately improving patient outcomes and optimizing clinical workflows.

METHODS

This retrospective study was approved by the ethics committee (Approval No. Y2023–1073), and the requirement for informed consent was waived due to its retrospective nature. All data provided to GPT-4o were anonymized, ensuring that no patient-identifying information was included. The study was conducted in accordance with the Declaration of Helsinki (1964) and its later amendments. No identifiable participant information was included in the study, and thus, specific participant consent for publication was not required.

Study Design and Data Set

A total of 180 original ophthalmic imaging reports, including 100 FFA reports and 80 ocular B-scan ultrasound reports, covering a wide range of pathologic abnormalities, were collected from Eye Center, The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China, between July 2024 and December 2024. These reports were selected and randomized using a freeware research data randomization tool (https://www.randomizer.org). The dataset was divided into two sets: correct and incorrect, each containing 90 reports. In the incorrect set, a total of 200 errors were deliberately introduced by an ophthalmology resident doctor, with a maximum of three errors per report (see Fig. 1).

Error Classification and Verification

The deliberate introduction of errors was designed to mimic common error patterns identified in prior literature (e.g., omissions, insertions, misspellings) to standardize the evaluation of error detection across human and AI reviewers. Errors were categorized into four primary types based on prior research to encompass the most common error types in ophthalmic reports [2–4]:

(a) Omission: The exclusion of relevant words or phrases, including deletions and missing terms related to eye laterality, lesion details, phase, or location (e.g., omitting "left" in a report describing findings in the left eye).
(b) Insertion: The unintentional addition of incorrect words or phrases, such as inappropriate terms, incorrect substitutions, or word confusions affecting eye laterality, lesion descriptions, phase, or location (e.g., inserting "right" instead of "left" eye).
(c) Misspelling: Spelling mistakes, including truncated or incorrectly spelled words related to eye laterality, lesion, phase, or location (e.g., "lft" instead of "left").

(d) Other Errors: Errors that do not fit into the above categories, including incorrect date entries, numbering mistakes for images or series, unit measurement errors (e.g., centimeters vs. millimeters), template errors, miswrites, and punctuation mistakes.

The error content specifically focused on four aspects: eye laterality, lesion, phase, and location to ensure critical clinical information was accurately captured. Only the errors intentionally introduced into the reports were used as the reference standard. To verify that no additional errors were present, the reports were reviewed by three independent readers (Y.X, D.K., and K.J.), who have more than 5 years of training experience. Any discrepancies identified during the review were resolved through consensus among all three reviewers.

Error Detection and Time Measurement

Six eye care professionals with varying levels of clinical experience participated in evaluating the reports. This group included two retinal specialists, each with over 10 years of experience (later named as Experts), two general ophthalmologists, each with more than 5 years of experience (later named as Ophthalmologists), and two ophthalmic technicians, each with over 5 years of experience (later named as Technicians). Each one reviewed the ophthalmic imaging reports to identify potential errors. To mitigate fatigue, reviewers evaluated reports in sessions of 30 reports each, with mandatory 10-min breaks between sessions. The time taken to evaluate each report was recorded using a digital stopwatch.

GPT-4o (https://openai.com/research/gpt-4o) was accessed through the OpenAI application programming interface (https://platform.openai.com/docs/api-reference/introduction). The study used a paid version of GPT-4o to ensure consistent access and performance. The reports were organized in Microsoft Excel (version 16.8) and individually processed using a Python script (version 3.11; Python Software Foundation) via the API.

Similar to the human readers, GPT-4o was tasked with identifying potential errors in each report using zero-shot prompting. This approach involves providing GPT-4o with a single instruction without iterative refinements or examples. The prompt used was: "In the following, I will provide you with an ophthalmology report, divided into a 'findings' and an 'impression' section. Please evaluate the report for mistakes and assess and validate the consistency between the 'findings' and the 'impression' sections, highlighting any discrepancies or notable points." GPT-4o flagged errors but did not automatically generate corrected reports, requiring human oversight to validate and finalize corrections.

The time required for GPT-4o to correct each report was measured by recording the duration from when the prompt was sent to when the final response was received. This measurement was conducted on 15 randomly selected reports of varying lengths.

Cost Analysis

For the cost analysis, the study referenced the 2021 Annual Salary Survey Report for Chinese Hospitals to determine the salaries based on predetermined monthly working hours. Salaries were extracted for each category of ophthalmologists, including senior ophthalmologists, attending physicians, and resident doctors. Costs were converted from Chinese Yuan (CNY) to U.S. dollars (USD) using the exchange rate as of January of 2025.

At the time of the study, the cost for GPT-4o was $0.06 per 1000 tokens for prompts and $0.12 per 1000 tokens for overall usage. The total cost of evaluating the reports was calculated based on the token usage for the 15 randomly selected reports analyzed. This calculation included both the cost of sending prompts to GPT-4o and processing the responses, ensuring an accurate estimation of the financial implications of integrating GPT-4o into the reporting workflow.

Real-World Evaluation

In addition to simulated error seeding, we also conducted a small real-world validation study.

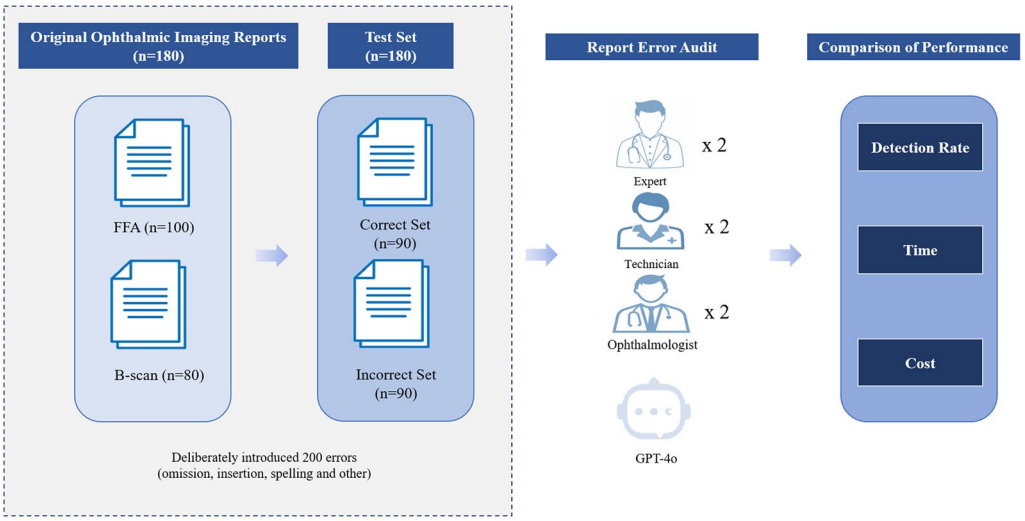

Fig. 1 Study flowchart. **A** Initially, 180 original ophthalmic imaging reports, including 100 fundus fluorescein angiography (FFA) reports and 80 ocular B-scan ultrasound reports, were selected. **B** These reports were randomized into two sets: a correct set and an incorrect set, each containing 90 reports. Within the incorrect set, 200 errors across four categories (omission, insertion, spelling errors, other errors) were deliberately introduced by an ophthalmology resident doctor, with a maximum of three errors per report. **C** GPT-4o and six eye care professionals were tasked with evaluating each report to identify potential errors, enabling a comparative analysis of their performance

We prospectively collected 20 ophthalmic diagnostic reports containing known errors (ten FFA and ten B-scan reports). Each report contained exactly one error, equally distributed across four categories: omission ($n=5$), insertion ($n=5$), spelling errors ($n=5$), and other errors ($n=5$). Reports were independently reviewed by GPT-4o, one retinal specialist, one general ophthalmologist, and one ophthalmic technician. The number of correctly identified errors, missed errors, and false positives was recorded for each rater.

Statistical Analysis

All statistical analyses were performed using IBM SPSS Statistics version 26.0 and Python version 3.11. The primary outcomes measured were the number of correctly detected errors and the time taken to process each report. The number of errors correctly identified by GPT-4o was compared to those identified by the ophthalmologists using Wald χ^2 tests. The Wilson method was employed to calculate 95% confidence intervals [28]. Analyses were conducted for all report types combined and separately for FFA and ocular B-scan reports.

The average time and cost for processing reports corrected by GPT-4o were compared with those of the human readers using paired-sample *t* tests. Bonferroni correction was applied to adjust for multiple comparisons. A two-sided *P* value of less than 0.05 was considered statistically significant. Cohen's *d* was used to measure effect sizes for differences in processing time, with values of 0.2, 0.5, and 0.8 representing small, medium, and large effects, respectively. Interrater reliability was assessed using Cohen's κ, categorized as follows: 0.01–0.20 (none to slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial), and 0.81–1.00 (almost perfect agreement). Due to the exploratory nature of the research, a power analysis was not performed initially. However, a post hoc power analysis was conducted to determine the sample size adequacy, assuming an effect size of 0.2, 80% power, and a 5% significance level, confirming that the sample of 180 reports was sufficient to detect significant differences in error detection rates.

FFA Report	Error and Error type(s)	Error Detection	
FFA Findings: Dot microaneurysms observed in the retina with retinal detachment noted on the temporal side Hemorrhage obscuring fluorescence Capillary dilation and leakage Significant leakage observable on the temporal side of the optic disc in the late stage **Impression:** Diabetic retinopathy in the right eye without retinal detachment	**Error Analysis:** The Impression section states that the right eye has diabetic retinopathy without retinal detachment. However, the FFA Findings do not specify the eye and indicate the presence of retinal detachment. **Error types:** Omission: laterality Insert: with vs without	**Error 1** Expert 1：yes Expert 2：yes Technician 1: yes Technician 2: yes Ophthalmologist 1: yes Ophthalmologist 2: yes GPT-4o: yes	**Error 2** Expert 1：no Expert 2：Yes Technician 1: yes Technician 2: yes Ophthalmologist 1: yes Ophthalmologist 2: no GPT-4o: yes
FFA Findings: Extensive capillary non-perfusion is observed in the mid-peripheral and peripheral retina of the right eye, presenting as large areas of non-fluorescence. Significant fluorescein exudation is noted around the optic disc, along with cystoid fluorescein accumulation in the macula. **Impression:** Non-Ischemic CRVO in the Right Eye	**Error Analysis:** The FFA finding describes extensive capillary non-perfusion and areas of non-fluorescence, suggesting ischemic CRVO, and " fluorescein exudation " is incorrect; it should be " fluorescein leakage" to accurately reflect the FFA findings. **Error types:** Insert: Ischemic vs Non-Ischemic Miswrite: exudation vs leakage	Expert 1：yes Expert 2：no Technician 1: yes Technician 2: yes Ophthalmologist 1: yes Ophthalmologist 2: no GPT-4o: yes	Expert 1：no Expert 2：yes Technician 1: yes Technician 2: no Ophthalmologist 1: yes Ophthalmologist 2: yes GPT-4o: no
B-scan Findings: Reticular high echoes are observed in the subretinal area on the temporal side of the rght eye, accompanied by localized retinal elevation. The edges of the detached region appear irregular, indicating a possible retinal tear. A small amount of echoes is visible in the vitreous, with no significant hemorrhage detected. Additionally, dynamic observation shows wave-like movements of the detached retina when the probe angle is adjusted. **Impression:** Right Eye Retinal Detachment	**Error Analysis:** In the Findings section, "rght" is a misspelling of "right." **Error types:** Misspelling: rght vs right	Expert 1：yes Expert 2：yes Technician 1: yes Technician 2: yes Ophthalmologist 1: yes Ophthalmologist 2: yes GPT-4o: yes	

◂**Fig. 2** Comparative proofreading examples by GPT-4o and the human readers illustrate incorrect ophthalmic reports with their respective errors and error types, alongside the corresponding proofreading outcomes. *FFA* fundus fluorescein angiography, *B-scan* ocular B-scan ultrasound

RESULTS

Performance in Detecting Errors in Ophthalmic Reports

In the comprehensive analysis, GPT-4o achieved a detection rate of 79.0% (158 of 200; 95% CI 73.0, 85.0) in total examinations, which was comparable to the average detection performance of Experts (84.0% [167.5 of 200; 95% CI 79.0, 89.0]), Technicians (86.0% [172 of 200; 95% CI 81.0, 91.0]), and Ophthalmologists (78.0% [155 of 200; 95% CI 72.0, 83.0]); *P* value range, 0.09–0.81. Notably, Technician 1 outperformed GPT-4o in Total examinations (88.0% [176 of 200; 95% CI 83.0, 93.0] vs 79.0%; $P=0.02$) (Figs. 2 and S1; Table 1).

Subgroup Analysis by Imaging Modality

In fundus fluorescein angiography (FFA), GPT-4o demonstrated a detection rate of 79.0% (100 of 126; 95% CI 72.0, 86.0), aligning with the average performance of Experts (86.0% [108 of 126; 95% CI 80.0, 92.0]), Technicians (87.0% [110 of 126; 95% CI 81.0, 93.0]), and Ophthalmologists (78.0% [98.5 of 126; 95% CI 71.0, 85.0]); *P* value range, 0.13–0.94. Technician 1 also significantly outperformed GPT-4o in FFA (90.0% [113 of 126; 95% CI 84.0, 95.0] vs 79.0%; $P=0.04$).

For B-scan ultrasonography (B-SCAN), GPT-4o attained a detection rate of 78.0% (58 of 74; 95% CI 78.0, 78.0), which was similar to the average detection rates of Experts (80.0% [59.5 of 74; 95% CI 80.0, 80.0]), Technicians (84.0% [62 of 74; 95% CI 84.0, 84.0]), and Ophthalmologists (76.0% [56.5 of 74; 95% CI 76.0, 76.0]); *P* value range, 0.53–0.92 (Table S1). Technician performance was also evaluated, with Technician 1 achieving a detection rate of 86% (64 of 74; 95% CI 78.0, 94.0; $P=0.28$ compared to GPT-4o), confirming their superior performance across modalities.

Error Categories

GPT-4o was less effective than the top-performing Expert in identifying omission errors (detection rate, 75% [51 of 68; 95% CI 65, 85] vs 83% [56.5 of 68; 95% CI 74, 92]; $P=0.34$) (Table S1). However, there was no significant difference between GPT-4o and the other Experts, Technicians, and Ophthalmologists in detecting omission errors (*P* value range, 0.28–0.40).

For misspelling errors, GPT-4o achieved a detection rate of 79% [49 of 62; 95% CI 69, 89], which was not significantly different from Ophthalmologists (79% [49 of 62; 95% CI 69, 89]; $P=1.00$) but was lower than Experts (83% [51.5 of 62; 95% CI 74, 92]; $P=0.73$) and Technicians (88% [54.5 of 62; 95% CI 80, 96]; $P=0.28$).

In detecting insertion errors, GPT-4o had a detection rate of 84% [42 of 50; 95% CI 74, 94], which was comparable to Experts (84% [42 of 50; 95% CI 74, 94]; $P=1.00$) but lower than Technicians (91% [45.5 of 50; 95% CI 83, 99]; $P=0.45$) and higher than Ophthalmologists (80% [40 of 50; 95% CI 69, 91]; $P=0.79$).

For other types of errors, GPT-4o achieved a detection rate of 80% [16 of 20; 95% CI 62, 98], which was similar to Technicians (80% [16 of 20; 95% CI 62, 98]; $P=1.00$) and Ophthalmologists (78% [15.5 of 20; 95% CI 59, 96]; $P=1.00$) but was lower than Experts (88% [17.5 of 20; 95% CI 73, 102]; $P=0.83$) (Table S1). Overall, there was no significant difference between GPT-4o and the other ophthalmologists in detecting errors across different report categories (*P* value range, 0.28–1.00) (Table S1).

GPT-4o demonstrated a false-positive error rate of 5.6% (10 of 180). When compared to human readers, GPT-4o's false-positive rate was significantly higher than that of Expert 2 (1.1% [2/180]; $P=0.04$*) and both Technician 1 and Technician 2 (1.1% [2/180] each; $P=0.04$*). The Average Technician also exhibited a significantly lower false-positive of 1.1% (2/180; $P=0.04$*) compared to GPT-4o. In contrast, the false-positive for Expert 1 (1.7% [3/180]; $P=0.09$), the

Table 1 Error detection rate analysis: GPT-4o versus ophthalmologists (*$P < 0.05$)

Reader	Total		FFA		B-SCAN	
	Detection rate (%)	*P*	Detection rate (%)	*P*	Detection rate (%)	*P*
Expert 1	86 (81–90) 171/200	0.12	87 (81–92) 109/126	0.18	84 (84–84) 62/74	0.53
Expert 2	82 (77–87) 164/200	0.53	85 (79–91) 107/126	0.32	77 (77–77) 57/74	> 0.99
Average expert	84 (79–89) 167.5/200	0.27	86 (80–92) 108/126	0.25	80 (80–80) 59.5/74	0.92
Technician 1	88 (83–93) 176/200	0.02*	90 (84–95) 113/126	0.04*	85 (85–85) 63/74	0.39
Technician 2	84 (79–89) 168/200	0.25	85 (79–91) 107/126	0.32	82 (82–83) 61/74	0.68
Average technician	86 (81–91) 172/200	0.09	87 (81–93) 110/126	0.13	84 (84–84) 62/74	0.53
Ophthalmologist 1	77 (71–83) 154/200	0.72	79 (71–86) 99/126	1	74 (74–74) 55/74	0.7
Ophthalmologist 2	78 (72–84) 156/200	0.9	78 (71–85) 98/126	0.88	78 (78–78) 58/74	> 0.99
Average ophthalmologist	78 (72–83) 155/20	0.81	78 (71–85) 98.5/126	0.94	76 (76–76) 56.5/74	0.92
GPT-4o	79 (73–85) 158/200		79 (72–86) 100/126		78 (78–78) 58/74	

Average Expert (1.4% [2.5/180]; $P=0.06$), Ophthalmologist 1 (2.2% [4/180]; $P=0.17$), Ophthalmologist 2 (1.7% [3/180]; $P=0.09$), and the Average Ophthalmologist (1.9% [3.5/180]; $P=0.13$) did not differ significantly from that of GPT-4o (Table S2). The higher false-positive rate of GPT-4o may be attributed to its sensitivity to subtle linguistic variations.

Interrater Agreement

Agreements among human raters ranged from slight to substantial, with most falling between fair and moderate. Notably, there was substantial agreement between the two ophthalmologists and between Expert 1 and Ophthalmologist 1, while lower agreements among certain experts and technicians highlighted variability in interrater reliability across different professional roles. GPT-4o exhibited fair to substantial agreement with human raters, showing the highest consistency with Expert 1 ($\kappa=0.57$) and both ophthalmologists ($\kappa=0.45$ and $\kappa=0.47$), whereas lower agreement levels with Expert 2 and the technicians suggest areas for improvement in GPT-4o's performance relative to these raters (Table 2).

Reading Time

GPT-4o was markedly faster in processing all 180 ophthalmic reports compared to human ophthalmologists. The total reading time for GPT-4o was approximately 0.27 h. In contrast, the fastest ophthalmologist completed the 180 reports in approximately 2.17 h, and the slowest took around 3.19 h (Fig. 3A). The average time taken by GPT-4o to review each ophthalmic report was significantly shorter than that of the quickest human ophthalmologist (mean reading time, 5.33 s ± 2.24 [SD] vs 41.96 s ± 7.64; $P<0.001$; Cohen's $d=-4.96$) (Fig. 3B; Table S3).

Cost Analysis

GPT-4o incurred significantly lower mean costs per ophthalmic report compared to all human readers. The mean cost per report for GPT-4o was $0.03 ± 0.01, whereas Experts averaged $0.24 ± 0.04, Technicians $0.16 ± 0.03, and Ophthalmologists $0.23 ± 0.05 (all $P<0.001$; Cohen d ranging from −3.13 to −5.95) (Table S3).

The estimated average cost for proofreading all 180 ophthalmic reports by the six human reviewers was $37.46. This includes an average

Table 2 Interrater reliability between GPT-4o and ophthalmologists

Reader	Expert 1	Expert 2	Technician 1	Technician 2	Ophthalmology 1	Ophthalmology 2	GPT-4o
Expert 1		0.51	0.50	0.28	0.59	0.45	0.57
Expert 2	0.51		0.34	0.08	0.36	0.25	0.36
Technician 1	0.50	0.34		0.50	0.46	0.37	0.32
Technician 2	0.28	0.08	0.50		0.24	0.19	0.21
Ophthalmologist 1	0.59	0.36	0.46	0.24		0.63	0.45
Ophthalmologist 2	0.45	0.25	0.37	0.19	0.63		0.47
GPT-4o	0.57	0.36	0.32	0.21	0.45	0.47	

cost of $43.07 per retinal specialists, $28.24 per ophthalmic technicians, and $41.07 per ophthalmologist.

In sharp contrast, GPT-4o incurred a total cost of only $4.8 for proofreading the same set of reports (Fig. 3C). On a per-report basis, the total estimated average cost for proofreading one ophthalmology report by the six human reviewers was $0.21. Specifically, the average cost per retinal specialists was $0.24, per ophthalmic technicians was $0.16, and per ophthalmologist was $0.23.

Comparatively, GPT-4o required only $0.03 per report. The mean cost per report for GPT-4o was significantly lower than that of the most economical human ophthalmologist (mean cost per report, $0.03 ± 0.01 vs $0.16 ± 0.03, respectively; $P < 0.001$; Cohen's $d = -5.26$) (Fig. 3D).

Real-World Evaluation

In this real-world validation set of 20 erroneous reports, GPT-4o detected 18 of 20 errors (sensitivity 90%) with no false positives. Both the retinal specialist and ophthalmic technician identified all 20 errors (100% sensitivity, no false positives), whereas the general ophthalmologist detected 19 errors (95% sensitivity) with no false positives (Table S4).

DISCUSSION

This study aimed to assess ChatGPT's effectiveness in identifying common errors and discrepancies in FFA and B-scan reports and to evaluate its potential in reducing both time and associated costs. To achieve this, ChatGPT's performance in detecting a set of predefined errors within a collection of ophthalmology reports was compared against that of six eye care professionals (two retinal specialists, two general ophthalmologists, and two ophthalmic technicians) with varying levels of experience.

Overall, GPT-4o's detection rates in Total examinations and FFA were comparable to those of Experts and Technicians, except Technician 1, who significantly outperformed the AI model in these modalities. In B-scan examinations, GPT-4o's performance was on par with all reader categories, including Experts, Technicians, and Ophthalmologists. These findings indicate that while GPT-4o performs similarly to most human readers, certain experienced technicians exhibit superior detection capabilities in specific diagnostic modalities. This outperformance by Technician 1 may be due to their extensive experience with report-specific terminology and error patterns. Additionally, GPT-4o demonstrated a significantly faster processing time than all human readers, highlighting its efficiency in handling report reviews. From a cost perspective,

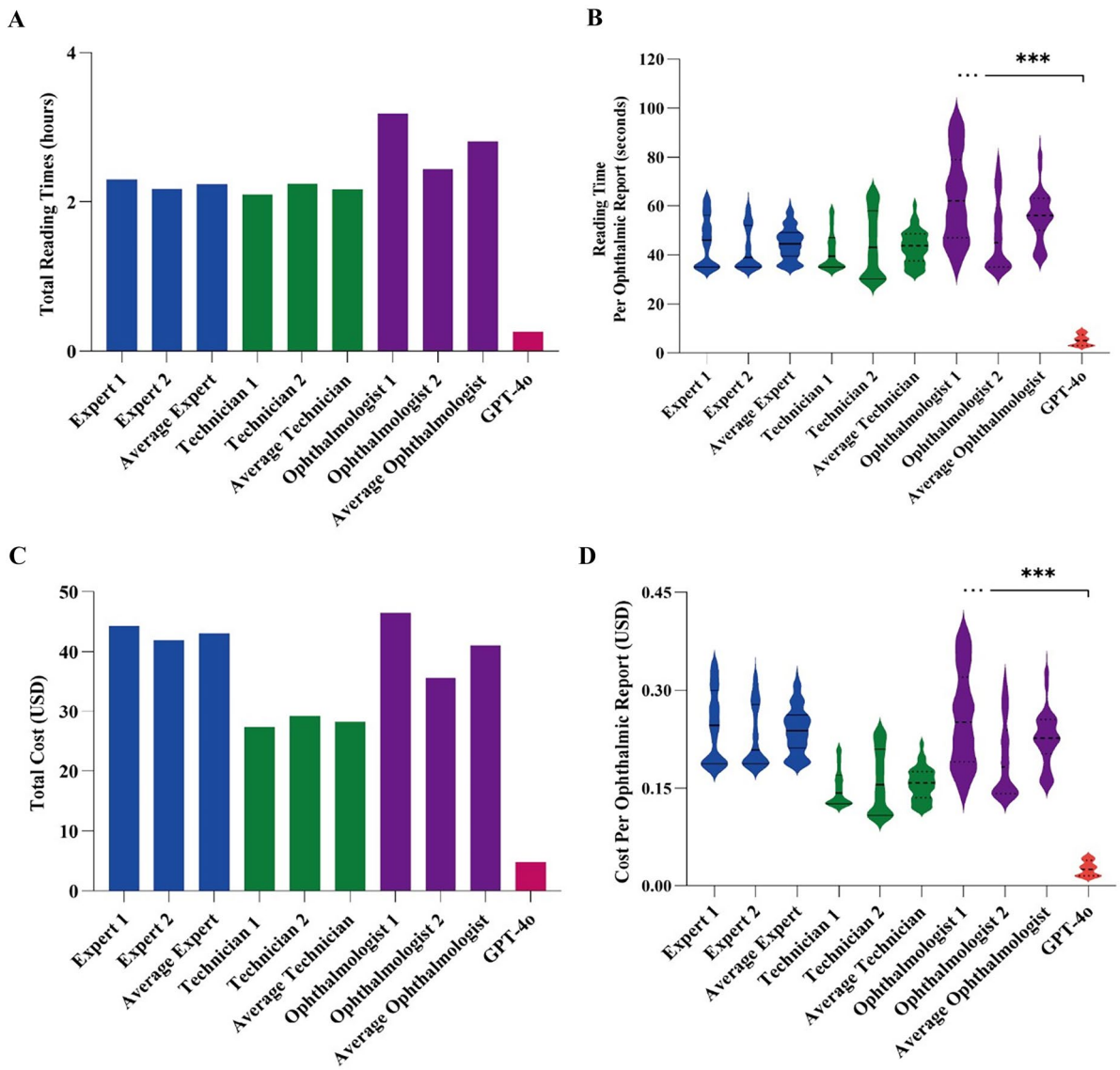

Fig. 3 **A** Bar graph displaying total reading time in seconds for GPT-4o and human readers. **B** Violin plot illustrating reading time per ophthalmic report in seconds. **C** Bar graph presenting total cost in U.S. dollars for GPT-4o and human readers. **D** Violin plot showing cost per ophthalmic report in U.S. dollars. *Dashed lines* represent medians and *dotted lines* indicate quartiles. ***$P < 0.001$

digital technology-driven hierarchical screening proved to be substantially more economical than human reviewers when considering both total and per-report expenses [20]. Our results complement recent evidence from bilingual ophthalmology reasoning tasks, where DeepSeek-R1 and other state-of-the-art LLMs demonstrated strong diagnostic and management accuracy, further highlighting the potential of advanced models to support complex clinical decision-making [21, 22].

GPT-4o demonstrated exceptional efficiency in both reading time and cost per ophthalmic report. The AI model processed all reports in a fraction of the time required by human ophthalmologists and did so at a substantially lower cost. These findings highlight GPT-4o's potential to significantly enhance workflow efficiency

and reduce operational expenses in clinical ophthalmology settings, making it a valuable tool for rapid and cost-effective report analysis. This underscores the potential for artificial intelligence to streamline ophthalmology workflows beyond the realm of image interpretation [23]. In terms of time efficiency, the findings are consistent with other research exploring the integration of large language models into ophthalmologic practices [24, 25]. When employed as a proofreading tool, GPT-4o not only maintains performance levels comparable to human reviewers but also offers significant cost savings, primarily due to reduced time investment and lower operational costs. Moreover, unlike human reviewers whose performance may decline under multitasking demands or during off-hours shifts, GPT-4o provides consistent and reliable performance [26]. As AI algorithm costs decrease, its cost-effectiveness will improve, making it increasingly viable for clinical use. This is especially true in regions with high labor costs, where GPT-4o can significantly reduce operational expenses by performing tasks consistently and efficiently. Additionally, GPT-4o can serve as an educational tool for training resident doctors by identifying common errors and providing real-time feedback. Our findings are consistent with recent work in medical laboratory reporting, where GPT models also demonstrated high accuracy in error detection and substantial agreement with senior experts, underscoring the broader applicability of LLMs across diverse clinical reporting domains [27]. Overall, the integration of AI in ophthalmology could streamline workflows, enhance productivity, and lower costs, particularly in high-cost labor environments.

This preliminary real-world validation highlights that ChatGPT-4o can achieve high sensitivity (90%) in detecting naturally occurring report errors across multiple modalities, with performance approaching that of human experts. Notably, no false positives were observed. However, the small sample size and restriction to single-error reports limit generalizability. Larger multi-center datasets will be necessary to better stratify performance across error types and to assess whether real-world error detection aligns with the simulated error seeding framework [28, 29].

However, this study is not without limitations. Firstly, it was conducted in an experimental setting using a predefined set of reports, some of which contained deliberately introduced errors. While these errors were based on common patterns identified in existing literature, the categorization may not encompass the full range of errors present in real-world ophthalmology reports. Therefore, the diversity and frequency of errors in actual clinical practice might differ from those represented in this study. The use of deliberately introduced errors limits the study's immediate applicability to real-world settings, where error patterns may be more varied and less predictable. Secondly, the assessment of time savings did not include a direct comparison between uncorrected and GPT-4o corrected reports to avoid potential bias from insufficient blinding of the reviewers. Thirdly, the cost analysis does not account for the additional expenses associated with integrating a large language model into clinical workflows, such as infrastructure and data security measures. These costs include server maintenance, electricity, and software licensing, which we estimate to be minimal but acknowledge as a limitation. Furthermore, the error detection capabilities of GPT-4o necessitate human oversight for validation, which is both a practical requirement and a legal obligation. GPT-4o flags errors but does not produce corrected reports, requiring human intervention to finalize corrections. Lastly, the experimental nature of the study may have introduced a Hawthorne effect, where participants altered their behavior due to awareness of being observed, potentially inflating the error detection rates compared to typical clinical settings [30]. To mitigate this, reviewers were not informed that their performance would be compared to AI until after the study. Additionally, the study did not explore the impact of domain-specific fine-tuning of GPT-4o or its seamless integration into existing clinical workflows, leaving these aspects for future research [31].

CONCLUSIONS

This study demonstrates that ChatGPT-4o is a highly efficient and cost-effective tool for auditing errors in ophthalmic imaging reports, achieving detection rates comparable to human reviewers while significantly reducing processing time and costs. Despite its higher false-positive rate compared to some human reviewers, its speed and cost advantages make it a promising adjunct tool in clinical workflows. By addressing common errors in FFA and B-scan reports, ChatGPT has the potential to enhance report accuracy, streamline workflows, and serve as an educational tool for residents resident doctors. However, its application in real-world settings requires further validation to ensure generalizability and seamless integration into clinical practice, which we plan to address in future studies.

ACKNOWLEDGEMENTS

We gratefully acknowledge the six eye care professionals, including two retinal specialists, two general ophthalmologists, and two ophthalmic technicians, for their invaluable contributions and expertise in evaluating the reports for this study. We thank the participants of the study.

Author Contributions. Yufeng Xu: Responsible for methodology, investigation, and original draft preparation. Daohuan Kang: Methodology, validation, and analysis. Danli Shi: Manuscript review and editing. Yih Chung Tham: Manuscript review and editing. Andrzej Grzybowski: Manuscript review and editing. Kai Jin: Led the conceptualization, manuscript review and editing, and secured funding.

Funding. This work was financially supported by the Natural Science Foundation of China (grant number 82201195). The journal's Rapid Service Fee was funded by the authors.

Data Availability. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflicts of Interest. Yufeng Xu, Daohuan Kang, Danli Shi, Yih Chung Tham, Andrzej Grzybowski, and Kai Jin have no conflicts of interest or competing interests to disclose. Andrzej Grzybowski is a Section Editor of Ophthalmology and Therapy. Andrzej Grzybowski was not involved in the selection of peer reviewers for the manuscript nor any of the subsequent editorial decisions.

Ethical Approval. This retrospective study was approved by the ethics committee (Approval No. Y2023–1073), and the requirement for informed consent was waived due to its retrospective nature. All data provided to GPT-4o were anonymized, ensuring that no patient-identifying information was included. The study was conducted in accordance with the Declaration of Helsinki (1964) and its later amendments. No identifiable participant information was included in the study, and thus, specific participant consent for publication was not required.

REFERENCES

1. Arrigo A, Aragona E, Battaglia Parodi M, Bandello F. Quantitative approaches in multimodal fundus imaging: state of the art and future perspectives. Prog Retin Eye Res. 2023;92:101111.

2. Kadakia A, Zhang J, Yao X. Ultrasound in ocular oncology: technical advances, clinical applications, and limitations. Exp Biol Med. 2023;248(5):371–9.

3. Ilginis T, Clarke J, Patel PJ. Ophthalmic imaging. Br Med Bull. 2014;111(1):77–88.

4. White SJ, Phua QS, Lu L, Yaxley KL, McInnes MDF, To MS. Heterogeneity in systematic reviews of medical imaging diagnostic test accuracy studies: a systematic review. JAMA Netw Open. 2024;7(2):e240649.

5. Jin K, Pan X, You K, Wu J, Liu Z, Cao J, et al. Automatic detection of non-perfusion areas in diabetic macular edema from fundus fluorescein angiography for decision making using deep learning. Sci Rep. 2020;10(1):15138.

6. Lexa FJ, Jha S. Artificial intelligence for image interpretation: counterpoint-the radiologist's incremental foe. AJR Am J Roentgenol. 2021;217(3):558–9.

7. Lee CS, Nagy PG, Weaver SJ, Newman-Toker DE. Cognitive and system factors contributing to diagnostic errors in radiology. AJR Am J Roentgenol. 2013;201(3):611–7.

8. Brady AP. Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging. 2017;8(1):171–82.

9. Sheng B, Guan Z, Lim LL, Jiang Z, Mathioudakis N, Li J, et al. Large language models for diabetes care: potentials and prospects. Sci Bull. 2024;69(5):583–8.

10. Jin K, Ye J. Artificial intelligence and deep learning in ophthalmology: current status and future perspectives. Adv Ophthalmol Pract Res. 2022;2(3):100078.

11. Parillo M, Vaccarino F, Beomonte Zobel B, Mallio CA. ChatGPT and radiology report: potential applications and limitations. Radiol Med. 2024;129(12):1849–63.

12. Soleimani M, Seyyedi N, Ayyoubzadeh SM, Kalhori SRN, Keshavarz H. Practical evaluation of ChatGPT performance for radiology report generation. Acad Radiol. 2024;31(12):4823–32.

13. Rundle CW, Szeto MD, Presley CL, Shahwan KT, Carr DR. Analysis of ChatGPT-generated differential diagnoses in response to physical exam findings for benign and malignant cutaneous neoplasms. J Am Acad Dermatol. 2024;90(3):615–6.

14. Sun SH, Huynh K, Cortes G, Hill R, Tran J, Yeh L, et al. Testing the ability and limitations of ChatGPT to generate differential diagnoses from transcribed radiologic findings. Radiology. 2024;313(1):e232346.

15. Yu T, Shao A, Wu H, Su Z, Shen W, Zhou J, et al. A systematic review of advances in AI-assisted analysis of fundus fluorescein angiography (FFA) images: from detection to report generation. Ophthalmol Ther. 2025;14(4):599–619.

16. Chen X, Zhang W, Xu P, Zhao Z, Zheng Y, Shi D, et al. Ffa-gpt: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. NPJ Digit Med. 2024;7(1):111.

17. Shao A, Liu X, Shen W, Li Y, Wu H, Pan X, et al. Generative artificial intelligence for fundus fluorescein angiography interpretation and human expert evaluation. NPJ Digit Med. 2025;8(1):396.

18. Wu H, Jin K, Yip CC, Koh V, Ye J. A systematic review of economic evaluation of artificial intelligence-based screening for eye diseases: from possibility to reality. Surv Ophthalmol. 2024;69(4):499–507.

19. Guerra GA, Hofmann HL, Le JL, Wong AM, Fathi A, Mayfield CK, et al. ChatGPT, Bard, and Bing Chat are large language processing models that answered orthopaedic in-training examination questions with similar accuracy to first-year orthopaedic surgery residents. Arthrosc J Arthrosc Relat Surg. 2025;41(3):557–62.

20. Wu X, Wu Y, Tu Z, Cao Z, Xu M, Xiang Y, et al. Cost-effectiveness and cost-utility of a digital technology-driven hierarchical healthcare screening pattern in China. Nat Commun. 2024;15(1):3650.

21. Xu P, Wu Y, Jin K, Chen X, He M, Shi D. Deepseek-r1 outperforms Gemini 2.0 Pro, OpenAI o1, and o3-mini in bilingual complex ophthalmology reasoning. Adv Ophthalmol Pract Res. 2025;5(3):189–95.

22. Srinivasan S, Ai X, Zou M, Zou K, Kim H, Lo TWS, et al. Ophthalmological question answering and reasoning using OpenAI o1 vs other large language models. JAMA Ophthalmol. 2025. https://doi.org/10.1001/jamaophthalmol.2025.2413.

23. Gao Z, Pan X, Shao J, Jiang X, Su Z, Jin K, et al. Automatic interpretation and clinical evaluation

for fundus fluorescein angiography images of diabetic retinopathy patients by deep learning. Br J Ophthalmol. 2023;107(12):1852–8.

24. Liu X, Wu J, Shao A, Shen W, Ye P, Wang Y, et al. Uncovering language disparity of ChatGPT on retinal vascular disease classification: cross-sectional study. J Med Internet Res. 2024;26:e51926.

25. Wu H, Su Z, Pan X, Shao A, Xu Y, Wang Y, et al. Enhancing diabetic retinopathy query responses: assessing large language model in ophthalmology. Br J Ophthalmol. 2025. https://doi.org/10.1136/bjo-2024-325861.

26. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open. 2023;6(8):e2330320.

27. Han W, Wan C, Shan R, Xu X, Chen G, Zhou W, et al. Evaluation of error detection and treatment recommendations in nucleic acid test reports using ChatGPT models. Clin Chem Lab Med (CCLM). 2025;63(9):1698–708.

28. Grzybowski A, Jin K, Wu H. Challenges of artificial intelligence in medicine and dermatology. Clin Dermatol. 2024;42(3):210–5.

29. Jin K, Yuan L, Wu H, Grzybowski A, Ye J. Exploring large language model for next generation of artificial intelligence in ophthalmology. Front Med Lausanne. 2023;10:1291404.

30. Benedetti F, Carlino E, Piedimonte A. Increasing uncertainty in CNS clinical trials: the role of placebo, nocebo, and Hawthorne effects. Lancet Neurol. 2016;15(7):736–47.

31. Betzler BK, Chen H, Cheng CY, Lee CS, Ning G, Song SJ, et al. Large language models and their impact in ophthalmology. Lancet Digit Health. 2023;5(12):e917–24.