• Intelligent Ophthalmology •

Bilingual evaluation of large language models for patient education in refractive surgery

Tsung-Hsien Tsai[1,2], *Chin-Ling Tsai*[3], *Jui-Hung Hsu*[4], *Ching-Hsi Hsiao*[2,5], *Hung-Chi Chen*[2,5]

[1]Department of Ophthalmology, Chang Gung Memorial Hospital, Keelung 204, Taiwan, China
[2]School of Medicine, College of Medicine, Chang Gung University, Taoyuan 333, Taiwan, China
[3]Shanghai Aier Eye Hospital, Shanghai 200031, China
[4]Department of Medical Education, Chang Gung Memorial Hospital, Keelung 204, Taiwan, China
[5]Department of Ophthalmology, Chang Gung Memorial Hospital, Linkou 333, Taiwan, China
Co-first Authors: Tsung-Hsien Tsai and Chin-Ling Tsai
Correspondence to: Hung-Chi Chen. Chang Gung Memorial Hospital, Linkou Main Branch, No.5, Fuxing Street, Guishan District, Taoyuan 333, Taiwan, China. mr3756@cgmh.org.tw
Received: 2025-11-15 Accepted: 2026-01-28

Abstract

● **AIM:** To evaluate the ability of six advanced large language models (LLMs)—in providing accurate, comprehensive, and readable patient education on corneal refractive surgeries [laser *in-situ* keratomileusis (LASIK), keratorefractive lenticule extraction (KLEx), and photorefractive keratectomy (PRK)] in both English and Chinese.

● **METHODS:** This is a cross-sectional, comparative study. Twenty-six questions, compiled from authoritative ophthalmologic sources and covering four domains (procedure basics and eligibility; safety, risks and long-term stability; recovery and postoperative experience; and practical concerns), were administered in both English and Chinese *via* fresh chat sessions with each LLM, respectively. Five performance metrics were evaluated: accuracy, comprehensiveness, word count, readability, and reproducibility, using appropriate statistical tests.

● **RESULTS:** OpenAI o1 and DeepSeek-R1 consistently achieved the highest accuracy and most comprehensive responses, significantly outperforming ChatGPT-4o, Gemini Advanced, Claude Sonnet, and Tongyi Qwen (Friedman $P<0.001$). Although overall accuracy and comprehensiveness were similar across languages, Chinese responses were significantly longer. Readability varied among the models, with Claude Sonnet generally producing the most readable English texts. Reproducibility analysis revealed moderate consistency, reflecting inherent variability in outputs to identical prompts.

● **CONCLUSION:** Reasoning-augmented LLMs, particularly OpenAI o1 and DeepSeek-R1, demonstrate superior performance in delivering bilingual patient education for corneal refractive surgery, with high accuracy and comprehensiveness. However, variations in response length, readability, and reproducibility indicate that further refinement is necessary before these tools can be reliably integrated into clinical practice.

● **KEYWORDS:** large language models; refractive surgery; patient education

DOI:10.18240/ijo.2026.06.01

Citation: Tsai TH, Tsai CL, Hsu JH, Hsiao CH, Chen HC. Bilingual evaluation of large language models for patient education in refractive surgery. *Int J Ophthalmol* 2026;19(6):1019-1027

INTRODUCTION

Large language models (LLMs) are revolutionizing patient education by enhancing clinical decision-making, communication, and personalized care. The growing interest in LLM applications is underscored by over 100 publications in top ophthalmology journals since 2022[1]. Yet, their reliability in specialized fields such as ophthalmology remains underexplored, especially when addressing the complexities of refractive surgery consultations. Prior research has highlighted the potential of LLMs in managing common ocular conditions and procedures[2-16]. However, concerns about accuracy, readability, and cultural relevance have been raised[2,4,17-19].

In corneal refractive surgery—which includes laser *in-situ* keratomileusis (LASIK), keratorefractive lenticule extraction (KLEx), and photorefractive keratectomy (PRK)—there is a need for explanations that are not only technically precise but also linguistically clear and culturally sensitive. Notably, previous studies have predominantly focused on English-language outputs[20-22], overlooking the unique challenges of refractive surgeries and the increasing global demand for non-English educational resources. This issue is particularly pertinent in regions such as Chinese mainland, Hong Kong, Taiwan regions, and Singapore, where myopia prevalence is

high and the need for corneal refractive surgery is substantial among Chinese-speaking populations.

Accordingly, we evaluated six state-of-the-art LLMs—ChatGPT-4o, Gemini Advanced 1.5 Pro, Claude 3.5 Sonnet, Tongyi Qwen 2.5, OpenAI o1, and DeepSeek-R1—for their ability to answer common patient questions about LASIK, KLEx, and PRK in both English and Chinese. We comprehensively assessed each model on accuracy, comprehensiveness, word count, readability, and reproducibility. Given the widespread use of LLMs and the clinical need for patient education, our objective was to determine whether these models can serve as reliable adjunct tools for refractive surgery counseling across languages, and to define their current limitations.

MATERIALS AND METHODS

Study Design This cross-sectional, comparative study evaluated six advanced LLMs for their effectiveness in providing patient education on corneal refractive surgery. The models assessed included ChatGPT-4o, Gemini Advanced 1.5 Pro, Claude 3.5 Sonnet, Tongyi Qwen 2.5, OpenAI o1, and DeepSeek-R1. The Institutional Review Board of Chang Gung Memorial Hospital waived the study review due to the use of publicly available online data. Twenty-six questions related to common corneal refractive surgeries—LASIK, KLEx, and PRK—were compiled from official sources, including the National Eye Institute and the American Academy of Ophthalmology. Questions were then translated into Traditional Chinese using ChatGPT-4o and reviewed by two bilingual ophthalmologists (Tsai TH and Tsai CL) for semantic and cultural appropriateness. For further clarity, the questions (26 English and 26 Chinese) were stratified into four thematic domains: procedure basics and eligibility; safety, risks and long-term stability; recovery and postoperative experience; and practical concerns.

On February 1st, 2025, the six LLMs were accessed *via* newly created user accounts from Taiwan, China to reduce potential bias from prior interactions. All models were tested using their default publicly available settings at the time of access. For models that natively incorporate internal reasoning mechanisms (*e.g.*, OpenAI o1 and DeepSeek-R1), these built-in reasoning capabilities were not disabled. However, no additional manual prompting strategies, step-by-step instructions, or special "thinking" or "reasoning" modes were explicitly invoked for any model. All questions were independently submitted in fresh chat sessions using a standardized prompt format in both English and Chinese. The initial responses were recorded, randomized, and anonymized by a single ophthalmologist (Tsai CL). Three independent ophthalmologists (Tsai TH, Hsiao CH, Chen HC) evaluated each anonymized response for accuracy and comprehensiveness using a 5-point Likert scale, where 1 indicated strong inaccuracy or lack of comprehensiveness and 5 indicated complete accuracy and high comprehensiveness. In addition, word counts were automatically extracted from each response. Readability of the English responses was evaluated using four online automated metrics—Flesch-Kincaid Grade Level, Gunning Fog Index, Coleman-Liau Index, and SMOG Index—accessible *via* https://app.readable.com/text/. For Chinese outputs, a dedicated readability score was employed using an online tool available at http://120.27.70.114:8000/analysis_a. For Chinese texts, we used the Chinese Readability Platform (CRP)[23-25], a research-based tool that estimates text difficulty by mapping linguistic features to educational grade levels. Unlike English readability formulas, there is currently no universally accepted or clinically validated readability index for Chinese medical text. Therefore, CRP was used in this study to enable relative, within-study comparisons across models under identical prompts, rather than for absolute interpretation of patient comprehension or reading level. Higher CRP scores indicate lower readability. To assess reproducibility, the same questions were re-submitted two weeks later by Dr. Tsai CL, and responses were classified identical or non-identical by Dr. Tsai TH according to whether any key information was added or removed relative to the original answer. A summary of the data collection process and evaluation metrics is depicted in Figure 1.

Statistical Analysis Statistical analyses were performed using R statistical software (version 4.2.2; R Core Team, R Foundation for Statistical Computing, Vienna, Austria). Differences among the six LLMs in ordinal Likert scale ratings and other continuous outcomes—such as word count and readability—were analyzed using Friedman tests, with subsequent pairwise comparisons conducted *via* Wilcoxon signed-rank tests adjusted with Bonferroni corrections, while thematic domain differences were evaluated using the Kruskal–Wallis tests, also with Bonferroni-adjusted post hoc analyses. Inter-rater reliability among the three graders for accuracy and comprehensiveness scores was assessed using the intraclass correlation coefficient (ICC) with a two-way random-effects model and absolute agreement definition [ICC(2,k)]. This approach evaluates the consistency of averaged ratings across multiple raters. ICC values were calculated for the overall dataset and additionally stratified by language (English and Chinese) to explore potential linguistic effects on rating consistency. Categorical differences in reproducibility were assessed using Chi-square and Fisher's exact tests. A two-sided P value <0.05 was considered statistically significant, unless otherwise specified.

RESULTS

Table 1 summarizes the performance metrics evaluated across the six LLMs.

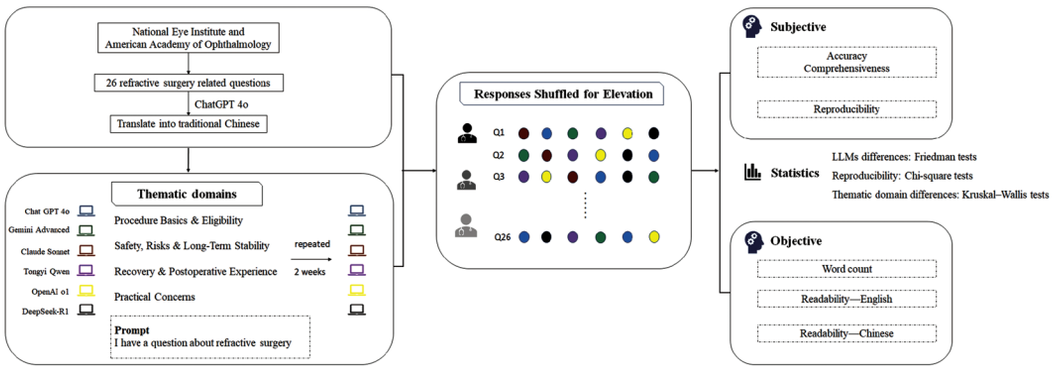

Figure 1 Study design Twenty-six refractive surgery-related questions from authoritative ophthalmology sources were translated into traditional Chinese and answered by six large language models. Responses were categorized into five thematic domains and shuffled for blinded evaluation by three graders. Both subjective (accuracy, comprehensiveness, reproducibility) and objective (word count, English/Chinese readability) outcomes were assessed using appropriate statistical tests. LLMs: Large language models.

Table 1 Performance metrics of LLMs—including accuracy, comprehensiveness, word count, reproducibility, and readability

n (%) or mean±SD

Model	Accuracy	Comprehensiveness	Wordcount	Reproducibility	Readability				
					Flesch-Kincaid	Gunning fog	Coleman-Liau	SMOG	Chinese readability
ChatGPT-4o	4.25±0.75	4.03±0.65	332.0±131.5	22 (42.3)	13.43±1.53	15.47±2.53	15.65±1.18	15.14±1.44	15.76±1.15
Gemini Advanced	4.38±0.79	4.29±0.69	443.0±182.8	21 (40.4)	10.90±1.40	13.44±1.67	13.51±1.56	13.23±1.17	17.24±1.54
Claude Sonnet	4.35±0.86	4.15±0.73	312.8±106.7	22 (42.3)	9.97±1.24	12.48±1.47	13.45±1.39	11.84±1.11	18.51±2.36
Tongyi Qwen	4.15±0.93	3.97±0.73	429.4±162.0	20 (38.5)	12.62±1.10	14.97±1.73	13.76±1.18	14.58±1.14	17.56±1.53
OpenAI o1	4.88±0.32	4.79±0.46	917.9±447.2	29 (55.8)	10.90±0.88	12.48±0.96	13.25±1.23	13.05±0.67	20.84±1.89
DeepSeek-R1	4.86±0.35	4.80±0.46	544.9±253.1	25 (48.1)	11.13±1.25	11.49±1.63	14.83±1.43	12.47±0.97	18.56±1.63

LLMs: Large language models; SD: Standard deviation.

Table 2 Pairwise comparisons of accuracy and comprehensiveness scores among the LLMs

Accuracy					Comprehensiveness				
Model 1	Model 2	Score	*P*	Adjusted *P*	Model 1	Model 2	Mean score	*P*	Adjusted *P*
ChatGPT-4o	Gemini Advanced	4.25 *vs* 4.38	0.154	NA	ChatGPT-4o	Gemini Advanced	4.03 *vs* 4.29	<0.001	0.004[a]
ChatGPT-4o	Claude Sonnet	4.25 *vs* 4.35	0.157	NA	ChatGPT-4o	Claude Sonnet	4.03 *vs* 4.15	0.080	1.199
ChatGPT-4o	Tongyi Qwen	4.25 *vs* 4.15	0.227	NA	ChatGPT-4o	Tongyi Qwen	4.03 *vs* 3.97	0.709	NA
ChatGPT-4o	OpenAI o1	4.25 *vs* 4.88	<0.001	<0.001[b]	ChatGPT-4o	OpenAI o1	4.03 *vs* 4.79	<0.001	<0.001[b]
ChatGPT-4o	DeepSeek-R1	4.25 *vs* 4.86	<0.001	<0.001[b]	ChatGPT-4o	DeepSeek-R1	4.03 *vs* 4.80	<0.001	<0.001[b]
Gemini Advanced	Claude Sonnet	4.38 *vs* 4.35	0.977	NA	Gemini Advanced	Claude Sonnet	4.29 *vs* 4.15	0.037	0.557
Gemini Advanced	Tongyi Qwen	4.38 *vs* 4.15	0.031	0.466	Gemini Advanced	Tongyi Qwen	4.29 *vs* 3.97	<0.001	0.004[a]
Gemini Advanced	OpenAI o1	4.38 *vs* 4.88	<0.001	<0.001[b]	Gemini Advanced	OpenAI o1	4.29 *vs* 4.79	<0.001	<0.001[b]
Gemini Advanced	DeepSeek-R1	4.38 *vs* 4.86	<0.001	<0.001[b]	Gemini Advanced	DeepSeek-R1	4.29 *vs* 4.80	<0.001	<0.001[b]
Claude Sonnet	Tongyi Qwen	4.35 *vs* 4.15	0.003	0.051	Claude Sonnet	Tongyi Qwen	4.15 *vs* 3.97	0.029	0.440
Claude Sonnet	OpenAI o1	4.35 *vs* 4.88	<0.001	<0.001[b]	Claude Sonnet	OpenAI o1	4.15 *vs* 4.79	<0.001	<0.001[b]
Claude Sonnet	DeepSeek-R1	4.35 *vs* 4.86	<0.001	<0.001[b]	Claude Sonnet	DeepSeek-R1	4.15 *vs* 4.80	<0.001	<0.001[b]
Tongyi Qwen	OpenAI o1	4.15 *vs* 4.88	<0.001	<0.001[b]	Tongyi Qwen	OpenAI o1	3.97 *vs* 4.79	<0.001	<0.001[b]
Tongyi Qwen	DeepSeek-R1	4.15 *vs* 4.86	<0.001	<0.001[b]	Tongyi Qwen	DeepSeek-R1	3.97 *vs* 4.80	<0.001	<0.001[b]
OpenAI o1	DeepSeek-R1	4.88 *vs* 4.86	0.589	NA	OpenAI o1	DeepSeek-R1	4.79 *vs* 4.80	0.763	NA

[a]*P*<0.01; [b]*P*<0.001. LLMs: Large language models; NA: Not available.

Accuracy The mean accuracy scores were highest for OpenAI o1 (4.88±0.32) and DeepSeek-R1 (4.86±0.35), both of which significantly outperformed ChatGPT-4o (4.25±0.75), Gemini Advanced (4.38±0.79), Claude Sonnet (4.35±0.86), and Tongyi Qwen (4.15±0.93; Friedman *P*<0.001). Pairwise Wilcoxon comparisons confirmed that both OpenAI o1 and DeepSeek-R1 were statistically superior to the other models (all *P*<0.003 Bonferroni-adjusted alpha), though there was no significant difference between them (*P*=0.589), as detailed in Table 2.

When the data were stratified by language, overall accuracy did not differ significantly between English (4.55±0.14) and Chinese (4.41±0.35) outputs (Wilcoxon *P*=0.089). Additionally, accuracy varied by thematic category (Kruskal-Wallis *P*<0.001), with the procedure basics and eligibility domain receiving higher ratings (4.62±0.65) than the domains

of safety, risks, long-term stability, and practical concerns (both $P<0.05$, Bonferroni-adjusted), as shown in Table 3.

Comprehensiveness Similarly, for comprehensiveness, OpenAI o1 (4.79±0.46) and DeepSeek-R1 (4.80±0.46) again achieved the highest scores, outperforming ChatGPT-4o (4.03±0.65), Gemini Advanced (4.29±0.69), Claude Sonnet (4.15±0.73), and Tongyi Qwen (3.97±0.73; Friedman $P<0.001$). Pairwise Wilcoxon tests confirmed that both top models were significantly more comprehensive than the others (all $P<0.003$), with no significant difference between OpenAI o1 and DeepSeek-R1 ($P=0.763$), as detailed in Table 2. Moreover, Gemini Advanced was significantly more comprehensive than both ChatGPT-4o and Tongyi Qwen (both $P<0.003$). When responses were analyzed by language, the overall comprehensiveness did not significantly differ between English (4.38±0.16) and Chinese (4.30±0.18) responses (Wilcoxon $P=0.116$). No significant differences in comprehensiveness were observed across the different thematic categories (Kruskal-Wallis $P=0.128$).

Inter-rater Reliability Inter-rater reliability among the three ophthalmologist graders was moderate. The ICC(2,k) was 0.540 for accuracy and 0.484 for comprehensiveness, indicating moderate agreement. In language-stratified analyses, agreement for accuracy was higher for Chinese responses than for English responses (English ICC=0.320 *vs* Chinese ICC=0.665), indicating greater consistency when evaluating Chinese outputs.

Word Count Word count analysis revealed that OpenAI o1 and DeepSeek-R1 produced longer responses compared to the other models (Friedman $P<0.001$), while ChatGPT-4o and Claude Sonnet generated shorter responses. Notably, Chinese responses (656.6±137.5 words) were significantly longer than English responses (336.7±70.8 words; $P<0.001$). Although overall comparisons across thematic categories reached significance, pairwise differences did not remain significant after Bonferroni correction.

Readability For the English responses, readability was evaluated using four established indices, as detailed in Figure 2.

The Flesch-Kincaid Grade Level revealed significant differences among models (Friedman $P<0.001$). Specifically, Gemini Advanced (10.90±1.40), Claude Sonnet (9.97±1.24), OpenAI o1 (10.90±0.88), and DeepSeek-R1 (11.13±1.25) produced notably more readable outputs than ChatGPT-4o (13.43±1.53) and Tongyi Qwen (12.62±1.10), with Claude Sonnet exhibiting superior readability compared to DeepSeek-R1. A similar pattern emerged using the Gunning Fog Index (Kruskal-Wallis $P<0.001$), where Gemini Advanced (13.44±1.67), Claude Sonnet (12.48±1.47), OpenAI o1 (12.48±0.96), and DeepSeek-R1 (11.49±1.63) again

Table 3 Pairwise comparisons of accuracy and readability scores across thematic categories

	Accuracy				Readability			
Category 1	Category 2	Score	*P*	Adjusted *P*	Category 1	Category 2	Score	*P*	Adjusted *P*
Procedure basics and eligibility	Safety, risks, and long-term stability	4.619 *vs* 4.413	0.002	0.010[a]	Procedure basics and eligibility	Safety, risks, and long-term stability	17.195 *vs* 14.910	0.069	0.411
Procedure basics and eligibility	Recovery and postoperative experience	4.619 *vs* 4.5	0.000	0.002[b]	Procedure basics and eligibility	Recovery and postoperative experience	17.195 *vs* 13.742	0.206	1
Procedure basics and eligibility	Practical concerns	4.619 *vs* 4.375	0.001	0.008[b]	Procedure basics and eligibility	Practical concerns	17.195 *vs* 13.322	0.024	0.144
Safety, risks, and long-term stability	Recovery and postoperative experience	4.413 *vs* 4.5	0.031	0.184	Safety, risks, and long-term stability	Recovery and postoperative experience	14.910 *vs* 13.742	0.001	0.005[b]
Safety, risks, and long-term stability	Practical concerns	4.413 *vs* 4.375	0.315	1	Safety, risks, and long-term stability	Practical concerns	14.910 *vs* 13.322	<0.001	<0.001[c]
Recovery and postoperative experience	Practical concerns	4.5 *vs* 4.375	0.001	0.007[b]	Recovery and postoperative experience	Practical concerns	13.742 *vs* 13.322	0.213	1

[a]$P<0.05$; [b]$P<0.01$; [c]$P<0.001$.

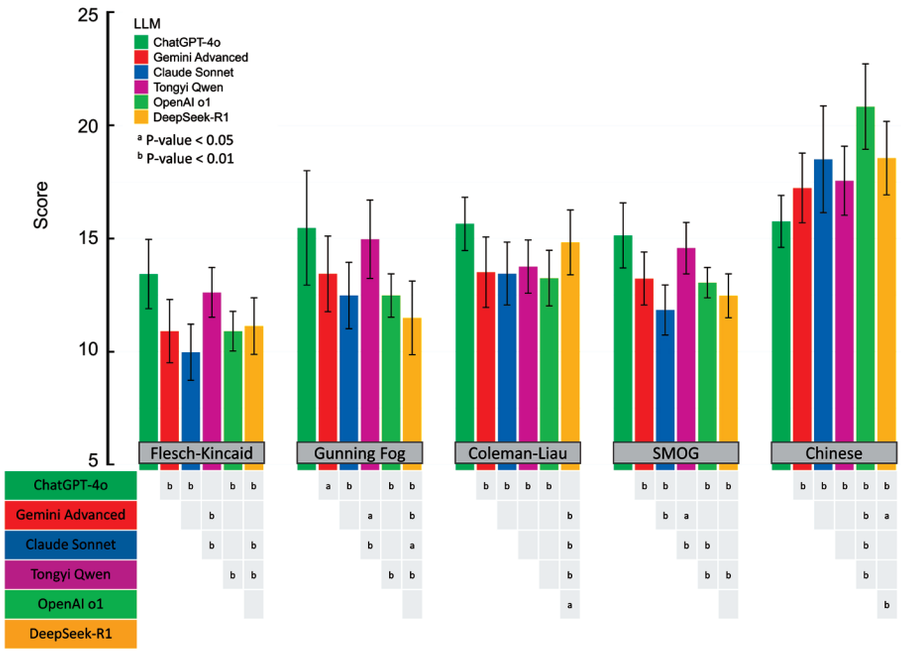

Figure 2 Readability scores of LLM-generated responses across five metrics Bar plot (mean±standard deviation) illustrating readability scores for six LLMs using four English metrics (Flesch–Kincaid, Gunning Fog, Coleman–Liau, and SMOG) and one Chinese readability metric. Higher scores represent more complex text and thus lower readability. ChatGPT-4o consistently produced the most complex English responses, whereas OpenAI o1 was rated the least readable in Chinese. Post hoc significance is as indicated ([a]*P*<0.05; [b]*P*<0.01). LLMs: Large language models.

outperformed ChatGPT-4o (15.47±2.53); in addition, Claude Sonnet, OpenAI o1, and DeepSeek-R1 scored significantly better than Tongyi Qwen (14.97±1.73), with DeepSeek-R1 even surpassing Gemini Advanced in pairwise comparisons. The Coleman-Liau Index further underscored inter-model variability (Friedman *P*<0.001), with Gemini Advanced (13.51±1.56), Claude Sonnet (13.45±1.39), Tongyi Qwen (13.76±1.18), and OpenAI o1 (13.25±1.23) generating more readable text than ChatGPT-4o (15.65±1.18) and DeepSeek-R1 (14.83±1.43). Moreover, comparisons across thematic categories revealed statistically significant differences (Kruskal-Wallis *P*<0.001), indicating that texts addressing practical concerns (13.32±1.33) and recovery/postoperative experience (13.74±1.38) were significantly more readable than those discussing safety, risks, and long-term stability (14.91±1.48; *P*<0.008, Bonferroni-adjusted), as shown in Table 3. Finally, the SMOG Index (Friedman *P*<0.001) favored Gemini Advanced (13.23±1.17), Claude Sonnet (11.84±1.11), OpenAI o1 (13.05±0.67), and DeepSeek-R1 (12.47±0.97) over ChatGPT-4o (15.14±1.44) and Tongyi Qwen (14.58±1.14), with Claude Sonnet outperforming both Gemini Advanced and OpenAI o1, though no significant differences were observed across thematic categories (Chi-square *P*=0.197).

In terms of Chinese readability, significant differences were also observed among models (Kruskal–Wallis *P*<0.001). Here, ChatGPT-4o demonstrated the best readability (15.76±1.15), outperforming Gemini Advanced (17.24±1.54), Claude Sonnet (18.51±2.36), Tongyi Qwen (17.56±1.53), OpenAI o1 (20.84±1.89), and DeepSeek-R1 (18.56±1.63). Both Gemini Advanced and the other leading models—Claude Sonnet, Tongyi Qwen, and DeepSeek-R1—showed superior performance compared to OpenAI o1, with Gemini Advanced also surpassed DeepSeek-R1. Importantly, no significant differences were detected across the various thematic categories (Chi-square *P*=0.083), as reported in Figure 2.

Reproducibility Over a two-week interval, OpenAI o1 had the highest reproducible response count (29/52; 55.8%), followed by DeepSeek-R1 (25/52; 48.1%), Claude Sonnet (22/52; 42.3%), ChatGPT-4o (22/52; 42.3%), Gemini Advanced (21/52; 40.4%), and Tongyi Qwen (20/52; 38.5%). Chi-square testing found no significant difference in reproducibility across models (*P*=0.511). English (44.9%) and Chinese (44.2%) responses had nearly identical reproducibility rates (*P*=1.000). There were no significant differences by thematic category either (Chi-square *P*=0.792).

DISCUSSION

Our results indicated that the reasoning-augmented models (OpenAI o1 and DeepSeek-R1) consistently generated the most accurate and comprehensive responses among the LLMs evaluated. However, overall reproducibility was suboptimal, and variability in readability was observed across models. These findings bear important implications for the integration of LLMs into clinical practice, particularly in the realm of patient counseling and education in refractive surgery.

Patient education is a critical step in the Enhanced Recovery after Surgery (ERAS®) protocols[26], with literature demonstrating that well-informed patients experience fewer postoperative complications and lower psychological stress[27]. In this context, LLMs offer transformative potential by delivering accurate, detailed, and timely information[28]. Given that corneal refractive surgeries are typically elective and self-paid procedures, effective patient education becomes even more crucial to meet the high expectations of patients.

Few studies have examined the role of LLMs in self-paid eye surgery consultations or patient education[15,20-22]. Beyond refractive surgery, a recent study evaluating LLM-generated patient education for upper blepharoplasty—another elective, largely self-paid ophthalmic procedure—found that several state-of-the-art models achieved high accuracy, but exhibited substantial variability in readability, with some models generating overly complex text. This finding mirrors our results and suggests that, even in elective surgical settings, current LLMs tend to prioritize informational completeness over patient accessibility. Together, these observations highlight that procedure-specific evaluation remains necessary before LLM-generated content can be safely adopted for patient counseling[15]. As for refractive surgery, Aydın *et al*[22] evaluated four earlier models—ChatGPT-3.5, ChatGPT-4.0, Google Gemini, and Microsoft Copilot—using 40 refractive surgery questions to assess appropriateness (accuracy), adequacy (comprehensiveness), and readability, finding that Google Gemini produced both more accurate answers and more readable texts. Demir[20] compared ChatGPT-4.0, Microsoft Copilot, and Google Gemini on 25 online-retrieved questions (with 13 questions similar to ours) and found that ChatGPT-4.0 achieved significantly higher accuracy (evaluated by Likert scores) and reliability, though its responses were less readable compared to the other models. In a further study by Doğan *et al*[21], earlier models including ChatGPT-3.5, Bard, and Bing Chat were assessed for accuracy, understandability, actionability, and readability in response to keratorefractive surgery questions; ChatGPT-3.5 demonstrated the highest accuracy, Bard provided higher understandability and actionability, and Bing Chat excelled in readability.

Our study expands upon these prior investigations by evaluating six LLMs—incorporating the latest reasoning-augmented models (OpenAI o1 and DeepSeek-R1), as well as Claude 3.5 Sonnet, Gemini Advanced 1.5 Pro, ChatGPT-4o, and Tongyi Qwen 2.5—to generate responses regarding corneal refractive surgery. We assessed these outputs in terms of accuracy, comprehensiveness, word count, readability, and reproducibility in both English and Chinese. Specifically, accuracy and comprehensiveness ratings reflect the models' ability to provide precise and complete information, while word count assesses response verbosity. Readability indices quantify the ease of understanding for users, and reproducibility evaluates consistency over time.

Notably, the reasoning-augmented models (OpenAI o1 and DeepSeek-R1) consistently delivered the most factually correct and comprehensive answers. A preliminary study of OpenAI o1 demonstrated an enhanced ability to understand complex medical scenarios, even surpassing GPT-4 in accuracy[29]. Similarly, DeepSeek-R1 exhibited exceptional reasoning capabilities, outperforming both ChatGPT and OpenAI o1 in tasks such as mathematics and coding[30], with a healthcare-focused study reporting that DeepSeek generates more comprehensive responses[31]. A preprint review summarized DeepSeek-R1's comparable or superior accessibility, accuracy, and readability, and emphasized its cost efficiency and adaptability in ophthalmology[32]. Recent evaluations further demonstrated that both OpenAI o1 and DeepSeek-R1 exhibit superior performance in answering ophthalmology questions, particularly in bilingual, complex reasoning tasks[33-34]. These studies echo our results.

Although three experienced ophthalmologists independently evaluated all responses in a blinded fashion, inter-rater agreement was moderate for both accuracy and comprehensiveness. This level of agreement is not unexpected given the inherently subjective nature of evaluating free-text, multi-topic medical counseling content using Likert scales, especially when responses differ in structure, emphasis, and level of detail rather than in clearly right-or-wrong factual statements. In language-stratified analyses, agreement for accuracy was higher for Chinese responses than for English responses. This suggests that differences in response heterogeneity between the two language settings may influence rating consistency. Importantly, despite variability in absolute scores, the relative ranking of model performance remained consistent across graders, with reasoning-augmented models demonstrating superior performance.

Whether LLM can perform equally across different languages had long been debated[35-36]. In field of ophthalmology for instance, previous work has shown that when using Chinese prompts, ChatGPT demonstrated promising support in clinical diagnosis of retinal diseases, though a performance gap with English prompts was noted[37]. In our study, however, the included LLMs showed comparable accuracy and comprehensiveness for both English and Chinese queries. One notable difference was that Chinese responses contained significantly more words than their English counterparts, echoing past studies reporting that some LLMs generate more characters in Chinese[38-39]. This disparity may reflect inherent

differences in communication style or model tuning; many of the included LLMs are primarily trained on English texts, which are naturally more concise.

Regarding English readability, we evaluated responses using several indices. The Flesch–Kincaid Grade Level estimates the U.S. school grade needed to understand a text[40], the Gunning Fog Index measures complexity based on average sentence length and the use of words with over three syllables, the Coleman–Liau Index calculates reading grade level based on letter density, and the SMOG Index estimates the years of education required by analyzing polysyllabic word frequency[41]. Our findings indicate that ChatGPT-4o generally produced responses that were more complex, whereas Claude Sonnet generated texts with higher readability. For example, using the Flesch–Kincaid metric, responses from Claude Sonnet were accessible to readers at a grade 10 level, while those from ChatGPT-4o required a reading level of grade 13. This is consistent with studies by Aydın *et al*[22] and Demir[20], which reported that ChatGPT-4.0 tends to generate more complex responses. Notably, our analysis also revealed that ChatGPT-4o's Chinese responses were more readable than the responses of other LLMs—a novel finding warranting further investigation. In general, patient education materials are recommended to be written at approximately a 6th to 8th grade reading level to ensure accessibility to the general population[41]. In this study, however, most LLM-generated responses substantially exceeded this recommended range, typically corresponding to approximately 10th to 15th grade reading levels across different models. This suggests that, despite acceptable factual accuracy, the current outputs of these models remain overly complex for direct use as patient-facing educational materials and would require simplification or clinician editing before clinical deployment.

Our domain-specific analysis revealed that LLM performance varied by the nature of the question. Questions concerning "procedure basics and eligibility" yielded the most accurate responses, likely due to the abundance and uniformity of factual information in the training data. Conversely, questions regarding "safety, risks, and long-term stability" or "recovery and practical concerns" were answered less accurately, possibly due to the contextual and individualized nature of these topics. In terms of readability, the Coleman-Liau Index indicated that texts addressing practical concerns (13.32±1.33) and recovery and postoperative experiences (13.74±1.38) were significantly more readable than those discussing safety and risks, suggesting that everyday language used in practical contexts may enhance readability compared to more technical, jargon-heavy discussions.

Moreover, our results also showed generally low reproducibility among all six LLMs, indicating that identical prompts can yield variable outputs. This low reproducibility likely stems from the stochastic nature of these models, which utilize random sampling during text generation[42]. Such inconsistency may undermine the reliability of LLMs in clinical applications, as patients could receive different information on subsequent interactions, potentially leading to confusion or misinterpretation. Consequently, while LLMs show promise in delivering detailed and accurate responses, enhancing output stability through refined sampling strategies or ensemble approaches is crucial for their dependable integration into patient education and clinical decision-making. Our study advances current understanding by incorporating several novel dimensions and provides clinically relevant insights. First, we evaluated six models, including advanced reasoning-augmented LLMs (OpenAI o1 and DeepSeek-R1) not previously examined, and showed that extended reasoning improves accuracy and comprehensiveness, which is clinically important for reducing the risk of missing critical information during patient counseling. Second, our bilingual evaluation demonstrated that these models maintain high performance across languages. Third, our reproducibility assessment, based on repeated querying, highlights the limited output stability of current LLMs, indicating that their responses are not yet suitable for direct, unsupervised patient-facing use and should be reviewed and edited by clinicians before being provided to patients. Finally, our granular, domain-specific analysis identifies specific strengths and weaknesses of LLMs across different aspects of refractive surgery counseling, which may help clinicians choose appropriate models and anticipate the level of post-editing required when using these tools as drafting.

Despite these advances, several limitations must be acknowledged. First, while we evaluated a broad range of models and metrics, our study remains limited by the static nature of the queries; LLM responses may vary over time with updates and new training data. Second, some LLM responses exhibited hallucinations[43]. For example, since "KLEx" is a relatively new term encompassing various lenticule-based procedures such as small incision lenticule extraction (SMILE), corneal lenticule extraction for advanced refractive correction (CLEAR) and smooth incision lenticule keratomileusis (SILK)[44], some responses distorted factual information regarding these techniques. Similarly, queries about the cost of procedures such as LASIK, KLEx, or PRK—where answers vary regionally—occasionally resulted in inaccurate or overly generalized responses. These issues underscore the need for cautious application of LLMs in patient education. Third, all models were accessed from Taiwan, China using newly created accounts. LLM outputs may differ across regions because of regulatory restrictions, region-specific deployments, localized

safety policies, or differences in service versions. Therefore, our results may not be fully generalizable to users in other regions. Finally, our reliance on publicly available online data may not fully capture the nuance of patient-specific clinical scenarios, which could affect the generalizability of our findings.

In conclusion, our study demonstrates that advanced reasoning-augmented LLMs, notably OpenAI o1 and DeepSeek-R1, show superior performance in terms of accuracy and comprehensiveness for refractive surgery patient education. While the models maintained consistent accuracy across English and Chinese queries, notable differences in response length and readability were observed. Our findings emphasize the transformative potential of LLMs in refractive surgery patient counseling, while also highlighting the need for further refinement to balance accuracy, readability, and consistency.

ACKNOWLEDGEMENTS

Authors' Contributions: Tsai TH: Conception and design of the research, analysis and interpretation of data, draft of the manuscript and final approval to the version to be published; Tsai CL: Acquisition, analysis and interpretation of data, and draft of the manuscript; Hsu JH: Acquisition, analysis and interpretation of data; Hsiao CH: Conception and design of the research, analysis of data and final approval to the version to be published; Chen HC: Conception and design of the research, analysis of data and final approval to the version to be published.

Conflicts of Interest: Tsai TH, None; **Tsai CL,** None; **Hsu JH,** None; **Hsiao CH,** None; **Chen HC,** None.

REFERENCES

1 Agnihotri AP, Nagel ID, Artiaga JCM, *et al*. Large language models in ophthalmology: a review of publications from top ophthalmology journals. *Ophthalmol Sci* 2025;5(3):100681.

2 Patil NS, Huang R, Mihalache A, *et al*. The ability of artificial intelligence chatbots ChatGPT and google bard to accurately convey preoperative information for patients undergoing ophthalmic surgeries. *Retina* 2024;44(6):950-953.

3 Caranfa JT, Bommakanti NK, Young BK, *et al*. Accuracy of vitreoretinal disease information from an artificial intelligence chatbot. *JAMA Ophthalmol* 2023;141(9):906.

4 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.

5 Tsui JC, Wong MB, Kim BJ, *et al*. Appropriateness of ophthalmic symptoms triage by a popular online artificial intelligence chatbot. *Eye (Lond)* 2023;37(17):3692-3693.

6 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.

7 Potapenko I, Malmqvist L, Subhi Y, *et al*. Artificial intelligence-based ChatGPT responses for patient questions on optic disc drusen. *Ophthalmol Ther* 2023;12(6):3109-3119.

8 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.

9 Daungsupawong H, Wiwanitkit V. Chatbot ChatGPT-4 and frequently asked questions about amblyopia and childhood myopia. *J Pediatr Ophthalmol Strabismus* 2024;61(2):151.

10 Cheong KX, Zhang CX, 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.

11 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.

12 Masalkhi M, Ong J, Waisberg E, *et al*. Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology. *Eye (Lond)* 2024;38(8):1412-1417.

13 Wu JH, Nishida T, Moghimi S, *et al*. Performance of ChatGPT on responding to common online questions regarding key information gaps in glaucoma. *J Glaucoma* 2024;33(7):e54-e56.

14 Su ZC, Jin K, Wu HK, *et al*. Assessment of large language models in cataract care information provision: a quantitative comparison. *Ophthalmol Ther* 2025;14(1):103-116.

15 Demir S, Türkeş İC. Evaluation of ChatGPT-4o, Claude 3.5 sonnet, and google gemini 2.0 flash as patient education resources for upper blepharoplasty patients. *J Craniofacial Surg* 2025;36(8): e1261-e1264.

16 Demir S. Evaluation of the reliability and readability of answers given by chatbots to frequently asked questions about endophthalmitis: a cross-sectional study on chatbots. *Health Inform J* 2024;30(4):14604582241304679.

17 García-Porta N, Vaughan M, Rendo-González S, *et al*. Are artificial intelligence chatbots a reliable source of information about contact lenses *Cont Lens Anterior Eye* 2024;47(2):102130.

18 Sensoy E, Citirik M. Assessing the proficiency of artificial intelligence programs in the diagnosis and treatment of cornea, conjunctiva, and eyelid diseases and exploring the advantages of each other benefits. *Cont Lens Anterior Eye* 2024;47(2):102125.

19 Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI* 2023;1(1):1-3.

20 Demir S. Investigating the role of large language models on questions about refractive surgery. *Int J Med Inform* 2025;195:105787.

21 Doğan L, Özer Özcan Z, Yılmaz E. The promising role of chatbots in keratorefractive surgery patient education. *J Fr Ophtalmol* 2025;48(2):104381.

22 Aydın FO, Aksoy BK, Ceylan A, *et al*. Readability and appropriateness of responses generated by ChatGPT 3.5, ChatGPT 4.0, gemini, and microsoft copilot for FAQs in refractive surgery. *Turk J Ophthalmol* 2024;54(6):313-317.

23 Cheng Y, Xu DK, Dong J. On key factors of text reading difficulty grading and readability formula based on Chinese textbook corpus. *Applied Linguistics* 2020;1:132-143.

24 Shi RH, Liu S, Xu XW, *et al*. Benchmarking four large language models' performance of addressing Chinese patients' inquiries about dry eye disease: a two-phase study. *Heliyon* 2024;10(14):e34391.

25 The Chinese readability platform. http://120.27.70.114:8000/analysis_a (Accessed on February 1, 2025).

26 Ljungqvist O, Francis NK, Urman RD. *Enhanced Recovery After Surgery: A Complete Guide to Optimizing Outcomes*. Cham: Springer International Publishing; 2020.

27 Brodersen F, Wagner J, Uzunoglu FG, *et al*. Impact of preoperative patient education on postoperative recovery in abdominal surgery: a systematic review. *World J Surg* 2023;47(4):937-947.

28 Aydin S, Karabacak M, Vlachos V, *et al*. Large language models in patient education: a scoping review of applications in medicine. *Front Med (Lausanne)* 2024;11:1477898.

29 Xie YF, Wu JC, Tu HQ, *et al*. A preliminary study of o1 in medicine: are we closer to an AI doctor? *arXiv preprint* 2024;arXiv:240915277.

30 Gibney E. Scientists flock to DeepSeek: how they're using the blockbuster AI model. *Nature* 2025. Epub ahead of print.

31 Temsah MH, Jamal A, Alhasan K, *et al*. OpenAI o1-preview *vs*. ChatGPT in healthcare: a new frontier in medical AI reasoning. *Cureus* 2024;16(10):e70640.

32 Egger J, De Paiva LF, Luijten G, *et al*. Is DeepSeek-R1 a game changer in healthcare-a seed review. *Authorea Preprints* 2025.

33 Srinivasan S, Ai X, Zou M, *et al*. Can OpenAI o1 Reason well in ophthalmology: A 6,990-question head-to-head evaluation study. *JAMA Ophthalmol* 2025;143(5):477-485.

34 Xu PS, Wu Y, Jin K, *et al*. DeepSeek-R1 outperforms Gemini 2.0 Pro, OpenAI o1, and o3-mini in bilingual complex ophthalmology reasoning. *Adv Ophthalmol Pract Res* 2025;5(3):189-195.

35 Lai V, Ngo N, Pouran Ben Veyseh A, *et al*. ChatGPT beyond English: towards a comprehensive evaluation of large language models in multilingual learning. *Findings of the Association for Computational Linguistics: EMNLP 2023.* Singapore. Stroudsburg, PA, USA: ACL, 2023:13171-13189.

36 Seghier ML. ChatGPT: not all languages are equal. *Nature* 2023;615(7951):216.

37 Liu XC, Wu JG, Shao A, *et al*. Uncovering language disparity of ChatGPT on retinal vascular disease classification: cross-sectional study. *J Med Internet Res* 2024;26:e51926.

38 Wang JY, Shi RH, Le QH, *et al*. Evaluating the effectiveness of large language models in patient education for conjunctivitis. *Br J Ophthalmol* 2025;109(2):185-191.

39 Huang YY, Shi RH, Chen C, *et al*. Evaluation of large language models for providing educational information in orthokeratology care. *Contact Lens Anterior Eye* 2025;48(3):102384.

40 Flesch R. A new readability yardstick. *J Appl Psychol* 1948;32(3):221-233.

41 Huang G, Fang CH, Agarwal N, *et al*. Assessment of online patient education materials from major ophthalmologic associations. *JAMA Ophthalmol* 2015;133(4):449-454.

42 Suh CH, Yi J, Shim WH, *et al*. Insufficient transparency in stochasticity reporting in large language model studies for medical applications in leading medical journals. *Korean J Radiol* 2024;25(11):1029-1031.

43 Yang ZF, Wang DM, Zhou FQ, *et al*. Understanding natural language: Potential application of large language models to ophthalmology. *Asia Pac J Ophthalmol (Phila)* 2024;13(4):100085.

44 Wang Y, Xie LX, Yao K, *et al*. Evidence-based guidelines for keratorefractive lenticule extraction surgery. *Ophthalmology* 2025;132(4):397-419.