Table S1. Inter-Rater Reliability for Evaluation Scores Using Intraclass Correlation Coefficient (ICC)

Evaluation Dimension	ICC	95% Confidence Interval	Level of Agreement
Quality	0.980	0.959 – 0.990	Excellent
Accuracy	0.963	0.929 – 0.981	Excellent
Bias	0.877	0.744 – 0.939	Good
Relevance	0.960	0.925 – 0.979	Excellent

Intraclass Correlation Coefficient (ICC) was calculated using a two-way random effects model based on an absolute agreement definition. The average measures ICC is reported, representing the reliability of the mean of the 7 raters.

Table S2. Detailed Statistical Results for Expert Evaluation Scores

Model	Metric	Quality	Accuracy	Bias	Relevance
DeepSeek	Mean (SD)	87.27 (5.70)	79.29 (7.82)	27.17 (7.25)	80.35 (3.37)
	Shapiro-Wilk *p*	0.572	0.462	0.123	0.022
	Mean Rank	158.94	98.28	109.06	53.09
Gemini	Mean (SD)	89.78 (5.20)	90.73 (4.04)	23.03 (6.28)	87.95 (3.71)
	Shapiro-Wilk *p*	0.002	0.070	0.032	<0.001
	Mean Rank	179.23	203.89	74.80	104.36
GPT	Mean (SD)	82.49 (5.78)	84.97 (4.41)	32.90 (8.73)	86.62 (3.98)
	Shapiro-Wilk *p*	0.172	0.215	0.014	0.035
	Mean Rank	112.40	145.06	155.17	149.19
Grok	Mean (SD)	74.27 (7.90)	74.62 (6.28)	35.44 (9.86)	86.27 (4.70)
	Shapiro-Wilk *p*	0.320	0.149	0.076	0.468
	Mean Rank	55.03	58.77	166.98	139.37
Kruskal-Wallis H Test		109.13	139.29	64.63	89.46

SD: Standard Deviation. This table provides the detailed descriptive and procedural statistics that support the primary analysis presented in Table 3. Shapiro-Wilk p-values were used to assess normality ($p < 0.05$ indicates non-normal data), justifying the use of non-parametric tests. The Kruskal-Wallis H statistic, calculated from the Mean Ranks, tests for overall differences among models.

Table S3. Dunn's Post-Hoc Test Pairwise Comparisons of Mean Ranks for Expert Evaluation Scores

Comparison Pair	Quality	Accuracy	Bias	Relevance
DeepSeek vs. Gemini	-1.595	3.046**	2.641*	-8.592**
DeepSeek vs. GPT	3.587**	-3.607**	-3.555**	-7.420**
DeepSeek vs. Grok	8.010**	-8.142**	-4.465**	-6.662**
Gemini vs. GPT	-5.182**	6.653**	6.196**	-1.171
Gemini vs. Grok	-9.605**	11.188**	-7.106**	1.930
GPT vs. Grok	4.423**	-4.535**	-0.910	0.759

Values presented are the test statistics from the Dunn's pairwise comparison test.
Statistical significance is denoted as:* $p < 0.05$, **$p < 0.001$.
A positive test statistic indicates that the mean rank of the first model in the pair is higher than the second. For Bias, a higher mean rank indicates a worse score (more bias).

Table S4. Examples of Fabricated References Generated by LLMs

Fabrication Type	Description	Example
Completely fabricated	This type of reference is completely fictional, from the title, author, and journal to the publication information, with no real counterparts.	Ivarsen A, Hjortdal J. Seven-year follow-up of small incision lenticule extraction for myopia. J Refract Surg. 2019;35(12):768-774.
Content fabricated	This type of reference uses real, existing journals or authors, but the content is fabricated and does not match the actually published articles.	Brar S, Sriganesh S, Sute SS, Ganesh S. Decade-Long Safety Outcomes of SMILE: Reduced Night Vision Issues in Follow-Ups. J Ophthalmol. 2022; 2022:4319785.[1]
Real but irrelevant	This type of reference cites real, published articles; however, the cited works are unrelated to the topic being discussed, and their content does not substantiate the claims made in the manuscript.	Kymionis GD, Diakonis VF, Kalyvianaki M, et al. One-year follow-up of corneal confocal microscopy after corneal cross-linking in patients with post laser in situ keratosmileusis ectasia and keratoconus. 2009;147(5):774-778.e1.[2]

[1] Note: As described in the original text, the authors and the journal are real; however, this article title does not exist in the journal's publication record.
[2] Note: As described in the original text, this article exists but focuses on LASIK rather than SMILE, making it irrelevant to the topic.
Abbreviation: LASIK= laser-assisted in situ keratomileusis; SMILE = small incision lenticule extraction.