Region - Educational Research and Reviews, 2026, Volume 8, Issue 4
DOI: 10.32629/rerr.v8i4.5344
ISSN Online: 2661-4634
ISSN Print: 2661-4626

LLM-based virtual patient versus traditional teaching in ophthalmology clerkship: a controlled trial

Dingqiao WANG, Jun MAO, Wei DU*

Department of Ophthalmology, The Eighth Affiliated Hospital, Sun Yat-Sen University, Shenzhen 518033, China
*Corresponding Author
Email address: foxfeiniao@163.com

Abstract: Background: Clinical history-taking is central to medical education, yet practice opportunities remain scarce in short specialty rotations such as ophthalmology. No study has directly compared two commercial large language models (LLMs) operating within the same virtual patient framework. Methods: We conducted a quasi-experimental sequential cohort study (TREND-reported). Forty-five medical students completing a five-day ophthalmology clerkship were allocated to one of three groups (n = 15 each): a control group receiving traditional teaching with real patients, and two intervention groups using a retrieval-augmented generation (RAG)-based virtual patient system powered by GPT-4o or Claude 3.5 Sonnet. Primary outcome was history-taking performance (100-point rubric, ICC = 0.89) on Day 2 and Day 5. Secondary outcomes were self-efficacy (SE-12 scale) and student satisfaction, assessed by paired t-test and Fisher's exact test. Results: History-taking improved by 17.11 points (GPT-4o, $p < 0.001$) and 15.17 points (Claude, $p = 0.001$), versus a non-significant 5.87-point change in controls ($p = 0.078$). Self-efficacy gains were 25.77 points (GPT-4o, $p < 0.001$), 21.19 points (Claude, $p = 0.001$), and 2.39 points (control, $p = 0.536$). All intervention students rated satisfaction as basically satisfied or above, versus 80% in the control group (Fisher's exact $p < 0.001$). The two LLM groups did not differ significantly on any measure. Conclusions: Both RAG-based virtual patient systems (GPT-4o and Claude 3.5 Sonnet) improved history-taking performance and clinical self-efficacy within five days; traditional teaching produced no significant change on either measure. The comparable outcomes across LLMs indicate that system architecture and case design matter more than model selection. These findings support the use of LLM-based virtual patients as a practical supplement to bedside teaching in compressed clerkships, subject to confirmation in randomized trials.

Keywords: large language model; virtual patient; history-taking; clinical self-efficacy; ophthalmology clerkship; medical education; retrieval-augmented generation

1 Introduction

Clinical history-taking generates the majority of diagnostic information in patient encounters and is widely regarded as a core competency for medical students [1]. During clerkship training, students acquire this skill primarily through direct patient contact. Suitable encounters are, however, unevenly distributed across rotations, and individual students may have few supervised opportunities within a brief placement. This problem is particularly acute in ophthalmology, where clerkships typically last one to two weeks and patients suitable for teaching cannot always be found.

Standardized patient (SP) programmes offer structured, repeatable encounters but carry substantial costs, require complex scheduling, and are subject to variability in actor performance; these constraints limit their routine use within short clerkship timetables [2-3]. LLM-based virtual patient systems avoid these difficulties: they allow students to practise open-ended history-taking at any time, without logistical overhead. A systematic review of 39 studies found broadly positive effects of such systems on history-taking outcomes but identified recurring limitations: small cohort sizes, heterogeneous outcome measures, and limited evidence in specialty settings [4]. Within ophthalmology specifically, one randomized controlled trial showed that an LLM-based digital patient improved history-taking scores by a mean of 10.50 points over conventional instruction ($p < 0.001$) [5]. To date, no trial has directly compared two distinct commercial LLMs within the same framework, a gap that leaves educators without empirical grounds for model selection.

The present study was designed to address this question directly. We compared two commercially available LLM-based virtual patient systems, one powered by GPT-4o (OpenAI) and one by Claude 3.5 Sonnet (Anthropic), against a conventional teaching control in an ophthalmology clerkship. History-taking performance served as the primary outcome; secondary outcomes were self-efficacy (SE-12) [6], and student satisfaction.

2 Methods

2.1 Study design

This quasi-experimental study used a sequential cohort design (Figure 1). Three successive cohorts of ophthalmology clerkship students were enrolled sequentially. The first cohort served as the control group, followed by the GPT-4o intervention cohort, and then the Claude 3.5 Sonnet intervention cohort. The study was reported in accordance with the TREND (Transparent Reporting of Evaluations with Nonrandomized Designs) statement. Ethical approval was obtained from the institutional review board. All participants provided written informed consent prior to enrollment.

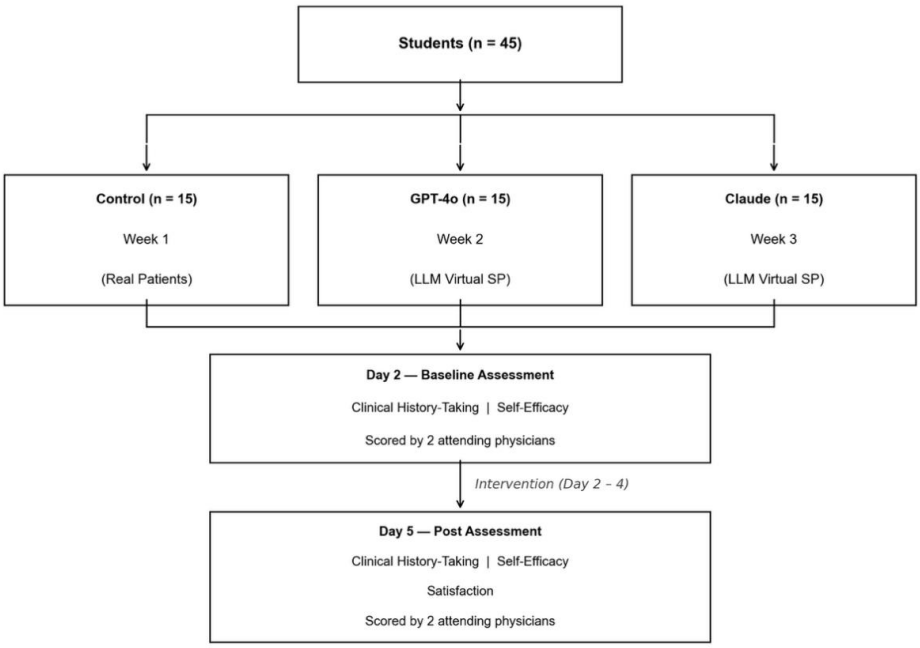

Figure 1. Study flow diagram

Students were allocated to three groups by clerkship week and assessed on Day 2 (pre-intervention) and Day 5 (post-intervention).

2.2 Setting and participants

The study was conducted at the Department of Ophthalmology, the Eighth Affiliated Hospital of Sun Yat-sen University, Shenzhen, China. Eligible participants were undergraduate medical students completing their scheduled

ophthalmology clerkship rotation. Students were excluded if they had prior ophthalmology clinical training, prior experience with LLM-based clinical simulation tools, or any technical inability to access the web application. Fifteen students were enrolled in each of the three sequential cohorts (total n = 45).

2.3 Intervention

Both LLM-based virtual patient systems were built on a RAG (retrieval-augmented generation) architecture [7]. Each system was populated with structured reference documents derived from a real clinical ophthalmology case. Before these documents were entered into the system, all patient-identifiable information (name, date of birth, contact details, and other personal identifiers) was removed and replaced with anonymised placeholders, in accordance with patient privacy requirements. The RAG framework restricted model outputs to the content of these documents, preserving clinical accuracy and reducing hallucinated responses. Students interacted with the virtual patient through free-text dialogue on a smartphone-based web application; the system returned responses derived from the de-identified case records. Each student completed one encounter per day across the five-day rotation; the same case was used throughout for all intervention students.

Students in the control group received conventional bedside teaching, in which two students per session interviewed a real inpatient directly while the others observed and took notes; no virtual patient tools were used.

2.4 Outcome measures

The primary outcome was clinical history-taking performance, assessed using a structured 100-point scoring rubric across seven domains: Chief Complaint (10 points), History of Present Illness (30 points), Ocular History (20 points), Past Medical History (15 points), Family History (10 points), Personal History (5 points), and Communication Skills (10 points). The rubric was developed with reference to established history-taking assessment frameworks [8-11] (see Supplementary appendix A). Two trained raters independently scored each student; inter-rater reliability was excellent (ICC = 0.89, 95% CI 0.84–0.93). Students in the LLM groups were scored based on their chat log records and written history submission; students in the control group were scored based on observed interview performance and written history submission. Assessments were conducted on Day 2 (pre-intervention baseline) and Day 5 (post-intervention).

Clinical self-efficacy was assessed using the SE-12 scale, a validated 12-item instrument (Supplementary appendix B). Each item is rated 1–10 (maximum raw score 120), rescaled to 100 for analysis.

Student satisfaction was assessed using a structured five-item questionnaire covering Learning Effectiveness, Clinical Relevance, Active Engagement, Confidence Building, and Overall Satisfaction, each scored 0–20 points (maximum 100; Supplementary appendix C). Satisfaction was categorized as Very Satisfied (≥ 80), Basically Satisfied (60–79), or Unsatisfied (< 60).

2.5 Statistical analysis

Data were analyzed using SPSS (version 26.0; IBM Corp., Armonk, NY). Continuous variables are presented as mean ± standard deviation. Normality was assessed by the Shapiro–Wilk test. Within-group changes from Day 2 to Day 5 were compared by paired t-test. Satisfaction distributions were compared by Fisher's exact test. A two-tailed $p < 0.05$ was considered statistically significant.

3 Results

3.1 Baseline characteristics

A total of 45 students completed the study with no dropouts (control: n = 15; GPT-4o: n = 15; Claude 3.5 Sonnet: n = 15). Baseline characteristics were comparable across the three cohorts. No significant between-group differences were observed in baseline history-taking scores (control: 55.73 ± 6.20; GPT-4o: 54.50 ± 7.86; Claude: 55.49 ± 9.09; p = 0.915)

or baseline self-efficacy scores (control: 51.27 ± 10.56; GPT-4o: 49.11 ± 11.63; Claude: 51.60 ± 12.23; p = 0.797), confirming group equivalence at the start of the intervention period.

3.2 Clinical history-taking performance

Changes in history-taking performance from Day 2 to Day 5 are presented in Table 1 and Figure 2. Students in the GPT-4o group demonstrated a mean improvement of 17.11 points (54.50 ± 7.86 to 71.61 ± 5.52; p < 0.001). Students in the Claude 3.5 Sonnet group demonstrated a mean improvement of 15.17 points (55.49 ± 9.09 to 70.66 ± 9.11; p = 0.001). The control group showed a mean change of 5.87 points (55.73 ± 6.20 to 61.61 ± 11.19; p = 0.078), which did not reach statistical significance.

Table 1. Pre- and post-intervention clinical history-taking scores

Group	Day 2 (mean ± SD)	Day 5 (mean ± SD)	Change (mean ± SD)	p value (paired t-test)
Control (n = 15)	55.73 ± 6.20	61.61 ± 11.19	5.87 ± 11.97	0.078
Virtual SP, GPT-4o (n = 15)	54.50 ± 7.86	71.61 ± 5.52	17.11 ± 10.47	< 0.001
Virtual SP, Claude 3.5 Sonnet (n = 15)	55.49 ± 9.09	70.66 ± 9.11	15.17 ± 13.37	0.001

Inter-rater reliability: ICC = 0.89 (95% CI 0.84–0.93). Virtual SP = virtual patient system powered by the indicated LLM. SD = standard deviation.

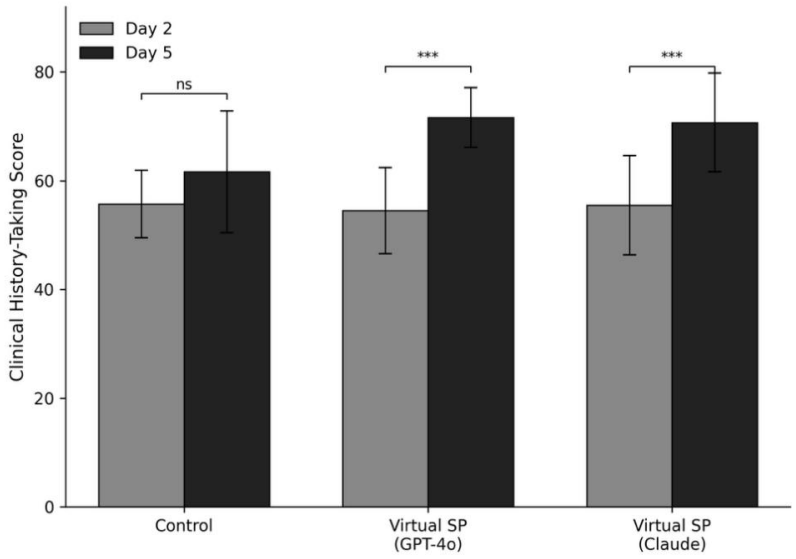

Figure 2. Clinical history-taking scores on Day 2 and Day 5 across groups

3.3 Clinical self-efficacy

SE-12 self-efficacy changes are presented in Table 2. The GPT-4o group showed a mean increase of 25.77 points (49.11 ± 11.63 to 74.88 ± 7.05; p < 0.001). The Claude 3.5 Sonnet group showed a mean increase of 21.19 points (51.60 ± 12.23 to 72.79 ± 9.82; p = 0.001). The control group showed a mean change of 2.39 points (51.27 ± 10.56 to 53.65 ± 11.16; p = 0.536).

Table 2. Pre- and post-intervention clinical self-efficacy scores (SE-12)

Group	Day 2 (mean ± SD)	Day 5 (mean ± SD)	Change (mean ± SD)	p value (paired t-test)
Control (n = 15)	51.27 ± 10.56	53.65 ± 11.16	2.39 ± 14.56	0.536
Virtual SP, GPT-4o (n = 15)	49.11 ± 11.63	74.88 ± 7.05	25.77 ± 13.79	< 0.001
Virtual SP, Claude 3.5 Sonnet (n = 15)	51.60 ± 12.23	72.79 ± 9.82	21.19 ± 18.23	0.001

Scores are out of 100 (SE-12 rescaled from raw maximum 120).

3.4 Student satisfaction

In the GPT-4o group, 10 students (66.7%) were Very Satisfied and 5 (33.3%) were Basically Satisfied; none were Unsatisfied. In the Claude 3.5 Sonnet group, 9 students (60.0%) were Very Satisfied, 6 (40.0%) were Basically Satisfied, and none were Unsatisfied. In the control group, 2 students (13.3%) were Very Satisfied, 10 (66.7%) were Basically Satisfied, and 3 (20.0%) were Unsatisfied. Fisher's exact test confirmed a significant difference in satisfaction distributions across the three groups ($p < 0.001$).

4 Discussion

Both RAG-based virtual patient systems produced significant improvements in history-taking performance and clinical self-efficacy over five days of an ophthalmology clerkship, whereas traditional teaching produced no significant improvement on either measure; satisfaction ratings were also markedly higher in both LLM groups ($p < 0.001$).

The improvement in the LLM groups most plausibly reflects the volume and regularity of practice those systems afforded. In conventional teaching, typically only two students per session interviewed the patient directly while others observed; with the virtual systems, every student completed an independent encounter each day, giving each individual substantially more active exposure over the five-day placement. Grounding the system's responses in de-identified real case documents through the RAG architecture helped maintain clinical accuracy across sessions, and the smartphone-based delivery meant that students could practise whenever time allowed, unconstrained by ward rounds or inpatient availability.

That the two LLMs performed similarly on all measures points to a more fundamental principle: it is the RAG framework architecture and the quality of the underlying case content, rather than the choice of language model, that drive educational outcomes. Institutions can therefore base model selection on practical considerations such as cost, data governance, and local IT infrastructure without expecting meaningful differences in student learning. Luo et al. reported a broadly comparable 10.50-point improvement in ophthalmology history-taking using a different LLM-based system ($p < 0.001$), which reinforces this interpretation.

The study has several limitations that constrain interpretation. The sequential cohort design cannot fully exclude the possibility that unmeasured temporal factors—changes in patient mix, staff supervision, or student cohort characteristics—contributed to the observed differences. Each group comprised only 15 students, providing limited statistical power for between-LLM comparisons; findings should therefore be treated as indicative rather than definitive. The study used a single standardized case at one institution, and no follow-up data were collected to assess whether gains persisted beyond the clerkship week. Confirmation will require randomized controlled trials with larger samples drawn from multiple centers and diverse specialties.

Within the constraints of a five-day ophthalmology clerkship, a RAG-based virtual patient system can improve students' history-taking performance and clinical self-efficacy in ways that conventional bedside teaching did not achieve over the same period. The close agreement between the two LLMs tested here suggests that framework design, not model identity, is the relevant variable for educators. Pending the results of randomized controlled trials at scale, these systems

appear to offer a clinically grounded and logistically feasible supplement to existing teaching provision in compressed specialty placements.

Conflicts of interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Götz S, Fournier J, Tisjlar K, et al. Effects of the quality of medical history taking on diagnostic accuracy[J/OL]. Signa Vitae, 2023, 19(5): 68-74. https://doi.org/10.22514/sv.2023.081

[2] Flanagan O L, Cummings K M. Standardized patients in medical education: A review of the literature[J/OL]. Cureus, 2023, 15(7): e42027. https://doi.org/10.7759/cureus.42027

[3] Hamilton A, Molzahn A, McLemore K. The evolution from standardized to virtual patients in medical education[J/OL]. Cureus, 2024, 16(10): e71224. https://doi.org/10.7759/cureus.71224

[4] Li D, Lutfi S L. Large language model–based virtual patient systems for history-taking in medical education: Comprehensive systematic review[J/OL]. JMIR Medical Informatics, 2026, 1: e79039. https://doi.org/10.2196/79039

[5] Luo M J, Bi S, Pang J, et al. A large language model digital patient system enhances ophthalmology history taking skills[J/OL]. npj Digital Medicine, 2025. https://doi.org/10.1038/s41746-025-01841-6

[6] Axboe M K, Christensen K S, Kofoed P E, et al. Development and validation of a self-efficacy questionnaire (SE-12) measuring the clinical communication skills of health care professionals[J/OL]. BMC Medical Education, 2016, 16(1): 272. https://doi.org/10.1186/s12909-016-0798-7

[7] Wang D, Liang J, Ye J, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study[J/OL]. J Med Internet Res, 2024, 26: e58041. https://doi.org/10.2196/58041

[8] Baker E A, Ledford C H, Fogg L, et al. The IDEA assessment tool: assessing the reporting, diagnostic reasoning, and decision-making skills demonstrated in medical students' hospital admission notes[J/OL]. Teach Learn Med, 2015, 27(2): 163-173. https://doi.org/10.1080/10401334.2015.1011654

[9] Malik T G, Mahboob U, Khan R A, et al. Virtual patients versus standardized patients for improving clinical reasoning skills in ophthalmology residents: a randomized controlled trial[J/OL]. BMC Med Educ, 2024, 24(1): 429. https://doi.org/10.1186/s12909-024-05241-4

[10] Cui Y, Wang L, Dong C, et al. A Bloom's Taxonomy-integrated rotation model enhances clinical reasoning and practical skills in optometry interns during ophthalmology rotation: a randomized controlled trial[J/OL]. Front Med (Lausanne), 2026, 13:1746533. https://doi.org/10.3389/fmed.2026.1746533

[11] Xu L, Xu Q, Liu C, et al. Virtual standardized patients for improving clinical thinking ability training in residents: Randomized controlled trial[J/OL]. JMIR Medical Education, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC12685284/

About the author

Wang Dingqiao (1995–), female, Shenzhen, Guangdong, China; PhD, Attending Physician, Department of Ophthalmology, the Eighth Affiliated Hospital of Sun Yat-sen University, Shenzhen, China. Research interests: glaucoma, oculoplastic surgery, and artificial intelligence in ophthalmology.

Mao Jun (1990–), male, Shenzhen, Guangdong, China; Master's degree, Attending Physician, Department of Ophthalmology, the Eighth Affiliated Hospital of Sun Yat-sen University, Shenzhen, China. Research interests: oculoplastic surgery and lacrimal duct disorders.

Du Wei (1987–), male, Shenzhen, Guangdong, China; PhD, Associate Chief Physician, Master's Supervisor, Deputy Director of the Department of Ophthalmology, the Eighth Affiliated Hospital of Sun Yat-sen University, Shenzhen, China. Research interests: cataract, retinal diseases, and ophthalmic surgery.

Supplementary materials

Supplementary appendix A: History-taking scoring rubric

Supplementary appendix B: SE-12 self-efficacy questionnaire

Supplementary appendix C: Student satisfaction questionnaire

Supplementary Appendix A

Clinical history-taking assessment rubric for ophthalmology clerkship

This rubric was developed by two senior ophthalmologists (associate chief physician or above). Each student is independently scored by two raters. Total score: 100 points.

Domain	Items Assessed	Max Score	Score Awarded	Notes
1. Chief Complaint	Accurately elicits and states the primary symptom(s) and duration. • Identifies the main ocular complaint (e.g., blurred vision, eye pain, redness, floaters) • Records onset and duration correctly	10		
2. History of Present Illness	Systematically explores the presenting complaint using key dimensions: • Onset (sudden vs. gradual) • Duration and progression • Character and severity • Location (unilateral/bilateral) • Aggravating and relieving factors • Associated ocular symptoms (pain, photophobia, discharge, visual changes)	30		
3. Ocular History	Inquires about prior ocular conditions and treatments: • Previous eye diseases or surgeries • Refractive status (glasses, contact lenses) • Prior treatments or medications for eye conditions • History of trauma to the eye	20		
4. Past Medical & Medication History	Explores systemic conditions and medications relevant to ocular health: • Systemic diseases (diabetes mellitus, hypertension, autoimmune disorders) • Current medications (steroids, antimalarials, glaucoma drops) • Known drug allergies	15		
5. Family History	Asks about hereditary ocular or systemic conditions in first-degree relatives: • Glaucoma, age-related macular degeneration, diabetic retinopathy • Hereditary retinal dystrophies	10		
6. Personal History	Collects relevant occupational and lifestyle information: • Occupation and visual demands • UV or chemical exposure • Smoking history (relevant to AMD, cataract)	5		
7. Communication & Organization	Evaluates the manner and structure of the interview: • Logical and systematic sequence of questioning • Appropriate use of open-ended and closed questions • Empathic, respectful communication with the patient • Avoids leading questions; summarizes findings appropriately	10		
TOTAL		**100**	___	

Scoring instructions: Each domain is scored based on the completeness and accuracy of the student's inquiry. Partial credit may be awarded within each domain at the rater's discretion. The final score is the mean of two independent raters' scores.

Rater 1: _____ (Attending Physician) Date: _____

Rater 2: _____ (Resident Physician) Date: _____

Supplementary Appendix B. Clinical Self-Efficacy Questionnaire (SE-12)

Adapted from: Axboe MK, Christensen KS, Kofoed PE, Ammentorp J. Development and validation of a self-efficacy questionnaire (SE-12) measuring the clinical communication skills of health care professionals. BMC Med Educ. 2016;16:272.

Purpose: To assess self-efficacy in clinical history-taking and communication skills before and after the teaching intervention.

Instructions: For each item below, please circle the number that best describes how certain you are that you are able to successfully perform the task.

Response scale: 1 = Very uncertain — 10 = Very certain

No.	How certain are you that you are able to successfully…	1	2	3	4	5	6	7	8	9	10
1	…identify the issues the patient wishes to address during the conversation?										
2	…make an agenda/plan for the conversation with the patient?										
3	…urge the patient to expand on his or her problems/worries?										
4	…listen attentively without interrupting?										
5	…encourage the patient to express thoughts and feelings?										
6	…structure the conversation with the patient?										
7	…demonstrate appropriate non-verbal behavior (eye contact, facial expression, placement, posture, and voicing)?										
8	…show empathy (acknowledge the patient's views and feelings)?										
9	…clarify what the patient knows in order to communicate the right amount of information?										
10	…check the patient's understanding of the information given?										
11	…make a plan based on shared decisions between you and the patient?										
12	…close the conversation by assuring that the patient's questions have been answered?										

Scoring: Sum of all 12 item scores (maximum raw score = 120); rescaled to a 100-point total for analysis.

Supplementary Appendix C. Student Satisfaction Questionnaire

Developed with reference to: (1) Ye et al. Virtual Standardized Patients for Improving Clinical Thinking Ability Training in Residents: Randomized Controlled Trial. *JMIR Med Educ*. 2025.

(2) Agha et al. Evaluation of Medical Students' Satisfaction with Using a Simulation-Based Learning Program. *Cureus*. 2024.

Purpose: To assess student satisfaction with the teaching modality received during the ophthalmology clerkship week.

Completion timing: End of clerkship week (Day 5), completed anonymously.

Instructions: For each item below, assign a score from 0 to 20 based on your personal experience during this clerkship week. The total score (sum of all 5 items) ranges from 0 to 100.

No.	Domain	Item	Score (0–20)
1	Learning Effectiveness	The teaching approach was effective in improving my clinical history-taking skills.	
2	Clinical Relevance	The teaching content was relevant and applicable to real clinical practice.	
3	Active Engagement	The teaching modality facilitated my active participation and engagement in learning.	
4	Confidence Building	The teaching approach helped build my confidence in conducting patient interviews.	
5	Overall Satisfaction	Overall, I am satisfied with my learning experience during this clerkship week.	
Score guide per item: 0 = Completely unsatisfied 5 = Partially unsatisfied 10 = Neutral 15 = Satisfied 20 = Completely satisfied			
		Total Score (0–100)	

Score interpretation:

- Very Satisfied: Total score ≥ 80
- Basically Satisfied: Total score 60–79
- Unsatisfied: Total score < 60

Note: All responses are anonymous and will not affect academic evaluations.