Original Article

narra j

Original Article

Chinese generative AI models (DeepSeek and Qwen) rival ChatGPT-4 in ophthalmology queries with excellent performance in Arabic and English

Malik Sallam[1,2*], Israa M. Alasfoor[3,4], Shahad W. Khalid[3,4], Rand I. Al-Mulla[3,4], Amwaj Al-Farajat[3,4], Maad M. Mijwil[5,6], Reem Zahrawi[7], Mohammed Sallam[8,9,10,11], Jan Egger[12,13,14,15] and Ahmad S. Al-Adwan[16]

[1]Department of Pathology, Microbiology and Forensic Medicine, School of Medicine, The University of Jordan, Amman, Jordan; [2]Department of Clinical Laboratories and Forensic Medicine, Jordan University Hospital, Amman, Jordan; [3]Section of Ophthalmology, Department of Special Surgery, School of Medicine, The University of Jordan, Amman, Jordan; [4]Section of Ophthalmology, Department of Special Surgery, Jordan University Hospital, Amman, Jordan; [5]College of Administration and Economics, Al-Iraqia University, Baghdad, Iraq; [6]Department of Computer Techniques Engineering, Baghdad College of Economic Sciences University, Baghdad, Iraq; [7]Department of Ophthalmology, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai, United Arab Emirates; [8]Department of Pharmacy, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai, United Arab Emirates; [9]Department of Management, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai, United Arab Emirates; [10]Department of Management, School of Business, International American University, Los Angeles, United States; [11]College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU), Dubai, United Arab Emirates; [12]Institute for Artificial Intelligence in Medicine (IKIM), Essen University Hospital (AöR), Girardetstraße, Germany; [13]Center for Virtual and Extended Reality in Medicine (ZvRM), Essen University Hospital (AöR), Hufelandstraße, Germany; [14]Cancer Research Center Cologne Essen (CCCE), University Medicine Essen (AöR), Hufelandstraße, Germany; [15]University of Duisburg-Essen, Faculty of Computer Science, Schützenbahn, Germany; [16]Department of Business Technology, Al-Ahliyya Amman University, Amman, Jordan

*Corresponding author: malik.sallam@ju.edu.jo

Abstract

The rapid evolution of generative artificial intelligence (genAI) has ushered in a new era of digital medical consultations, with patients turning to AI-driven tools for guidance. The emergence of Chinese-developed genAI models such as DeepSeek-R1 and Qwen-2.5 presented a challenge to the dominance of OpenAI's ChatGPT. The aim of this study was to benchmark the performance of Chinese genAI models against ChatGPT-4o and to assess disparities in performance across English and Arabic. Following the METRICS checklist for genAI evaluation, Qwen-2.5, DeepSeek-R1, and ChatGPT-4o were assessed for completeness, accuracy, and relevance using the CLEAR tool in common patient ophthalmology queries. In English, Qwen-2.5 demonstrated the highest overall performance (CLEAR score: 4.43±0.28), outperforming both DeepSeek-R1 (4.31±0.43) and ChatGPT-4o (4.14±0.41), with p=0.002. A similar hierarchy emerged in Arabic, with Qwen-2.5 again leading (4.40±0.29), followed by DeepSeek-R1 (4.20±0.49) and ChatGPT-4o (4.14±0.41), with p=0.007. Each tested genAI model exhibited near-identical performance across the two languages, with ChatGPT-4o demonstrating the most balanced linguistic capabilities (p=0.957), while Qwen-2.5 and DeepSeek-R1 showed a marginal superiority for English. An in-depth examination of genAI performance across key CLEAR components revealed that Qwen-2.5 consistently excelled in content completeness, factual accuracy, and relevance in both English and Arabic, setting a new benchmark for genAI in medical inquiries. Despite minor linguistic disparities, all three models exhibited robust multilingual capabilities, challenging the long-held assumption that genAI is inherently biased toward English. These findings highlight the evolving nature of AI-driven medical assistance, with Chinese genAI models being able to rival or even surpass ChatGPT-4o in ophthalmology-related queries.

Keywords: LLM, OpenAI, DeepSeek, Qwen, eye disease

Received: March 1, 2025 - Accepted: April 6, 2025 - Published online: April 8, 2025

Sallam et al. Narra J 2025; 5 (1): e2371 - http://doi.org/10.52225/narra.v5i1.2371

Introduction

In the year 2023, the abbreviation "AI" standing for artificial intelligence has been named the Collins Word of the Year [1]. This recent recognition is indicative of the far-reaching influence of generative AI (genAI) tools that have overthrown long-standing paradigms in various aspects of human activities [2-4]. Nowhere is this transformation more divisive than in the healthcare sector, where genAI's potential to augment medical education and clinical decision-making is met with equal measures of enthusiasm and apprehension [5-9]. The genAI models are characterized by a remarkable ability to synthesize, contextualize, and even infer knowledge with striking human-like sophistication [5,9,10]. OpenAI's ChatGPT models have taken the pole position of this technological revolution, with Generative Pre-training Transformer (GPT)-4 ascending to global prominence as a pioneering genAI model [5,9,11]. ChatGPT dominated applications ranging from casual conversation to complex problem-solving across various professional domains, such as healthcare [12-15]. Yet, as with all transformative technologies, dominance invites competition [16].

The emergence of Chinese-developed genAI models, particularly Qwen-2.5 (Alibaba) and DeepSeek-R1 (DeepSeek), has injected a new and formidable challenge into the genAI landscape [6,17-19]. In the rapidly evolving field of genAI, OpenAI's early lead raises the critical issue of whether its dominance will persist or if emerging challengers are poised to surpass it, particularly in the healthcare domain. Benchmarking demonstrated that DeepSeek-V3 consistently outperformed genAI models such as Llama 3.1 across various standard evaluations, particularly excelling in multilingual comprehension and coding [6]. Notably, DeepSeek-V3 matched OpenAI's GPT-4o models in accuracy, language generation quality, and specialized domains like medicine, despite being developed at a significantly lower cost [6].

The AI tools once seemed like a novelty; however, genAI models have rapidly become an indispensable tool in healthcare [7,20,21]. Recent evidence indicated that health professionals, educators, students, and patients are using genAI tools in their daily practices at a rapidly evolving pace [22-26]. Early genAI adopters have begun using ChatGPT and similar models to refine differential diagnoses, improve clinical workflow, and alleviate the administrative tasks burden [27-29]. Soon, healthcare practice will not merely adapt to the presence of genAI; it will find itself increasingly reliant on it, tightly woven into the fabric of patient care [21]. Nowhere is this shift more evident than in patient behavior. Patients, once bound by the constraints of scheduled consultations and the unpredictability of overburdened healthcare systems, are increasingly turning to genAI for guidance [30,31]. The trend is clear: patients now expect, and might soon prefer, immediate AI-generated medical insights over traditional healthcare encounters [23].

For much of its early dominance, OpenAI's ChatGPT models had a largely unchallenged position in genAI field. This dominance was manifested in ChatGPT's superior performance in different benchmarking studies compared to its Western genAI counterparts, such as Microsoft Copilot and Google Gemini, in different healthcare fields [32-35]. Therefore, the Western-developed genAI models, while notable, have not fundamentally altered OpenAI pioneering steps. The true challenge to OpenAI's dominance in genAI has not, as conventional wisdom might suggest, emerged from the familiar corridors of Silicon Valley, but rather from China [6,17,18]. The introduction of Chinese genAI models such as DeepSeek-R1 and Alibaba's Qwen-2.5 marked an escalation in genAI competition and a fundamental shift in genAI development capability [6,17,18]. Unlike the West's predominantly market-driven approach, China's AI strategy is distinguished by its sheer scale, centralized coordination, and an unambiguous national mandate to achieve technological advantage [36]. With state-backed funding, nationalized AI initiatives, and an aggressive push toward digital sovereignty, China has not merely entered the genAI race but has positioned itself as a formidable competitor at the global level [37]. The Chinese approach, in many respects, redefined genAI race trajectory by establishing an AI environment that now stands as a formidable rival to Western AI conglomerates [38]. The introduction of Qwen-2.5 and DeepSeek-R1 reflects the Chinese genAI development approach. These models are not mere replicas of ChatGPT but are designed to outperform it, which has been shown in recent benchmarking studies [6,19].

Among the many healthcare areas in which genAI has demonstrated utility, ophthalmology stands as a particularly compelling example [39-45]. Visual conditions are common health concerns worldwide, with disorders such as distance and near vision impairments, cataracts, glaucoma, and diabetic retinopathy affecting hundreds of millions of people [46-50]. The diagnostic process for many ophthalmology conditions is often dependent on pattern recognition which is an area where genAI models excel [51-53].

Recent studies which evaluated genAI's ability to provide guidance on eye diseases has shown that it can generate medically relevant responses, although with notable inconsistencies in factual accuracy [54,55]. For example, a study showed that ChatGPT matched or outperformed senior ophthalmology residents in diagnosing primary and secondary glaucoma from clinical cases [45]. Another study reported that ChatGPT-4 could provide rapid and safe medical guidance to patients inquiring about blepharoplasty, reinforcing its role in patient education [56]. Moreover, a study demonstrated that ChatGPT outperformed Google Bard in accuracy, empathy, and medical relevance, in medical eye conditions, postoperative healing, and medications [57]. Furthermore, a previous study found that ChatGPT provided high-quality information on myopia, although occasional inaccuracies highlight the need for careful evaluation [58].

It is important at this point to show that the utility of genAI into ophthalmology can extend beyond diagnostics, encompassing patient education and postoperative care instructions [59]. With many ophthalmologic conditions requiring long-term management, genAI models can be used to provide structured, step-by-step guidance on disease progression, treatment options, and surgical risks [60-62]. However, the effectiveness of genAI in ophthalmology among other health fields remains constrained by critical factors such as linguistic performance and training data biases [63,64]. Previous studies have revealed that ChatGPT's medical accuracy is markedly superior in English compared to other languages, such as Arabic [33,65,66], Chinese [67], French [68], Japanese [69], and Spanish [70]. This raises a critical concern regarding the reliability of genAI-generated medical information for the majority of the world's population that does not speak English as a first language.

It is known that the performance of genAI is inextricably linked to the datasets on which it is trained, with a majority of Western genAI models being trained heavily on English-language content [71]. While this approach has resulted in high-quality outputs in English, it has also led to notable deficits in non-English responses, with previous studies documenting reduced accuracy and completeness of Western genAI contents [72,73]. This linguistic bias is particularly consequential in healthcare, where inaccuracy can have grave consequences. Specifically, linguistic disparities in genAI performance could negatively impact patient safety and result in decreased quality of healthcare delivery [74,75]. Therefore, until rigorous benchmarking studies comprehensively assess genAI performance across diverse healthcare domains and languages, assertions of linguistic disparities remain largely speculative—a hypothesis awaiting empirical scrutiny [76].

The aim of this study was to address two key questions. First, can Chinese genAI models rival or surpass ChatGPT-4o in ophthalmology queries? Once the gold standard, ChatGPT now faces growing competition. By systematically evaluating Qwen-2.5, DeepSeek-R1, and ChatGPT-4o, we aimed to assess whether Chinese-developed AI has reached—or exceeded—parity in structured medical inquiries. Second, do genAI models exhibit language-based disparities? Previous research suggested that ChatGPT performs better in English than non-English languages including Arabic [33,65-67,77-79]. By comparing genAI responses in English and Arabic ophthalmology queries, this study aimed to examine whether the long-standing linguistic imbalance persists or whether newer genAI models have closed the gap.

Methods

Study design

This study was designed as a structured genAI performance evaluation following the METRICS checklist for evaluation of genAI models in healthcare [80]. GenAI model configurations, evaluation methodologies, prompt structures, language considerations (English and Arabic), and data source transparency were systematically documented. Query selection prioritized clinical relevance, with randomization applied where appropriate to minimize bias. GenAI-generated responses were assessed for completeness, factual accuracy, and relevance using the CLEAR tool for evaluation of genAI generated content, with interrater reliability measures to reduce subjectivity [81].

Features of the genAI models tested and prompting approach

This study evaluated the performance of three genAI models: ChatGPT-4o (OpenAI, paid version), DeepSeek-R1 (free version), and Qwen-2.5 (Alibaba Cloud, free version) [82-84]. ChatGPT-4o is OpenAI's advanced publicly available genAI model, accessible for free with limitations and fully unlocked via a paid subscription, whereas DeepSeek-R1 and Qwen-2.5 are freely available genAI models developed in China.

To ensure replicability and consistency, all three studied genAI models were tested under their default configurations, as provided by their respective platforms at the time of evaluation. No custom settings, Application Programming Interface (API) modifications, or manual adjustments were applied to prevent any variability arising from non-standard parameters. A standardized prompting procedure was implemented, with all queries executed on February 21, 2025, within a controlled timeframe to reduce performance fluctuations due to genAI models' updates or server-side optimizations. The same set of ophthalmology-related queries was submitted to each genAI model by three authors (IMA, SWK and RIA-M). Each genAI model's responses were recorded in full for subsequent evaluation.

Each genAI model was prompted using the exact ophthalmology-related queries without modification or follow-up feedback. To maintain the integrity of first-response accuracy, the "New Chat" feature was used for each query to ensure that every response was generated in isolation from previous interactions. Additionally, while the accounts had prior usage history, the tested genAI models do not retain past interactions across sessions, ensuring that all responses were generated without influence from previous queries.

To eliminate context carryover effects between languages, the same query was executed in separate, independent sessions when switching between English and Arabic. This process ensured that responses were generated without influence from prior answers, preventing genAI models from utilizing previous outputs to refine subsequent responses. Additionally, the "Regenerate Response" feature and Feedback feature were not used, preserving the spontaneity and reproducibility of the initial outputs.

Sample size of the queries used to test the genAI models

The selection of minimum sample size of the ophthalmology-related queries was based on a statistical power analysis to ensure sufficient sensitivity in detecting performance differences between genAI models across English and Arabic responses. Sample size estimation was conducted using Statulator, employing a two-sided significance level of 5% ($\alpha=0.05$), a power of 80% ($\beta=0.20$), and an expected effect size of 0.5 [85]. This calculation determined that a minimum of 34 pairs of queries (each presented in both Arabic and English) was required to achieve adequate statistical power.

To strengthen the study's reliability, robustness, and generalizability, the final query set was increased to 42 pairs, allowing for improved precision in estimating differences, accommodating potential variability in AI-generated responses, and reducing the risk of underpowering secondary analyses.

Formulation of ophthalmology queries

The selection of ophthalmology-related queries was conducted through a collaborative, expert-driven process involving five authors with expertise in ophthalmology, including a consultant and full professor of ophthalmology with over 30 years of clinical experience (RZ), as well as three senior ophthalmology residents (IMA, SWK, and RIA-M) and one junior resident (AA-F), all of whom are bilingual in Arabic and English. Collectively, the residents contributed a cumulative 130 months of ophthalmology training, ensuring that the queries reflected real-world clinical scenarios encountered in ophthalmology practice.

The queries were systematically designed to cover the most frequently encountered ophthalmologic conditions in outpatient and inpatient settings. The topics were classified into predefined categories to ensure comprehensive representation across four key domains of ophthalmology: refractive surgery (n=9), cataract (n=13), glaucoma (n=9), and eye infections (n=10) (**Underlying data**).

Following initial formulation in Arabic, the queries underwent a rigorous translation and validation process. They were first translated into English and then back-translated into Arabic to verify linguistic and conceptual consistency across both languages.

Any discrepancies identified during back-translation were discussed between two authors (IMA and SWK), leading to minor refinements for clarity and standardization. This approach ensured that the queries were equivalent in meaning, clinically relevant, and methodologically robust for cross-linguistic performance assessment of the genAI models.

Evaluation of genAI generated content

The evaluation of the AI-generated content was conducted independently by three authors (IMA, ASWK, and RIA-M) without blinding in relation to the genAI model being evaluated. To minimize subjectivity in the evaluation process, a consensus key response was formulated prior to assessment.

The evaluation was based on the CLEAR tool [81], modified and divided into three dimensions as follows: completeness (Is the content sufficient?); accuracy (Is the content accurate and evidence-based?); and relevance (Is the content clear, concise, easy to understand, and free from irrelevant information?). Each component was assessed using a 5-point Likert scale ranging from 5 (excellent) to 1 (poor) [81].

Statistical and data analyses

Statistical analyses were conducted using IBM SPSS Statistics for Windows, Version 26 (IBM, New York, USA), with a two-sided significance level set at p-value < 0.05 to determine statistical significance. The primary outcome measure was the average CLEAR score across three independent raters, encompassing both overall and component-specific scores for completeness, accuracy, and relevance [81].

To assess the consistency and reliability of AI-generated responses across multiple raters, intraclass correlation coefficient (ICC) was calculated at 0.665 for the average measures of the three CLEAR dimensions, which is indicative of moderate reliability based on previous study [86]. Comparisons of genAI model performance across the three tested models (ChatGPT-4o, DeepSeek-R1, and Qwen-2.5) were performed using Kruskal-Wallis (K-W) tests, a non-parametric method selected due to the non-normal distribution of data, as indicated by the Shapiro-Wilk test (p<0.001).

Pairwise post-hoc comparisons were conducted to evaluate language-based disparities. GenAI performance in English vs. Arabic was analyzed separately for each model using Mann-Whitney U (M-W) tests, a non-parametric alternative to the independent samples t-test due to non-normality of distribution. Effect sizes were reported using Cohen's d to quantify the magnitude of language-based performance disparities [87].

Results

General performance of the tested genAI models in ophthalmology

Across both English and Arabic queries, Qwen-2.5 demonstrated the highest overall performance, followed by DeepSeek-R1, with ChatGPT-4o consistently ranking lowest, albeit all were ranked as excellent based on the CLEAR scores' interpretations. In English, Qwen-2.5 achieved an average CLEAR score of 4.43±0.28, outperforming DeepSeek-R1 (4.31±0.43) and ChatGPT-4o (4.14±0.41), with a statistically significant difference (p=0.002) (**Figure 1**). A similar trend was observed in Arabic, where Qwen-2.5 scored 4.40±0.29, ahead of DeepSeek-R1 (4.20±0.49) and ChatGPT-4o (4.14±0.41), with statistical significance of p=0.007 (**Figure 1**).

Despite these differences across the three tested genAI models, language-based comparisons within each genAI model revealed no statistically significant discrepancies. Qwen-2.5 performed slightly better in English than in Arabic (p=0.447), DeepSeek-R1 exhibited a similar trend

(p=0.239), while ChatGPT-4o showed virtually identical performance across both languages (p=0.957).

Effect size calculations further supported this finding. ChatGPT-4o demonstrated no meaningful performance gap between English and Arabic (t=−0.042; p=0.966; Cohen's d=0.0093). Qwen-2.5 showed only a small advantage in English (t=0.696; p=0.488; Cohen's d=0.1537), while DeepSeek-R1 exhibited a slightly more pronounced but still minor preference for English (t=1.061; p=0.292; Cohen's d=0.2343).

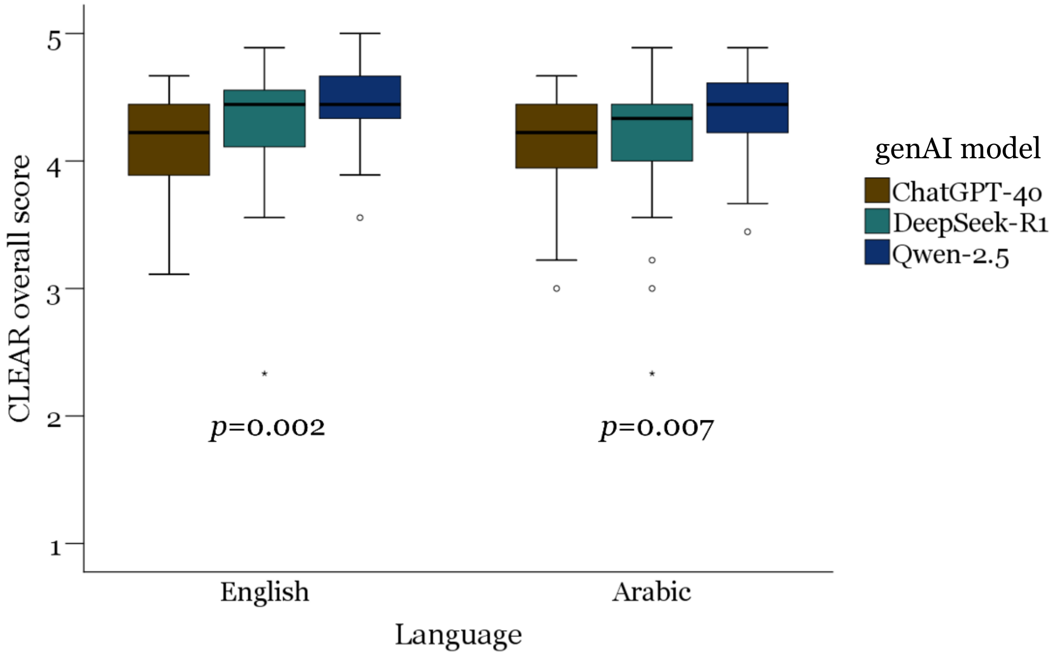

Figure 1. Comparison of generative AI (genAI) model performance in English and Arabic using CLEAR overall scores. The p-values were calculated using Kruskal-Wallis test.

CLEAR component-specific evaluation in ophthalmology per genAI model

Breaking down performance by CLEAR dimensions (completeness, accuracy, and relevance), Qwen-2.5 consistently achieved the highest scores across both languages, followed by DeepSeek-R1, with ChatGPT-4o ranking lowest (**Table 1**). Specifically, Qwen-2.5 had the strongest performance in completeness, accuracy, and relevance, with no statistically significant differences between the two languages. DeepSeek-R1 followed, demonstrating slightly lower scores across dimensions while maintaining comparable accuracy. ChatGPT-4o showed the lowest scores overall, particularly in completeness, though its accuracy scores remained relatively stable across the two languages (**Table 1**).

Table 1. Comparison of generative AI (genAI) model performance across English and Arabic using the three CLEAR evaluation dimensions

genAI model	Language	CLEAR overall score	Completeness score	Accuracy score	Relevance score
		Mean±SD	Mean±SD	Mean±SD	Mean±SD
ChatGPT-4o	English	4.14±0.41	4.11±0.48	4.16±0.58	4.15±0.39
	Arabic	4.14±0.41	4.04±0.45	4.22±0.55	4.17±0.43
p-value[a]		0.957	0.454	0.585	0.700
DeepSeek-R1	English	4.31±0.43	4.29±0.56	4.33±0.57	4.31±0.34
	Arabic	4.20±0.49	4.12±0.66	4.20±0.60	4.28±0.38
p-value[a]		0.239	0.133	0.221	0.810
Qwen-2.5	English	4.44±0.28	4.44±0.44	4.38±0.40	4.50±0.23
	Arabic	4.40±0.29	4.36±0.45	4.37±0.42	4.45±0.24
p-value[a]		0.477	0.350	0.868	0.438

[a]Analyzed using Mann-Whitney test

Comparative analysis of genAI models' performance per CLEAR components across combined languages

While no statistically significant language-based disparities were observed for any of the three genAI models, differences in overall performance between the three models were prominent when results from both languages were combined (**Table 2**). Qwen-2.5 consistently outperformed DeepSeek-R1 and ChatGPT-4o across all evaluation metrics, with significantly higher overall CLEAR scores (p<0.001) (**Table 2**). While accuracy domain did not differ significantly across models (p=0.147), Qwen-2.5 maintained the highest mean scores (4.38±0.41), followed by DeepSeek-R1 (4.26±0.58) and ChatGPT-4o (4.19±0.56). In terms of content relevance, Qwen-2.5 significantly outperformed both DeepSeek-R1 and ChatGPT-4o (p<0.001), achieving the highest scores (4.47±0.23), followed by DeepSeek-R1 (4.29±0.36) and ChatGPT-4o (4.16±0.41) (**Table 2**).

Table 2. Comparison of generative AI (genAI) model overall performance per CLEAR dimensions, for combined languages (English and Arabic)

genAI model	ChatGPT-4o	DeepSeek-R1	Qwen-2.5	p-value[a]
	Mean±SD	Mean±SD	Mean±SD	
Overall CLEAR	4.14±0.41	4.25±0.46	4.42±0.29	<0.001*
Completeness	4.07±0.46	4.21±0.62	4.40±0.45	<0.001*
Accuracy	4.19±0.56	4.26±0.58	4.38±0.41	0.147
Relevance	4.16±0.41	4.29±0.36	4.47±0.23	<0.001*

[a]Analyzed using Kruskal Wallis test
*Statistically significant at p=0.05

Comparative analysis of genAI models' overall performance across combined languages per ophthalmology topic

The overall performance of ChatGPT-4o, DeepSeek-R1, and Qwen-2.5 was then evaluated across four ophthalmology query topics: refractive surgery, cataract, glaucoma, and eye infections. Kruskal-Wallis analysis revealed statistically significant differences in genAI performance for refractive surgery (p<0.001) and eye infections (p=0.015), with better performance noted for Qwen-2.5 (**Figure 2**). No significant differences were observed for cataract (p=0.424) and glaucoma (p=0.347) (**Figure 2**).

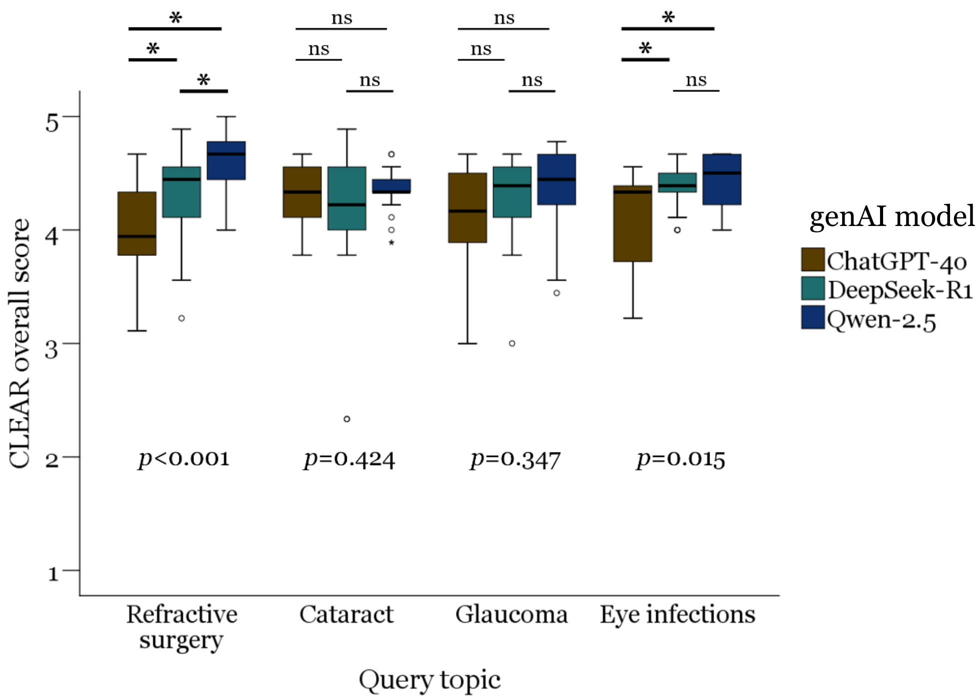

Figure 2. Comparison of generative AI (genAI) model performance across ophthalmology query topics. p-values were calculated using Kruskal-Wallis test. Post-hoc analysis results using Mann-Whitney U tests are indicated by the horizontal lines between genAI models, with significant results indicated by asterisk, while statistically insignificant results are indicated by ns.

Post-hoc Mann-Whitney *U* tests demonstrated that for refractive surgery, Qwen-2.5 significantly outperformed both DeepSeek-R1 (U=78.0; Z=−2.68; p=0.007) and ChatGPT-4o (U=30.0; Z=−4.20; p<0.001), while DeepSeek-R1 also outperformed ChatGPT-4o (U=99.0; Z=−2.01; p=0.045). For eye infection-related queries, Qwen-2.5 and DeepSeek-R1 outperformed ChatGPT-4o. Qwen-2.5 produced significantly more comprehensive responses than ChatGPT-4o (U=102.5; Z=−2.66; p=0.008), while DeepSeek-R1 also outperformed ChatGPT-4o (U=120.5; Z=−2.19; p=0.029). However, no significant differences were observed between Qwen-2.5 and DeepSeek-R1 (U=172.5; p=0.449), suggesting that while both models were superior to ChatGPT-4o in this category, they performed at similar levels to one another.

Comparative analysis of genAI models' performance for each CLEAR component per ophthalmology topic

The K-W test revealed statistically significant inter-model differences in at least one CLEAR dimension for refractive surgery, cataract, and eye infections, while glaucoma-related queries showed no significant differences across models (**Table 3**). For refractive surgery, significant disparities were found in all components: completeness (p<0.001), accuracy (p=0.009), and relevance (p<0.001) (**Table 3**). Post-hoc analysis indicated that Qwen-2.5 significantly outperformed both ChatGPT-4o and DeepSeek-R1 in all three dimensions (p<0.05). ChatGPT-4o scored significantly lower than both Chinese genAI models, particularly in completeness (p=0.006 vs DeepSeek-R1, p<0.001 vs Qwen-2.5).

For cataract, only relevance showed a significant difference across models (p=0.022), with Qwen-2.5 scoring higher than ChatGPT-4o (p=0.020) and DeepSeek-R1 (p=0.016). No significant differences were found in completeness (p=0.558) or accuracy (p=0.802), indicating similar performance across models for content depth and factual accuracy.

Table 3. Comparative performance of generative AI (genAI) models across ophthalmology topics based on CLEAR evaluation dimensions

Query topic	genAI model	CLEAR overall score	Completeness score	Accuracy score	Relevance score
		Mean±SD	Mean±SD	Mean±SD	Mean±SD
Refractive surgery	ChatGPT-4o	3.98±0.42	3.89±0.38	4.07±0.71	3.96±0.41
	DeepSeek-R1	4.26±0.47	4.30±0.64	4.22±0.68	4.26±0.31
	Qwen-2.5	4.59±0.26	4.69±0.27	4.52±0.59	4.56±0.20
p-value[a]		<0.001*	<0.001*	0.009*	<0.001*
Cataract	ChatGPT-4o	4.30±0.24	4.24±0.33	4.33±0.38	4.33±0.25
	DeepSeek-R1	4.15±0.61	4.05±0.70	4.18±0.74	4.21±0.47
	Qwen-2.5	4.35±0.21	4.26±0.34	4.31±0.28	4.49±0.17
p-value		0.424	0.558	0.802	0.022*
Glaucoma	ChatGPT-4o	4.16±0.45	4.13±0.51	4.28±0.42	4.05±0.54
	DeepSeek-R1	4.27±0.42	4.10±0.71	4.35±0.41	4.35±0.37
	Qwen-2.5	4.34±0.39	4.25±0.69	4.38±0.31	4.38±0.31
p-value		0.347	0.497	0.730	0.097
Eye infections	ChatGPT-4o	4.07±0.46	3.95±0.55	4.03±0.7	4.23±0.33
	DeepSeek-R1	4.38±0.22	4.43±0.22	4.32±0.41	4.38±0.16
	Qwen-2.5	4.43±0.23	4.48±0.20	4.33±0.43	4.47±0.23
p-value		0.015*	<0.001*	0.392	0.032*

[a]Analyzed using Kruskal Wallis test
*Statistically significant at p=0.05

For glaucoma-related queries, none of the three genAI models demonstrated statistically significant differences across completeness (p=0.497), accuracy (p=0.730), or relevance (p=0.097), suggesting that AI-generated responses in this domain are relatively uniform across models (**Table 3**). For eye infections, significant inter-model differences were observed in completeness (p<0.001) and relevance (p=0.032), but not in accuracy (p=0.392) (**Table 3**). Post-hoc analysis revealed that Qwen-2.5 scored significantly higher than ChatGPT-4o in completeness (p<0.001) and relevance (p=0.018), while DeepSeek-R1 also outperformed ChatGPT-4o in completeness (p=0.001). No significant differences were found between DeepSeek-R1 and Qwen-2.5 (**Table 3**).

Discussion

The findings of this study mark a major shift in the landscape of genAI in healthcare—one in which Chinese genAI models have not only rivaled, but in some domains, surpassed the pioneering model, namely OpenAI's ChatGPT-4o. For a few years, Western-developed AI models, particularly those from technology conglomerates such as OpenAI, Google, Microsoft, and Meta, have maintained an uncontested intellectual and technological power over large language models (LLMs). However, the emergence of Qwen-2.5 and DeepSeek-R1 as formidable challengers, particularly in healthcare domain, signals a dramatic disruption of this status quo.

In this study, the findings delineated a clear hierarchy of performance in responding to common queries frequently asked by the patients in ophthalmology practice. Specifically, the findings of this study showed that Qwen-2.5 led across all evaluation metrics, followed by DeepSeek-R1, while ChatGPT-4o consistently ranked lowest, albeit all of the tested genAI models demonstrated excellent performance. Notably, this trend held true across both English and Arabic queries, which signals a remarkable milestone in multilingual genAI competency. In the initial phase of genAI models' availability, genAI performance in non-English languages in the context of healthcare has suffered from linguistic bias, with models demonstrating notable declines in accuracy, completeness, and relevance outside of English.

For example, a recent study revealed a marked linguistic disparity in genAI performance in infectious disease queries [65]. The four tested genAI models were consistently rated as "excellent" in English, yet only "above average" in Arabic [65]. Similarly, in endocrine queries, ChatGPT-4o outperformed Microsoft Copilot—earning "excellent" rating in English and "very good" rating in Arabic—while Copilot achieved only "very good" to "good" ratings [88]. Comparable linguistic deficits in genAI performance have also been observed in other languages. For example, in a study on ChatGPT's diagnostic accuracy in Chinese language for retinal vascular diseases, Liu *et al.* found language disparities in its performance [89]. In contrast, our findings suggest that this linguistic gap is rapidly narrowing, highlighting the expanding multilingual generalizability of the next-generation of genAI models, especially the recently released Chinese genAI models.

In this study, the field of ophthalmology was chosen as the subject for benchmarking to assess genAI competence in health content generation. Unlike some other healthcare specialties that rely to a degree on subjective clinical interpretation, ophthalmology is characterized by highly structured diagnostic criteria, well-defined surgical interventions, and a vast body of standardized literature. Therefore, genAI models trained in ophthalmology must exhibit factual accuracy and a capacity for precise, structured, and contextually relevant reasoning. Based on the intricate nature of ophthalmology practice and education, numerous recent studies assessed the performance of genAI in this field with variable results [90-103]. For example, a study found that ChatGPT-4o outperformed other genAI models in myopia care-related queries, with 81% of responses rated as "good", compared to 61% for ChatGPT-3.5 and 55% for Google Bard [104]. Another study reported that GPT-3.5 excelled in delivering reliable, in-depth patient information, while GPT-4o provided strong general knowledge but lacked depth in specialized topics, and Gemini proved adequate for basic inquiries but insufficient for detailed medical guidance in thyroid eye disease [105]. Balas *et al.* demonstrated the potential of ChatGPT in generating medical retinal disease content that aligns closely with the American Academy of Ophthalmology Preferred Practice Pattern guidelines [106]. In the same vein, another study demonstrated that practicing ophthalmologists overwhelmingly preferred ChatGPT over Google for answering cataract-related frequently asked questions (FAQs), with ChatGPT's responses containing fewer inaccuracies [107]. Collectively, this growing body of literature, combined with our results, suggests that genAI models emerged as a valuable resource for eye health education and may serve as a tool for ophthalmologists to create customizable educational materials tailored to varied literacy levels of patients.

On the negative side, a study highlighted ChatGPT's limitations in ophthalmology, noting that while it received positive ratings, it still generates incomplete, incorrect, or potentially harmful recommendations [108]. These recommendations included invasive procedures not endorsed by the American Academy of Ophthalmology [108]. Additionally, a study examined AI chatbots in ophthalmic outpatient registration and differential diagnosis, finding GPT-4.0

outperformed GPT-3.5, approaching and occasionally surpassing resident-level diagnostic accuracy [90]. However, another study also showed that while AI chatbots show promise in triaging patients, their role in clinical decision-making requires further validation [90]. One scoping review acknowledged ChatGPT's potential in retinal conditions, aiding in clinical decision-making, patient education, and administrative automation [109]. Nevertheless, the same review highlighted ChatGPT's susceptibility to misinformation and clinical inaccuracies, necessitating careful oversight and responsible integration into practice [109].

Our results confirmed that all of the three tested genAI models—ChatGPT-4o, DeepSeek-R1, and Qwen-2.5—demonstrated excellent overall performance in ophthalmology-related queries. This suggests that genAI has reached a level of maturity where it can serve as a credible resource for health education and patient counseling. However, the performance gap among models in this study indicated that not all genAI models are created equal—with Qwen-2.5 consistently delivering the most comprehensive and contextually relevant responses, particularly in refractive surgery and eye infections.

The implications of our findings extend far beyond mere genAI benchmarking—they raise important questions about the role of genAI in the future of ophthalmology [110,111]. Will genAI remain a passive, informational tool, assisting physicians and patients in knowledge retrieval? Or are we on the cusp of a profound transformation, where genAI becomes an active participant in clinical workflows, diagnostic reasoning, and even treatment planning? While this study confirms that genAI excels at ophthalmology-related content generation, it also highlights the variability in genAI performance across different ophthalmologic domains. Although Qwen-2.5 outperformed its counterparts in refractive surgery and eye infections, all three models performed comparably in cataract and glaucoma-related queries. This suggests that certain domains of ophthalmology lend themselves more readily to genAI assistance, while others may require further refinement in training methodologies.

The next frontier in AI-assisted ophthalmology lies in multimodal AI—models that can integrate LLMs with deep learning algorithms capable of analyzing fundus images, optical coherence tomography scans, and visual field reports [112-114]. Future AI iterations may not merely answer patient inquiries but actively assist ophthalmologists in real-time diagnostic decision-making [115]. Moreover, AI-powered ophthalmology assistant chatbots could soon become embedded into electronic health records, streamlining clinical documentation, summarizing case histories, and offering evidence-based treatment recommendations [116-119]. However, for this vision to be realized safely and ethically, genAI models must be continuously benchmarked, validated against real-world clinical outcomes, and optimized for domain-specific accuracy [76].

Finally, despite the rigorous methodology employed in this study, several limitations must be acknowledged as follows. First, the evaluation of genAI models relied on a structured set of ophthalmology-related queries, which, while designed to reflect common clinical and patient inquiries, may not fully capture the variability and complexity of real-world ophthalmic consultations. The formulation, translation, and back-translation process ensured linguistic consistency between English and Arabic queries; however, subtle nuances in medical phrasing and interpretation across languages may still have influenced AI-generated responses. Second, the CLEAR tool used for assessing genAI performance—while validated for evaluating medical AI outputs—remains a subjective metric, dependent on human raters' interpretations of completeness, accuracy, and relevance. Additionally, to ensure assessment objectivity, future evaluations should incorporate blinded assessments where evaluators are unaware of which model generated each response, thereby reducing potential bias in performance comparisons. Future studies should explore more objective validation methods, such as direct patient outcomes or expert consensus panels, to refine genAI evaluation in clinical contexts. Third, the genAI models were tested under default configurations without fine-tuning. While this approach enhances replicability, it does not account for the potential impact of model updates, API modifications, or customized prompting strategies, all of which could influence response quality. Fourth, the study design focused exclusively on text-based AI responses and did not assess multimodal capabilities, such as AI models integrated with ophthalmic imaging analysis (fundus photography or optical coherence tomography scans). As genAI advances toward multimodal integration, future research should investigate how LLMs perform when coupled with deep learning models for image-based diagnostics and decision support. Lastly, this study did not assess patient perceptions or usability factors, which are crucial in determining whether AI-generated ophthalmic information is actionable, comprehensible, and trusted by patients and clinicians. Future studies should incorporate qualitative assessments, exploring patient and provider trust in AI-driven medical information, as well as its impact on health decision-making and adherence to treatment recommendations. Additionally, it is important to note that benchmarking outcomes may be affected by temporal and domain-specific biases, potentially favoring genAI models trained on newer or more representative data. Given these limitations, while our findings provide strong evidence for the growing capabilities of Chinese genAI models in ophthalmology, further research is required to validate these results in clinical practice, optimize AI integration, and ensure safe, equitable, and effective AI-driven ophthalmic care.

Conclusion

The findings of this study highlighted a remarkable shift in genAI dominance, with Chinese genAI models rivaling the pioneering and leading Western genAI model (ChatGPT-4o) in ophthalmology-related queries. Qwen-2.5 and DeepSeek-R1 have demonstrated that multilingual genAI models can now rival, and in some domains surpass OpenAI's (ChatGPT-4o). This milestone is not merely an academic curiosity—it has profound implications for global ophthalmology, patient education, and equitable access to medical knowledge. GenAI is on the cusp of transforming ophthalmic care, from multilingual patient education to AI-assisted diagnostics. However, its integration into clinical practice must be approached with caution, ensuring that these models are validated against real-world clinical standards and optimized for domain-specific accuracy. The future of ophthalmology will not be AI versus human expertise—it will be AI augmenting human expertise. If responsibly developed and ethically deployed, genAI could become one of the most powerful tools in modern ophthalmology, democratizing access to high-quality medical knowledge and reshaping the way eye care is delivered worldwide. The question is no longer whether genAI will revolutionize ophthalmology, but which genAI model will lead the way—and as this study suggests, the future of AI-driven ophthalmic innovation may very well be Chinese.

Ethics approval

Not required.

Acknowledgments

Not applicable.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

The datasets analyzed during the current study are available in Open Science Framework (OSF), using the following link: https://osf.io/9xw4y/with DOI: 10.17605/OSF.IO/9XW4Y.

Declaration of artificial intelligence use

This study used artificial intelligence (AI) tool and methodologies in the following capacities. ChatGPT-4o was employed for language refinement (improving grammar, sentence structure, and readability of the manuscript) and technical writing assistance (providing suggestions for structuring complex technical descriptions more effectively). We confirm that all AI-assisted processes were critically reviewed by the authors to ensure the integrity and reliability of the results. The final decisions and interpretations presented in this article were solely made by the authors.

How to cite

Sallam M, Alasfoor IM, Khalid SW, *et al*. Chinese generative AI models (DeepSeek and Qwen) rival ChatGPT-4 in ophthalmology queries with excellent performance in Arabic and English. Narra J 2025; 5 (1): e2371 - http://doi.org/10.52225/narra.v5i1.2371.

References

1. The British Broadcasting Corporation (BBC). AI named word of the year by Collins Dictionary. Available from: https://www.bbc.com/news/entertainment-arts-67271252. Accessed: 27 February 2025.
2. Mbizo T, Oosterwyk G, Tsibolane P, *et al*. Cautious optimism: The influence of generative AI tools in software development projects. In: Gerber A, editor. South African computer science and information systems research trends. Cham: Springer Nature Switzerland; 2024.
3. Yusuf A, Pervin N, Román-González M. Generative AI and the future of higher education: A threat to academic integrity or reformation? Evidence from multicultural perspectives. Int J Educ Technol High Educ 2024;21(1):21.
4. Cohen J, Lee G, Greenbaum L, *et al*. The generative world order: AI, geopolitics, and power. Goldman Sachs 2023. Available from: https://www.goldmansachs.com/insights/articles/the-generative-world-order-ai-geopolitics-and-power. Accessed: 27 February 2025.
5. Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. Healthcare 2023;11(6):887.
6. Sallam M, Al-Mahzoum K, Sallam M, *et al*. DeepSeek: Is it the end of generative AI monopoly or the mark of the impending doomsday? Mesopotam J Big Data 2025;2025:26-34.
7. Reddy S. Generative AI in healthcare: An implementation science informed translational path on application, integration and governance. Implement Sci 2024;19(1):27.
8. Sallam M, Al-Mahzoum K, Alaraji H, *et al*. Apprehension toward generative artificial intelligence in healthcare: A multinational study among health sciences students. Preprints 2024.
9. Li J, Dada A, Puladi B, *et al*. ChatGPT in healthcare: A taxonomy and systematic review. Comput Methods Programs Biomed 2024;245:108013.
10. Sengar SS, Hasan AB, Kumar S, *et al*. Generative artificial intelligence: A systematic review and applications. Multimed Tools Appl 2024.
11. Jandhyala V. GPT-4 and beyond : Advancements in AI language models. Int J Sci Res Comput Sci Eng Inf Technol 2024;10(5):274-285.
12. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595.
13. Einarsson H, Lund SH, Jónsdóttir AH. Application of ChatGPT for automated problem reframing across academic domains. Comput Educ Artif Intell 2024;6:100194.
14. Sallam M, Salim NA, Barakat M, *et al*. ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. Narra J 2023;3(1):e103.
15. Sallam M. Bibliometric top ten healthcare-related ChatGPT publications in the first ChatGPT anniversary. Narra J 2024;4(2):e917.
16. George AS. AI supremacy at the price of privacy: Examining the Tech Giants' Race for data dominance. Partners Univ Innov Res Publ 2025;3:26-43.
17. Gibney E. Scientists flock to DeepSeek: How they're using the blockbuster AI model. Nature 2025.
18. Normile D. Chinese firm's large language model makes a splash. Science 2025;387(6731):238.
19. Sallam M, Al-Mahzoum K, Eid H, *et al*. Chinese generative AI models challenge western AI in clinical chemistry MCQs: A benchmarking follow-up study on AI use in health education. Babyl J Artif Intell 2025;2025:1-14.
20. Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. NPJ Digit Med 2019;2:77.
21. Yim D, Khuntia J, Parameswaran V, *et al*. Preliminary evidence of the use of generative AI in health care clinical services: Systematic narrative review. JMIR Med Inform 2024;12:e52073.
22. Ibrahim H, Liu F, Asim R, *et al*. Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. Sci Rep 2023;13(1):12187.
23. Armbruster J, Bussmann F, Rothhaas C, *et al*. "Doctor ChatGPT, can you help me?" The patient's perspective: Cross-sectional study. J Med Internet Res 2024;26:e58831.

24. Mater W. Intention to use ChatGPT from healthcare workers: Jordan case study. In: 2024 2nd International Conference on Cyber Resilience (ICCR). Dubai; 2024.
25. Sallam M, Al-Mahzoum K, Almutairi YM, *et al.* Anxiety among medical students regarding generative artificial intelligence models: A pilot descriptive study. Int Med Educ 2024;3(4):406-425.
26. Robertson C, Woods A, Bergstrand K, *et al.* Diverse patients' attitudes towards artificial intelligence (AI) in diagnosis. PLOS Digit Health 2023;2(5):e0000237.
27. Berg HT, van Bakel B, van de Wouw L, *et al.* ChatGPT and generating a differential diagnosis early in an emergency department presentation. Ann Emerg Med 2024;83(1):83-86.
28. Mese I, Taslicay CA, Sivrioglu AK. Improving radiology workflow using ChatGPT and artificial intelligence. Clin Imaging 2023;103:109993.
29. Neha F, Bhati D, Shukla DK, *et al.* ChatGPT: Transforming healthcare with AI. AI 2024;5(4):2618-2650.
30. Singh JP. Quantifying healthcare consumers' perspectives: An empirical study of the drivers and barriers to adopting generative AI in personalized healthcare. RRST 2022;2(1):171-193.
31. Esmaeilzadeh P, Maddah M, Mirzaei T. Using AI chatbots (e.g., CHATGPT) in seeking health-related information online: The case of a common ailment. Comput Hum Behav Artif Humans 2025;3:100127.
32. Sallam M, Al-Salahat K, Eid H, *et al.* Human versus artificial intelligence: ChatGPT-4 outperforming bing, bard, ChatGPT-3.5 and humans in clinical chemistry multiple-choice questions. Adv Med Educ Pract 2024;15:857-871.
33. Sallam M, Al-Mahzoum K, Almutawaa RA, *et al.* The performance of OpenAI ChatGPT-4 and Google Gemini in virology multiple-choice questions: A comparative analysis of English and Arabic responses. BMC Res Notes 2024;17(1):247.
34. Salman I, Ameer O, Khanfar M, *et al.* Artificial intelligence in healthcare education: Evaluating the accuracy of ChatGPT, Copilot, and Google Gemini in cardiovascular pharmacology. Front Med 2025;12:1495378.
35. Rossettini G, Rodeghiero L, Corradi F, *et al.* Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: A cross-sectional study. BMC Med Educ 2024;24(1):694.
36. Zeng J. China's AI approach: A top-down nationally concerted strategy? In: Zeng J, editor. Artificial intelligence with Chinese characteristics national strategy, security and authoritarian governance; 2022.
37. McInerney K. Yellow Techno-Peril: The 'Clash of Civilizations' and anti-Chinese racial rhetoric in the US–China AI arms race. Big Data Soc 2024;11(2):20539517241227873.
38. Craddock M. The AI superpower showdown: Inside the US-China race for technological supremacy. Available from: https://medium.com/@mcraddock/inside-the-us-china-race-for-technological-supremacy-52cb5c3df063. Accessed: 28 February 2025.
39. Feng X, Xu K, Luo M-J, *et al.* Latest developments of generative artificial intelligence and applications in ophthalmology. Asia Pac J Ophthalmol 2024;13(4):100090.
40. Waisberg E, Ong J, Kamran SA, *et al.* Generative artificial intelligence in ophthalmology. Surv Ophthalmol 2025;70(1):1-11.
41. Sonmez SC, Sevgi M, Antaki F, *et al.* Generative artificial intelligence in ophthalmology: Current innovations, future applications and challenges. Br J Ophthalmol 2024;108(10):1335-1340.
42. Li Z, Wang L, Wu X, *et al.* Artificial intelligence in ophthalmology: The path to the real-world clinic. Cell Rep Med 2023;4(7):101095.
43. Wawer MPA, Reimer RP, Rokohl AC, *et al.* Artificial intelligence in ophthalmology - status quo and future perspectives. Semin Ophthalmol 2023;38(3):226-237.
44. Delsoz M, Madadi Y, Raja H, *et al.* Performance of ChatGPT in diagnosis of corneal eye diseases. Cornea 2024;43(5):664-670.
45. Delsoz M, Raja H, Madadi Y, *et al.* The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. Ophthalmol Ther 2023;12(6):3121-3132.
46. Burton MJ, Ramke J, Marques AP, *et al.* The lancet global health commission on global eye health: Vision beyond 2020. Lancet Glob Health 2021;9(4):e489-e551.
47. Teo ZL, Tham YC, Yu M, *et al.* Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis. Ophthalmology 2021;128(11):1580-1591.
48. Lundeen EA, Burke-Conte Z, Rein DB, *et al.* Prevalence of diabetic retinopathy in the US in 2021. JAMA Ophthalmol 2023;141(8):747-754.
49. Hashemi H, Pakzad R, Yekta A, *et al.* Global and regional prevalence of age-related cataract: A comprehensive systematic review and meta-analysis. Eye 2020;34(8):1357-1370.

50. Lin Y, Jiang B, Cai Y, *et al*. The global burden of glaucoma: Findings from the global burden of disease 2019 study and predictions by Bayesian age-period-cohort analysis. J Clin Med 2023;12(5):1828.
51. Costin H-N, Fira M, Goraș L. Artificial intelligence in ophthalmology: Advantages and limits. Appl Sci 2025;15(4):1913.
52. Oshika T. Artificial intelligence applications in ophthalmology. JMA J 2025;8(1):66-75.
53. Yang Z, Wang D, Zhou F, *et al*. Understanding natural language: Potential application of large language models to ophthalmology. Asia Pac J Ophthalmol 2024;13(4):100085.
54. Chen JS, Reddy AJ, Al-Sharif E, *et al*. Analysis of ChatGPT responses to ophthalmic cases: Can ChatGPT think like an ophthalmologist? Ophthalmol Sci 2025;5(1):100600.
55. Momenaei B, Wakabayashi T, Shahlaee A, *et al*. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. Ophthalmol Retina 2023;7(10):862-868.
56. Cox A, Seth I, Xie Y, *et al*. Utilizing ChatGPT-4 for providing medical information on blepharoplasties to patients. Aesthet Surg J 2023;43(8):NP658-NP662.
57. Al-Sharif EM, Penteado RC, Dib El Jalbout N, *et al*. Evaluating the accuracy of ChatGPT and Google BARD in fielding oculoplastic patient queries: A comparative study on artificial versus human intelligence. Ophthalmic Plast Reconstr Surg 2024;40(3):303-311.
58. Biswas S, Logan NS, Davies LN, *et al*. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. Ophthalmic Physiol Opt 2023;43(6):1562-1570.
59. Ittarat M, Cheungpasitporn W, Chansangpetch S. Personalized care in eye health: Exploring opportunities, challenges, and the road ahead for chatbots. J Pers Med 2023;13(12):1679.
60. Baig MM, Hobson C, GholamHosseini H, *et al*. Generative AI in improving personalized patient care plans: Opportunities and barriers towards its wider adoption. Appl Sci 2024;14(23):10899.
61. Bellanda VCF, Santos MLd, Ferraz DA, *et al*. Applications of ChatGPT in the diagnosis, management, education, and research of retinal diseases: A scoping review. Int J Retina Vitreous 2024;10(1):79.
62. Honavar SG. Eye of the AI storm: Exploring the impact of AI tools in ophthalmology. Indian J Ophthalmol 2023;71(6):2328-2340.
63. Celi LA, Cellini J, Charpignon ML, *et al*. Sources of bias in artificial intelligence that perpetuate healthcare disparities-A global review. PLOS Digit Health 2022;1(3):e0000022.
64. Hanna MG, Pantanowitz L, Jackson B, *et al*. Ethical and bias considerations in artificial intelligence/machine learning. Mod Pathol 2025;38(3):100686.
65. Sallam M, Al-Mahzoum K, Alshuaib O, *et al*. Language discrepancies in the performance of generative artificial intelligence models: An examination of infectious disease queries in English and Arabic. BMC Infect Dis 2024;24(1):799.
66. Sallam M, Mousa D. Evaluating ChatGPT performance in Arabic dialects: A comparative study showing defects in responding to Jordanian and Tunisian general health prompts. Mesopotam J Artif Intell Healthc 2024;2024:1-7.
67. Yu P, Fang C, Liu X, *et al*. Performance of ChatGPT on the Chinese postgraduate examination for clinical medicine: Survey study. JMIR Med Educ 2024;10:e48514.
68. Guigue PA, Meyer R, Thivolle-Lioux G, *et al*. Performance of ChatGPT in French language Parcours d'Accès Spécifique Santé test and in OBGYN. Int J Gynaecol Obstet 2024;164(3):959-963.
69. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the national nurse examinations in Japan: Evaluation study. JMIR Nurs 2023;6:e47305.
70. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, *et al*. Evaluating the efficacy of ChatGPT in navigating the Spanish medical residency entrance examination (MIR): Promising horizons for AI in clinical medicine. Clin Pract 2023;13(6):1460-1487.
71. Roumeliotis KI, Tselikas ND. ChatGPT and Open-AI Models: A preliminary review. Future Internet 2023;15(6):192.
72. Tian S, Jin Q, Yeganova L, *et al*. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform 2024;25(1):bbad493.
73. Fleisig E, Smith G, Bossi M, *et al*. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. arXiv:2406.08818;2024.
74. Al Shamsi H, Almutairi AG, Al Mashrafi S, *et al*. Implications of language barriers for healthcare: A systematic review. Oman Med J 2020;35(2):e122.
75. Bélisle-Pipon JC. Why we need to be careful with LLMs in medicine. Front Med 2024;11:1495582.
76. Sallam M, Khalil R, Sallam M. Benchmarking generative AI: A call for establishing a comprehensive framework and a generative AIQ test. Mesopotam J Artif Intell Healthc 2024;2024:69-75.

77. Samaan JS, Yeo YH, Ng WH, *et al*. ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. Arab J Gastroenterol 2023;24(3):145-148.
78. Faisal S, Kamran TE, Khalid R, *et al*. Evaluating the comprehension and accuracy of ChatGPT's responses to diabetes-related questions in Urdu compared to English. Digit Health 2024;10:20552076241289730.
79. Ozturk N, Yakak I, Ağ MB, *et al*. Is ChatGPT reliable and accurate in answering pharmacotherapy-related inquiries in both Turkish and English? Curr Pharm Teach Learn 2024;16(7):102101.
80. Sallam M, Barakat M, Sallam M. A preliminary checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence-based models in health care education and practice: Development study involving a literature review. Interact J Med Res 2024;13:e54704.
81. Sallam M, Barakat M, Sallam M. Pilot testing of a tool to standardize the assessment of the quality of health information generated by artificial intelligence-based models. Cureus 2023;15(11):e49373.
82. Alibaba Cloud. Qwen 2.5-Max: A large language model for advanced reasoning and domain-specific applications. Available from: https://chat.qwenlm.ai/. Accessed: 28 February 2025.
83. DeepSeek. DeepSeek. Available from: https://www.deepseek.com/. Accessed: 28 February 2025.
84. OpenAI. GPT-4o. Available from: https://openai.com/index/hello-gpt-4o/. Accessed: 28 February 2025.
85. Dhand NK, Khatkar MS. Statulator, an online statistical calculator for paired comparisons Available from: https://statulator.com/SampleSize/ss2PM.html. Accessed: 20 February 2025.
86. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155-163.
87. Sullivan GM, Feinn R. Using effect size-or why the P value is not enough. J Grad Med Educ 2012;4(3):279-282.
88. A-Abbasi H, Al-Qudheeby M, Kheyami ZA, *et al*. Cross-linguistic evaluation of generative AI models for diabetes and endocrine queries. Jordan Med J 2024;58(4):311-326.
89. Liu X, Wu J, Shao A, *et al*. Uncovering language disparity of ChatGPT on retinal vascular disease classification: Cross-sectional study. J Med Internet Res 2024;26:e51926.
90. Ming S, Yao X, Guo X, *et al*. Performance of ChatGPT in ophthalmic registration and clinical diagnosis: Cross-sectional study. J Med Internet Res 2024;26:e60226.
91. Tan YMC, Rojas-Carabali W, Cifuentes-González C, *et al*. The potential role of large language models in Uveitis Care: perspectives after ChatGPT and Bard launch. Ocul Immunol Inflamm 2024;32(7):1435-1439.
92. Rampat R, Debellemanière G, Gatinel D, *et al*. Artificial intelligence applications in cataract and refractive surgeries. Curr Opin Ophthalmol 2024;35(6):480-486.
93. Dihan Q, Chauhan MZ, Eleiwa TK, *et al*. Using large language models to generate educational materials on childhood glaucoma. Am J Ophthalmol 2024;265:28-38.
94. Cheong KX, Zhang C, Tan TE, *et al*. Comparing generative and retrieval-based chatbots in answering patient questions regarding age-related macular degeneration and diabetic retinopathy. Br J Ophthalmol 2024;108(10):1443-1449.
95. Ali MJ. ChatGPT and lacrimal drainage disorders: Performance and scope of improvement. Ophthalmic Plast Reconstr Surg 2023;39(3):221-225.
96. Gupta AS, Sulewski ME, Armenti ST. Performance of ChatGPT in cataract surgery counseling. J Cataract Refract Surg 2024;50(4):424-425.
97. Wu G, Lee DA, Zhao W, *et al*. ChatGPT and Google Assistant as a source of patient education for patients with amblyopia: Content analysis. J Med Internet Res 2024;26:e52401.
98. Hu X, Ran AR, Nguyen TX, *et al*. What can GPT-4 do for Diagnosing Rare Eye Diseases? A Pilot Study. Ophthalmol Ther 2023;12(6):3395-3402.
99. Shiraishi M, Tomioka Y, Miyakuni A, *et al*. Performance of ChatGPT in answering clinical questions on the practical guideline of blepharoptosis. Aesthetic Plast Surg 2024;48(13):2389-2398.
100. Rojas-Carabali W, Cifuentes-González C, Wei X, *et al*. Evaluating the diagnostic accuracy and management recommendations of ChatGPT in Uveitis. Ocul Immunol Inflamm 2024;32(8):1526-1531.
101. Rasmussen MLR, Larsen AC, Subhi Y, *et al*. Artificial intelligence-based ChatGPT chatbot responses for patient and parent questions on vernal keratoconjunctivitis. Graefes Arch Clin Exp Ophthalmol 2023;261(10):3041-3043.
102. Giannuzzi F, Carlà MM, Hu L, *et al*. Artificial intelligence with ChatGPT 4: A large language model in support of ocular oncology cases. Int Ophthalmol 2025;45(1):59.
103. Özer Özcan Z, Doğan L, Yilmaz IE. Artificial doctors: Performance of Chatbots as a tool for patient education on Keratoconus. Eye Contact Lens 2025;51(3):e112-e116.

104. Lim ZW, Pushpanathan K, Yew SME, *et al*. Benchmarking large language models' performances for myopia care: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023;95:104770.

105. Bahir D, Hartstein M, Zloto O, *et al*. Thyroid eye disease and artificial intelligence: A comparative study of ChatGPT-3.5, ChatGPT-4o, and Gemini in patient information delivery. Ophthalmic Plast Reconstr Surg 2024.

106. Balas M, Mandelcorn ED, Yan P, *et al*. ChatGPT and retinal disease: A cross-sectional study on AI comprehension of clinical guidelines. Can J Ophthalmol 2025;60(1):e117-e123.

107. Cohen SA, Brant A, Fisher AC, *et al*. Dr. Google vs. Dr. ChatGPT: Exploring the use of artificial intelligence in ophthalmology by comparing the accuracy, safety, and readability of responses to frequently asked patient questions regarding cataracts and cataract surgery. Semin Ophthalmol 2024;39(6):472-479.

108. Cappellani F, Card KR, Shields CL, *et al*. Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients. Eye 2024;38(7):1368-1373.

109. Bellanda VCF, Santos MLD, Ferraz DA, *et al*. Applications of ChatGPT in the diagnosis, management, education, and research of retinal diseases: A scoping review. Int J Retina Vitreous 2024;10(1):79.

110. Keskinbora KH. Current roles of artificial intelligence in ophthalmology. Explor Med 2023;4(6):1048-1067.

111. Choi JY, Yoo TK. New era after ChatGPT in ophthalmology: Advances from data-based decision support to patient-centered generative artificial intelligence. Ann Transl Med 2023;11(10):337.

112. Wang S, He X, Jian Z, *et al*. Advances and prospects of multi-modal ophthalmic artificial intelligence based on deep learning: a review. Eye Vis 2024;11(1):38.

113. Jin K, Yuan L, Wu H, *et al*. Exploring large language model for next generation of artificial intelligence in ophthalmology. Front Med 2023;10:1291404.

114. Sevgi M, Keane PA. Ophthalmology's new horizon: Moving from reactive care to proactive artificial intelligence solutions. Saudi J Ophthalmol 2023;37(3):171-172.

115. Sabaner MC, Anguita R, Antaki F, *et al*. Opportunities and challenges of Chatbots in ophthalmology: A narrative review. J Pers Med 2024;14(12):1165.

116. Singer M, Fu J, Chow J, *et al*. Development and evaluation of aeyeconsult: A novel ophthalmology Chatbot leveraging verified textbook knowledge and GPT-4. J Surg Educ 2023;81(3):438-443.

117. Ramjee P, Sachdeva B, Golechha S, *et al*. CataractBot: An LLM-powered expert-in-the-loop chatbot for cataract patients. arXiv:2402.04620;2025.

118. Lyons R, Arepalli S, Fromal O, *et al*. Artificial intelligence chatbot performance in triage of ophthalmic conditions. Can J Ophthalmol 2024;59(4):e301-e308.

119. Madadi Y, Delsoz M, Khouri A, *et al*. Applications of artificial intelligence-enabled robots and chatbots in ophthalmology: Recent advances and future trends. Curr Opin Ophthalmol 2024;35(3):238-243.