Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Large language model use and hallucination awareness among medical students: a single-centre cross-sectional survey

Le Gao
Sun Yat-sen University
Jiamin Ouyang
Sun Yat-sen University
Lingxuan Peng
Sun Yat-sen University
Yuxian Zou
Sun Yat-sen University
Zitian Liu
liuzt25@mail.sysu.edu.cn

Sun Yat-sen University

Research Article

Keywords: large language models, hallucinations, medical education, ophthalmology, credibility, artificial intelligence literacy

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11390972/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Large language models (LLMs) are increasingly used in medical learning, but their answers may contain plausible yet incorrect information. It remains unclear whether students' awareness of these hallucinations is associated with how they judge the credibility of specialty-specific answers.

Methods

This single-centre cross-sectional survey was conducted at Sun Yat-sen University, China, from January to July 2026. Students completed a 17-item questionnaire on site and were classified into three groups by ophthalmic educational background. Students and five blinded senior ophthalmologists rated five examination-derived LLM answers on a five-point credibility scale. Analyses included descriptive statistics, Kendall's concordance, Friedman and Kruskal–Wallis tests, and exploratory Spearman correlations with bootstrap confidence intervals and false-discovery-rate (FDR) correction.

Results

Of 145 invited students, 138 provided complete questionnaires (95.2%): 22 in the basic-knowledge group, 80 in the course-level group, and 36 in the specialist-level group. Overall, 127 (92.0%) had used an LLM and 121 (87.7%) perceived improved learning efficiency. Although 98 (71.0%) reported familiarity with hallucinations and 108 (78.3%) reported encountering them, only 35 (25.4%) felt able to identify them clearly. Student mean credibility ratings ranged from 3.51 to 3.67, compared with expert means of 3.0–5.0. Students rated the retinal-neovascularization answer 0.51 points above the expert mean and the traumatic-lens-dislocation answer 1.43 points below it. Hallucination familiarity was associated with LLM familiarity ($\rho = 0.342$), use frequency ($\rho = 0.421$), and self-rated detection ability ($\rho = 0.274$; all FDR-adjusted $P < 0.01$). Neither hallucination familiarity nor self-rated ability was significantly associated with closer agreement with experts.

Conclusions

Students commonly used LLMs and recognised hallucination risks, but few felt able to identify hallucinations clearly. Awareness was not significantly associated with closer agreement with expert credibility judgments. These findings support further evaluation of supervised, case-based teaching in source verification and appraisal of LLM-generated information. Objective hallucination detection requires separate assessment.

Background

Large language models (LLMs), such as ChatGPT and DeepSeek, are increasingly used in medical education. They can explain medical concepts, summarise learning materials, and answer clinical questions, offering convenient support for independent learning. Surveys suggest that medical students value these applications while recognising concerns about the reliability of generated information [1, 2]. Evaluations using specialist question banks have also shown that response accuracy varies between models and that apparently convincing explanations may contain incorrect reasoning or important omissions [3]. A major concern is hallucination, in which an LLM generates information that appears plausible but is factually incorrect or unsupported [4].

Such information may affect both students' immediate learning and their subsequent clinical practice. In a randomised study of novice medical students, misleading AI-generated explanations reduced diagnostic accuracy compared with receiving no explanation [5]. Beyond these immediate effects, misconceptions acquired during learning could persist into clinical training and influence later decisions. This potential risk is particularly relevant to specialty-specific knowledge, such as ophthalmology, where errors concerning treatment indications, contraindications, or complications could affect patient assessment, management, and counselling. Expert assessments of LLM-generated ophthalmic answers have identified content with the potential to cause harm, although these studies did not demonstrate harm to actual patients [3]. Students therefore need to recognise the possibility of hallucinations and critically assess generated information before incorporating it into their learning or clinical reasoning.

Several learner characteristics may be relevant to this process, including familiarity with LLMs, awareness of hallucinations, and subject-specific knowledge. Research among general practice trainees has linked AI familiarity and use frequency to how readily trainees accept or reject generated answers [6]. A separate pilot study using simulated AI decision support found that clinicians with greater subject knowledge were better able to reject incorrect advice [7]. This provides indirect evidence because the study did not evaluate LLMs. Structured teaching may also be relevant: a controlled educational study among pre-medical students reported improvements in AI and pathology knowledge following an instructional programme, although it did not directly assess hallucination detection [8]. These findings provide a rationale for examining students' experience with LLMs and their educational background. However, it remains unclear whether greater familiarity or awareness is associated with credibility judgments that more closely agree with those of specialists.

Ophthalmic questions provide a practical setting for examining these judgments within a broader survey of LLM use and hallucination awareness. Previous research has used independent expert assessment to evaluate the accuracy, usefulness, and safety of generated ophthalmic answers [9]. Less is known about how students assess the credibility of such answers or how their ratings compare with expert judgments.

This study therefore aimed to describe medical students' LLM use and hallucination awareness and examine factors associated with their credibility ratings of LLM-generated specialty-specific answers. Using ophthalmic responses as study materials, we explored associations with LLM familiarity, use

frequency, and hallucination awareness, and compared student ratings and their agreement with expert judgments across three educational-background groups.

Methods

Study design and setting

This institution-based cross-sectional survey was conducted at Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China, from January to July 2026. Reporting followed the STROBE statement and relevant CHERRIES items for electronic questionnaire administration [10, 11].

Ethics approval and consent

The study was approved by the Institutional Review Board of Zhongshan Ophthalmic Center, Sun Yat-sen University. Students were informed about the research purpose, voluntary participation, and confidentiality of their responses. Individual questionnaire results were not returned to participants.

Participants and questionnaire administration

The sampling frame comprised undergraduate medical students and professional postgraduate students in ophthalmology. A total of 145 students were invited; 138 provided complete questionnaires and seven incomplete questionnaires were excluded. Participants were classified by ophthalmic educational background: basic knowledge (second-year undergraduates, n = 22), course-level knowledge (fourth-year undergraduates who had completed ophthalmology coursework, n = 80), and specialist-level knowledge (ophthalmology postgraduates, n = 36).

Students completed the questionnaire on site. Participation was voluntary and no incentives were offered. Personal information was kept confidential, and individual results were not fed back to students.

Questionnaire content

The locally developed questionnaire contained 17 items in four domains. It was administered in Chinese; an English translation is provided in Additional file 1. Formal cognitive interviewing and psychometric validation before administration were not documented.

The first domain (Q1–Q4) covered educational level, familiarity with LLMs, daily use frequency, and settings of use. Educational level was categorical and settings of use allowed multiple selections. The second domain (Q5–Q8) assessed perceived learning efficiency, trust relative to teaching physicians, trust when using LLMs under teacher supervision, and whether LLMs could completely replace teachers.The third domain (Q9–Q12) covered familiarity with hallucinations, frequency of encountering them, self-rated ability to identify hallucinated medical content, and perceived harm to learning. The

fourth domain (Q13–Q17) asked students to rate the credibility of five complete LLM-generated ophthalmic answers. The scale was coded 1 = completely untrustworthy, 2 = relatively untrustworthy, 3 = neutral, 4 = relatively trustworthy, and 5 = completely trustworthy. These items measured perceived credibility rather than the detection of predefined factual errors.

For descriptive reporting, familiarity with hallucinations meant any response above completely unfamiliar, and having encountered hallucinations meant any response above never. Agreement combined agree and strongly agree. LLM use was defined as any reported use rather than selecting never.

Generation of ophthalmic answers

Five multiple-response questions were randomly selected from a question bank for the Chinese ophthalmology standardized residency-training completion examination. They concerned glaucoma risk factors, keratitis treatment, complications of hyphema, diseases associated with retinal neovascularization, and traumatic lens dislocation. Each Chinese prompt asked the model to answer the question and provide an explanation.

Each question was submitted once through the official DeepSeek web interface using DeepSeek-V3 in January 2026 [12]. The final response from that initial submission was retained without prompt tuning, regeneration, or selection among alternative outputs. The retained answers were shown to students without subsequent correction. English translations of the prompts and responses are included in Additional file 1.

Student and expert ratings

Students rated each of the five answers independently within the questionnaire. They evaluated the credibility of the model's response rather than selecting the correct examination options. Their five ratings were summed to give a total credibility score from 5 to 25.

The same retained answers were emailed to five senior ophthalmologists: one glaucoma specialist, one cornea specialist, two retina specialists, and one cataract specialist. Experts used the same five-point credibility scale and completed their ratings independently. All 25 individual expert ratings were retained (Additional file 1, Table S1).

Expert means, medians, and ranges were summarised for each answer. Directional differences were calculated as the student mean minus the expert mean. Exact agreement and agreement within one point were calculated against the expert median. For each student, the mean absolute deviation was the average of the five absolute differences from the corresponding expert medians; its possible range was 0–4, with lower values indicating closer agreement. These comparisons described agreement in credibility judgments and were not measures of factual accuracy.

Statistical analysis

Analyses used Python 3.12 with pandas, NumPy, and SciPy. Category coding and item completeness were checked. Categorical variables were reported as counts and percentages; selected proportions had Wilson 95% confidence intervals (CIs). Ordered responses were summarised using medians and interquartile ranges (IQRs), with means and standard deviations (SDs) provided to describe rating differences. Figure 2 uses t-based 95% CIs for student means.

The Friedman test compared students' five repeated credibility ratings, with Kendall's W as an effect size. Cronbach's alpha described the consistency of these five ratings; it was not used to establish a single underlying construct or to validate the questionnaire. Expert agreement was described using the proportion of ratings equal to the item mode and tie-corrected Kendall's W. The latter measures agreement in the ranking of answers, rather than absolute agreement in scores, and was treated as exploratory because there were only five answers.

Exploratory Spearman correlations examined LLM familiarity and use frequency in relation to hallucination familiarity; hallucination familiarity and encounter frequency in relation to self-rated detection ability; and seven questionnaire variables in relation to each of the total credibility score and mean absolute deviation. The two correlations with hallucination familiarity were added during manuscript revision to address the stated study aim and were not prespecified. Percentile 95% CIs for Spearman coefficients were calculated from 5,000 bootstrap resamples of participants.

The three educational groups were compared using Kruskal–Wallis tests for six outcomes: hallucination familiarity, encounter frequency, self-rated detection ability, perceived harm, total credibility score, and mean absolute deviation. Five additional Kruskal–Wallis tests compared the groups on individual credibility items. The effect size η^2_H was calculated as $(H - k + 1)/(N - k)$, where $k = 3$ groups and N was the analysed sample size; negative estimates were set to zero. Benjamini–Hochberg FDR adjustment was applied separately to the two correlations with hallucination familiarity, the two correlations with self-rated ability, the seven correlations for each credibility outcome, the six educational-group comparisons, and the five item-specific group comparisons. Two-sided P values were used, with an FDR-adjusted threshold of 0.05 for these exploratory association analyses. No prospective sample-size calculation was documented; all complete responses were included in the primary analysis.

Results

Participants and LLM use

Of 145 invited students, 138 provided complete questionnaires (95.2%); seven incomplete questionnaires were excluded. The analysed sample included 22 students in the basic-knowledge group (15.9%), 80 in the course-level group (58.0%), and 36 in the specialist-level group (26.1%). These

corresponded to second-year undergraduates, fourth-year undergraduates, and ophthalmology postgraduates, respectively (Table 1).

Table 1
Participant characteristics and LLM use (N = 138)

Characteristic	n	%
Ophthalmic knowledge background		
Basic knowledge	22	15.9
Course-level knowledge	80	58.0
Specialist-level knowledge	36	26.1
LLM familiarity (Q2)		
Completely unfamiliar	11	8.0
Slightly familiar	85	61.6
Moderately familiar	32	23.2
Very familiar	10	7.2
Daily LLM question frequency (Q3)		
Never	11	8.0
Fewer than 1 per day	17	12.3
1–10 per day	86	62.3
10–50 per day	23	16.7
More than 50 per day	1	0.7
Settings of use (Q4; multiple responses)		
Learning	127	92.0
Daily life	85	61.6
Work	56	40.6
Chat or companionship	36	26.1
Never used an LLM	11	8.0

Basic knowledge: second-year undergraduates; course-level knowledge: fourth-year undergraduates; specialist-level knowledge: ophthalmology postgraduates (24 professional master's and 12 professional doctoral students).

Overall, 127 students had used an LLM (92.0%; 95% CI, 86.3–95.5%). Most described themselves as slightly familiar with LLMs (85/138, 61.6%); 32 (23.2%) were moderately familiar, 10 (7.2%) were very familiar, and 11 (8.0%) were completely unfamiliar. The most common daily use category was 1–10 questions (86/138, 62.3%). Learning was the most frequently reported use setting (127/138, 92.0%), followed by daily life (85/138, 61.6%), work (56/138, 40.6%), and chat or companionship (36/138, 26.1%).

Educational attitudes and hallucination awareness

Most students agreed or strongly agreed that LLMs improved learning efficiency (121/138, 87.7%). Fewer than half considered LLM-generated medical answers as trustworthy as answers from teaching physicians (57/138, 41.3%). More than two-thirds considered LLM use more trustworthy with teacher supervision (94/138, 68.1%), while only 10 students (7.2%) believed that LLMs could completely replace teachers (Table 2 and Fig. 1).

Table 2
Agreement with statements about LLM use and hallucinations (N = 138)

Statement	Disagree or strongly disagree, n (%)	Neutral, n (%)	Agree or strongly agree, n (%)
Improves learning efficiency (Q5)	4 (2.9)	13 (9.4)	121 (87.7)
As trustworthy as teaching physicians (Q6)	27 (19.6)	54 (39.1)	57 (41.3)
More trustworthy with teacher supervision (Q7)	15 (10.9)	29 (21.0)	94 (68.1)
Can completely replace teachers (Q8)	102 (73.9)	26 (18.8)	10 (7.2)
Can clearly identify hallucinations (Q11)	33 (23.9)	70 (50.7)	35 (25.4)
Undetected hallucinations harm learning (Q12)	10 (7.2)	30 (21.7)	98 (71.0)

Bars show the five ordered response categories for 138 students. Labels within sufficiently large segments are rounded percentages.

Ninety-eight students reported some familiarity with hallucinations (71.0%; 95% CI, 63.0–77.9%), and 108 reported having encountered them during medical learning (78.3%; 95% CI, 70.7–84.3%). Thirty-four (24.6%) reported encounters often or always. Only 35 students agreed that they could clearly identify hallucinated medical content (25.4%; 95% CI, 18.8–33.2%); 70 (50.7%) were neutral and 33 (23.9%) disagreed. Ninety-eight (71.0%) agreed that undetected hallucinations could harm learning.

Credibility ratings of the five answers

Student mean ratings ranged from 3.51 to 3.67, and all five answers had a median rating of 4 (IQR, 3–4). The mean total score was 17.96 (SD, 3.61), with a median of 18 (IQR, 15–20) and an observed range of

5–25. Differences among the five repeated ratings did not reach statistical significance (Friedman χ^2=9.01, df = 4, P = 0.061; Kendall's W = 0.016). Cronbach's alpha was 0.87.

Expert mean ratings for Q13–Q17 were 4.0, 3.8, 4.0, 3.0, and 5.0; the corresponding medians were 4, 4, 4, 3, and 5. Experts gave identical scores for four of the five answers. For the keratitis answer, the cornea specialist scored 3 and the other four specialists scored 4. Overall, 24 of 25 ratings (96.0%) equalled the item mode, and rank concordance was high (Kendall's W = 0.941; Table S1).

Student means were 0.33, 0.24, and 0.36 points below the expert means for the glaucoma, keratitis, and hyphema answers, respectively. Students rated the retinal-neovascularization answer 0.51 points above the unanimous expert score of 3. The largest difference was for traumatic lens dislocation: the student mean was 1.43 points below the unanimous expert score of 5 (Table 3 and Fig. 2). Exact agreement with the expert median ranged from 9.4% to 50.7%. Across students, the mean absolute deviation from the five expert medians was 0.85 points (SD, 0.43).

Table 3
Student and expert credibility ratings for the five ophthalmic answers

Item/topic	Student mean (SD)	Student median (IQR)	Expert mean (range)	Expert median	Student minus expert mean	Exact agreement, n (%)	Within 1 point, n (%)
Q13 Glaucoma risk factors	3.67 (0.90)	4 (3–4)	4.0 (4–4)	4	-0.33	70 (50.7)	124 (89.9)
Q14 Keratitis treatment	3.56 (0.97)	4 (3–4)	3.8 (3–4)	4	-0.24	57 (41.3)	121 (87.7)
Q15 Hyphema complications	3.64 (0.91)	4 (3–4)	4.0 (4–4)	4	-0.36	61 (44.2)	123 (89.1)
Q16 Retinal neovascularization	3.51 (0.85)	4 (3–4)	3.0 (3–3)	3	+ 0.51	48 (34.8)	123 (89.1)
Q17 Traumatic lens dislocation	3.57 (0.81)	4 (3–4)	5.0 (5–5)	5	-1.43	13 (9.4)	79 (57.2)

N = 138 students and 5 experts. All ratings used the same 1–5 scale. IQR, interquartile range; SD, standard deviation.

Blue circles show student means with 95% confidence intervals; red diamonds show expert means and ranges. Expert ranges collapse to a point for unanimous ratings. Q13, glaucoma; Q14, keratitis; Q15, hyphema; Q16, retinal neovascularization; Q17, traumatic lens dislocation.

Factors associated with awareness and credibility ratings

Students who were more familiar with LLMs also reported greater familiarity with hallucinations (Spearman's ρ = 0.342; 95% CI, 0.169–0.499). More frequent LLM use was similarly associated with greater hallucination familiarity (ρ = 0.421; 95% CI, 0.252–0.568); both FDR-adjusted P values were below 0.001. Hallucination familiarity was associated with higher self-rated detection ability (ρ = 0.274; 95% CI, 0.104–0.428; FDR-adjusted P = 0.002), as was more frequent experience of hallucinations (ρ = 0.191; 95% CI, 0.017–0.361; FDR-adjusted P = 0.025; Table S2).

Neither hallucination familiarity nor self-rated detection ability was significantly associated with the total credibility score or with mean absolute deviation from expert medians. For deviation from expert medians, the correlations were ρ = 0.075 (FDR-adjusted P = 0.733) for hallucination familiarity and ρ=−0.129 (FDR-adjusted P = 0.457) for self-rated ability. None of the seven examined variables was significantly associated with either credibility outcome after FDR correction (Table 4).

Table 4
Correlations with total credibility and deviation from expert medians

Variable	Total credibility: ρ (95% CI)	FDR-adjusted P	Mean absolute deviation: ρ (95% CI)	FDR-adjusted P
LLM familiarity (Q2)	-0.117 (-0.291 to 0.053)	0.385	-0.042 (-0.198 to 0.122)	0.733
LLM use frequency (Q3)	-0.105 (-0.287 to 0.086)	0.385	-0.001 (-0.174 to 0.173)	0.987
Trust relative to teaching physicians (Q6)	0.178 (-0.003 to 0.349)	0.257	-0.133 (-0.297 to 0.034)	0.457
Hallucination familiarity (Q9)	-0.121 (-0.291 to 0.052)	0.385	0.075 (-0.097 to 0.249)	0.733
Hallucination encounters (Q10)	-0.081 (-0.247 to 0.093)	0.487	0.057 (-0.122 to 0.231)	0.733
Self-rated detection ability (Q11)	-0.065 (-0.245 to 0.123)	0.525	-0.129 (-0.280 to 0.024)	0.457
Perceived educational harm (Q12)	-0.034 (-0.214 to 0.148)	0.696	-0.045 (-0.221 to 0.133)	0.733

All analyses included 138 students. Spearman correlations; 95% CIs are percentile intervals from 5,000 bootstrap resamples. Lower mean absolute deviation indicates closer agreement with expert medians. CI, confidence interval; FDR, false discovery rate.

Comparisons across ophthalmic knowledge-background groups

Mean total credibility scores were 17.41 (SD, 2.17) in the basic-knowledge group, 17.90 (SD, 3.78) in the course-level group, and 18.42 (SD, 3.94) in the specialist-level group. The corresponding mean absolute deviations from expert medians were 0.83 (SD, 0.38), 0.86 (SD, 0.42), and 0.83 (SD, 0.48) points (Table 5).

No statistically significant group differences were found for total credibility (H = 1.42, df = 2; FDR-adjusted P = 0.590) or mean absolute deviation (H = 0.67, df = 2; FDR-adjusted P = 0.715). Group differences in hallucination familiarity, encounter frequency, self-rated detection ability, and perceived harm also did not reach significance after correction (all FDR-adjusted P ≥ 0.208; Table S3). None of the five item-specific group comparisons was significant after correction (all FDR-adjusted P ≥ 0.156; Table S4).

Table 5
Credibility ratings across the three knowledge-background groups

Group	n	Total score: mean (SD)	Total score: median (IQR)	Deviation: mean (SD)	Deviation: median (IQR)
Basic knowledge	22	17.41 (2.17)	18 (16–19)	0.83 (0.38)	0.80 (0.45–1.15)
Course-level knowledge	80	17.90 (3.78)	18.5 (15–20)	0.86 (0.42)	0.90 (0.60–1.00)
Specialist-level knowledge	36	18.42 (3.94)	19 (16.75–20)	0.83 (0.48)	0.80 (0.60–1.00)

Total credibility scores range from 5 to 25. Deviation is each student's mean absolute difference from the five expert medians (possible range, 0–4). Overall three-group comparisons gave H = 1.42 and FDR-adjusted P = 0.590 for total scores, and H = 0.67 and FDR-adjusted P = 0.715 for deviation.

Sensitivity analyses

After omitting the three potentially inconsistent records (n = 135), the mean total credibility score was 17.87 (SD, 3.58) and the mean absolute deviation was 0.85 (SD, 0.43). LLM familiarity and use frequency remained associated with hallucination familiarity (both FDR-adjusted P < 0.001); hallucination familiarity and encounters remained associated with self-rated detection ability (FDR-adjusted P = 0.004 and 0.043, respectively). None of the examined variables was significantly associated with either credibility outcome after correction (all FDR-adjusted P ≥ 0.276), and none of the six educational-group comparisons was significant (all FDR-adjusted P ≥ 0.136). The Friedman test remained non-significant (P = 0.072).

Discussion

Principal findings

This survey identified a gap between students' awareness of LLM hallucinations and their confidence in recognising them. LLM use was common, and most students saw benefits for learning. Although 71.0% were familiar with hallucinations and 78.3% reported encountering them, only 25.4% felt able to identify them clearly. Greater LLM familiarity and more frequent use were associated with greater hallucination awareness. However, neither awareness nor self-rated detection ability was significantly associated with closer agreement with expert credibility ratings.

The answers that students trusted most or least also differed from the experts' judgments. Student mean ratings were similar across the five answers, whereas expert means ranged from 3 to 5. The clearest differences involved the retinal-neovascularization and traumatic-lens-dislocation answers. These findings describe differences in perceived credibility; they do not establish whether students could identify particular factual errors.

LLMs as learning tools and the role of teachers

The high level of use in this sample is consistent with growing interest in LLMs in medical education. Ganjavi and colleagues reported that 52% of students surveyed in May 2023 used ChatGPT for medical school coursework [1]. That estimate measures coursework-related ChatGPT use, whereas our 92.0% estimate measures any reported LLM use. Differences in definitions, survey dates, models, and sampling prevent a direct comparison of adoption rates. Across studies, however, students commonly report convenience and educational benefits while recognising the need to check information [1, 2].

Our findings also show that perceived usefulness did not imply unrestricted trust. Only 41.3% considered LLM answers as trustworthy as answers from teaching physicians, while 68.1% believed LLM use was more trustworthy under teacher supervision. Few thought that LLMs could completely replace teachers. These responses support using LLMs alongside teaching, with educators helping students examine explanations, identify errors, and judge when further advice is needed. The questionnaire measured attitudes towards supervision, rather than the effectiveness of a supervised teaching intervention.

Awareness, confidence, and evaluation of answers

More familiar and frequent users were more aware of hallucinations, and students with greater awareness reported higher detection confidence. These associations are plausible, but they cannot show whether LLM use increased awareness, whether aware students used LLMs more often, or whether both reflected other experiences. Self-reported familiarity and confidence were not significantly associated with smaller differences from expert ratings. General knowledge that LLMs can make errors may therefore be an incomplete indicator of how a student evaluates a particular answer.

Previous studies using labelled correct and incorrect AI outputs have also found difficulties in evaluating medical content [6, 13]. Those studies measured discrimination more directly than the present survey.

Our five-point credibility ratings cannot be interpreted as an error-detection test, and agreement with a small expert panel is not a validated measure of clinical competence.

The item-level results nevertheless help identify questions for teaching. Student ratings were close to expert means for glaucoma, keratitis, and hyphema, but were higher for retinal neovascularization and lower for traumatic lens dislocation. The latter two findings suggest that agreement with specialists can vary in both directions. They should not be described as a general tendency to trust all AI answers excessively. Topic knowledge, uncertainty, or features of the generated explanations might have influenced ratings, but students were not asked to explain their decisions.

The rationale for comparing ophthalmic backgrounds is supported by indirect evidence from simulated decision support that subject expertise may help clinicians reject incorrect AI advice [7]. However, our data did not demonstrate a significant group difference in credibility judgments. This does not contradict the possibility that knowledge matters: our groups reflected educational stages, whereas the judgments concerned only five answers and were not an objective knowledge test. The findings also cannot separate the effects of coursework, clinical exposure, and individual study.

Implications for medical education

Teaching should give students opportunities to practise checking medical information, in addition to explaining that hallucinations exist. In supervised cases, students could identify the claims in an LLM answer, compare them with reliable references, explain which claims they accept, and discuss missing information or uncertainty. Cases should include accurate as well as inaccurate answers so that students learn when trust is justified. Expert feedback can then address the reasons behind each judgment. Students have rated interactive case-based teaching as helpful for evaluating AI outputs [13], and this approach is consistent with role-based proposals for AI education [14]. The controlled educational study by Hopson and colleagues provides preliminary evidence that structured instruction can improve relevant knowledge [8], while the experiment by Teng and colleagues illustrates the risks of misleading explanations for novice learners [5]. Whether instruction improves objective detection of ophthalmic hallucinations remains to be tested.

Strengths and limitations

Our current study include retaining the first generated answer to each question without selecting preferred outputs, using examination-derived prompts across five topics, and preserving all individual ratings from five blinded subspecialists. The analysis included complete responses from 138 of 145 invited students and reported effect sizes and corrections for multiple comparisons.

However, several limitations affect interpretation. This single-centre sample may not represent other institutions or curricula, and the educational groups differed in size. Attitudes, hallucination encounters, and detection ability were self-reported and may be affected by recall and social-desirability bias. The three knowledge-background categories were based on training stage and did not establish each individual’s mastery.

The five stimuli and five experts provided a limited assessment. Global credibility ratings did not separately measure factual accuracy, completeness, or clinical safety. High observed expert concordance across five answers does not establish a universal reference standard. The stimuli came from one model at one time, and results may change with other questions or outputs.

Future multicentre studies should use larger item sets with independently verified accurate and inaccurate answers. They should record exact prompts and model settings, assess students' subject knowledge and prior AI training, and ask participants to explain their judgments. This would allow objective error detection and confidence to be compared before and after educational interventions.

Conclusions

Medical students commonly used LLMs and recognised their educational value, but only a minority felt able to identify hallucinated medical content clearly. Familiarity with hallucinations was associated with self-rated detection ability, yet neither was significantly associated with closer agreement with expert credibility ratings. These findings support supervised practice in evaluating LLM-generated information and checking reliable sources. They do not establish students' objective ability to detect hallucinations, which should be assessed using independently verified outputs.

Abbreviations

AI
artificial intelligence
CHERRIES
Checklist for Reporting Results of Internet E-Surveys
CI
confidence interval
FDR
false discovery rate
IQR
interquartile range
LLM
large language model
SD
standard deviation
STROBE
Strengthening the Reporting of Observational Studies in Epidemiology.

Declarations

Ethics approval and consent to participate

The study was conducted in accordance with the Declaration of Helsinki.The study was approved by the Institutional Review Board of Zhongshan Ophthalmic Center, Sun Yat-sen University, which waived the requirement for written informed consent. All participants were informed of the voluntary nature of participation, the confidentiality of their information, and the exclusive use of their responses for research purposes. Verbal informed consent was obtained from all participants.

Consent for publication

The authors consent to publication with BMC Medical Education.

Availability of data and materials

The deidentified dataset and analysis code are available from the corresponding author on reasonable request, subject to institutional and ethics requirements. The English questionnaire, translated study stimuli, expert ratings, and supplementary analyses are provided in Additional file 1.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the Teaching Quality and Teaching Reform Project of Zhongshan Ophthalmic Center, Sun Yat-sen University (No. 91017-12263011).

Authors' contributions

Study conception and design: Zitian Liu. Data collection: Jiamin Ouyang, Lingxuan Peng. Analysis and interpretation: Zitian Liu, Le Gao. Manuscript drafting and revision: Le Gao, Zitian Liu.

Acknowledgements

We would like to express their sincere gratitude to all the medical students who participated in this study. We also thank the ophthalmology experts for their valuable professional insights and contributions to the study, as well as our colleagues for their assistance and support throughout the research process.

References

1. Ganjavi C, Eppler M, O'Brien D, Ramacciotti LS, Ghauri MS, Anderson I, et al. ChatGPT and large language models (LLMs) awareness and use. A prospective cross-sectional survey of U.S. medical students. PLOS Digit Health. 2024;3(9):e0000596. 10.1371/journal.pdig.0000596.
2. Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGPT for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. JMIR Med

Educ. 2023;9:e50658. 10.2196/50658.

3. Chen JL, Lu AJ, Verma R, Wang L, Koch DD, Chen AJ. Assessment of correctness, content omission, and risk of harm in large language model responses to ophthalmology continuing medical education questions. Ophthalmol Sci. 2026;6(5):101130. 10.1016/j.xops.2026.101130.
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29:1930–40. 10.1038/s41591-023-02448-8.
5. Teng D, Tan L, Cao Q, Xia Y, Zhang N, Li J, et al. Impact of AI misinformation on diagnostic accuracy and confidence calibration in novice medical students. npj Digit Med. 2026;9:356. 10.1038/s41746-026-02547-z.
6. Zhou J, Zhang J, Wan R, Cui X, Liu Q, Guo H, et al. Integrating AI into clinical education: evaluating general practice trainees' proficiency in distinguishing AI-generated hallucinations and impacting factors. BMC Med Educ. 2025;25:406. 10.1186/s12909-025-06916-2.
7. Micocci M, Borsci S, Thakerar V, Walne S, Manshadi Y, Edridge F, et al. Attitudes towards trusting artificial intelligence insights and factors to prevent the passive adherence of GPs: a pilot study. J Clin Med. 2021;10(14):3101. 10.3390/jcm10143101.
8. Hopson S, Mildon C, Hassard K, Kubalek C, Laverty L, Urie P, et al. Enhancing AI literacy in undergraduate pre-medical education through student associations: an educational intervention. BMC Med Educ. 2025;25:999. 10.1186/s12909-025-07556-2.
9. Luo MJ, Pang J, Bi S, Lai Y, Zhao J, Shang Y, et al. Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. JAMA Ophthalmol. 2024;142(9):798–805. 10.1001/jamaophthalmol.2024.2513.
10. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med. 2007;147(8):573–7. 10.7326/0003-4819-147-8-200710160-00010.
11. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res. 2004;6(3):e34. 10.2196/jmir.6.3.e34.
12. DeepSeek-AI. DeepSeek-V3 Technical Report. arXiv [Preprint]. 2024. arXiv:2412.19437. 10.48550/arXiv.2412.19437
13. Waldock WJ, Lam G, Baptista A, Walls R, Sam AH. Which curriculum components do medical students find most helpful for evaluating AI outputs? BMC Med Educ. 2025;25:195. 10.1186/s12909-025-06735-5.
14. Ng FYC, Thirunavukarasu AJ, Cheng H, Tan TF, Gutierrez L, Lan Y, et al. Artificial intelligence education: an evidence-based medicine approach for consumers, translators, and developers. Cell Rep Med. 2023;4(10):101230. 10.1016/j.xcrm.2023.101230.

Figures

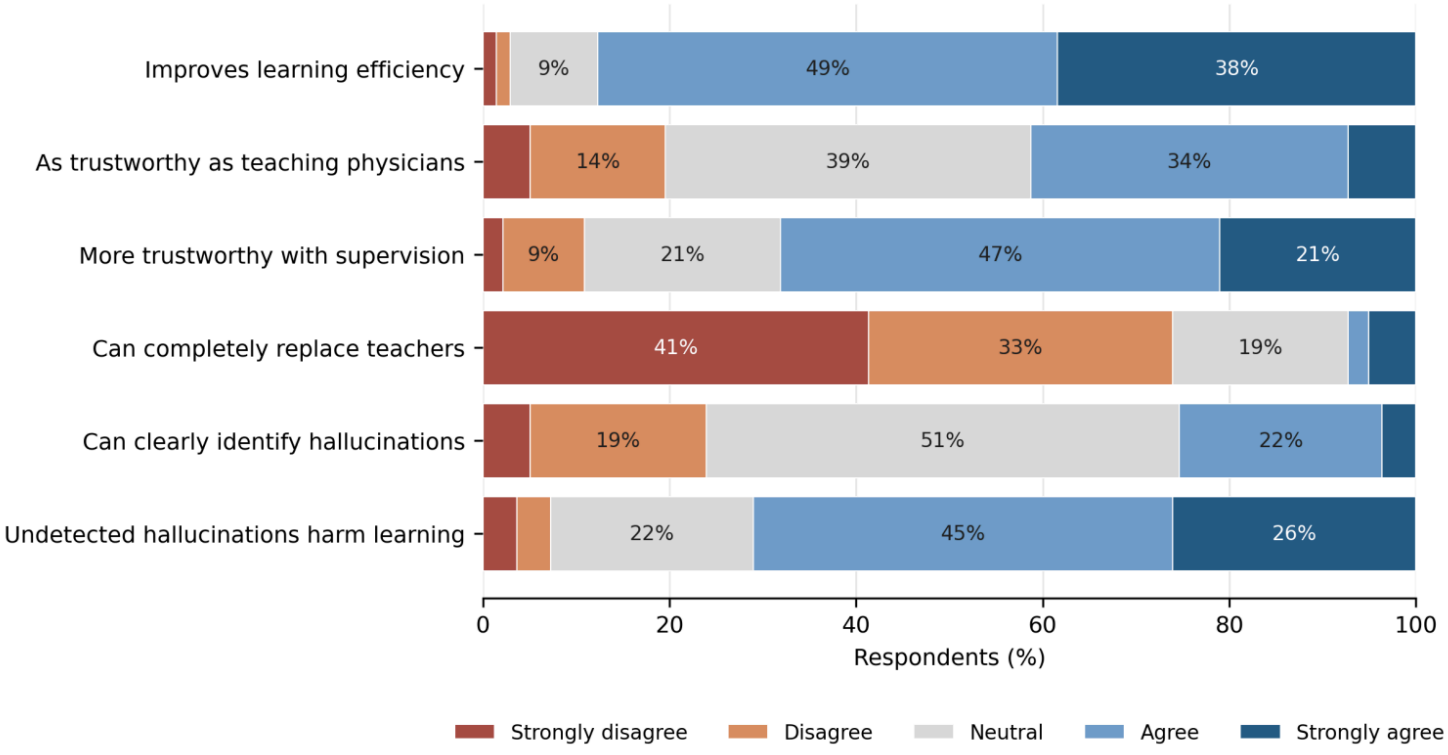


Figure 1

Responses to statements about LLM use and hallucinations

Bars show the five ordered response categories for 138 students. Labels within sufficiently large segments are rounded percentages.

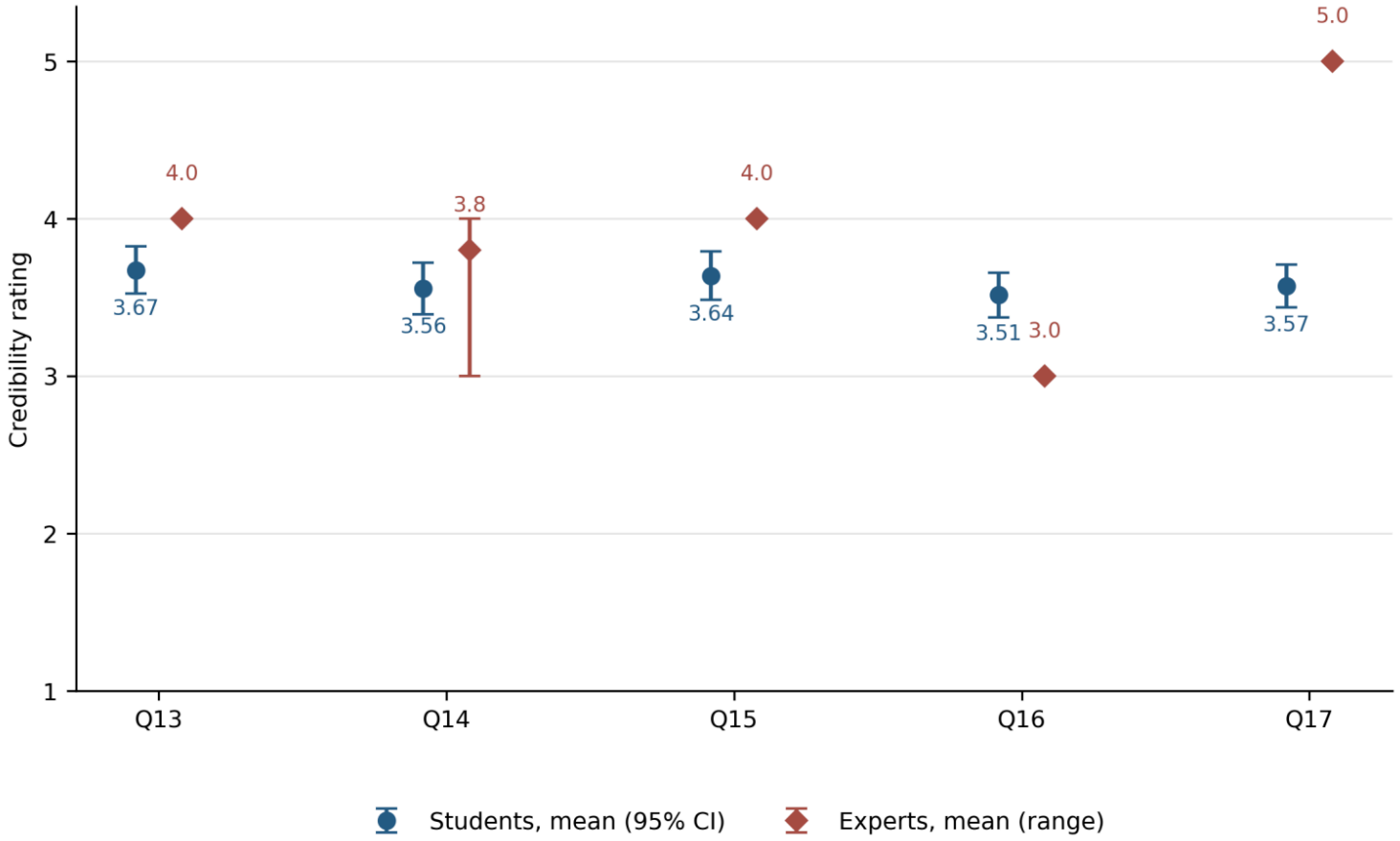


Figure 2

Student and expert credibility ratings

Blue circles show student means with 95% confidence intervals; red diamonds show expert means and ranges. Expert ranges collapse to a point for unanimous ratings. Q13, glaucoma; Q14, keratitis; Q15, hyphema; Q16, retinal neovascularization; Q17, traumatic lens dislocation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.docx