DOI: 10.66069/ojspub.1137260809

Evaluation of the Teaching Effectiveness of Generative Artificial Intelligence-Assisted Surgical Clerkship

Wei Gao[1], Wenjun Zhu[1], Song Li[1], Deyan Fan[2], Zhe Peng[1], Beibei He[1], Shile Wu[1,*]

[1]Department 1 of Hepatobiliary Pancreatic Surgery/Hernia Vascular Surgery, Qinghai Provincial People's Hospital, Xining 810007, Qinghai, China
[2]Qinghai Red Cross Hospital, Xining 810001, Qinghai, China
*Correspondence Author

Abstract: ***Background:*** ***Clinical clerkship is a critical stage linking preclinical theory with clinical practice in medical education. However, traditional surgical clerkship faces limitations in case resources, teaching staff capacity, and opportunities for skill training. This study developed a four-step generative artificial intelligence (GenAI)-assisted surgical clerkship teaching model and evaluated its effects on medical students' theoretical knowledge, operational skills, and clinical thinking. Methods: A randomized controlled design was adopted. A total of 47 undergraduate clinical medicine students who underwent surgical clerkship in the Department of General Surgery of Qinghai University Affiliated Hospital between July 2025 and February 2026 were enrolled and randomly assigned to an experimental group (n = 24) and a control group (n = 23). The experimental group adopted a four-step GenAI (DeepSeek)-assisted teaching model consisting of AI scenario introduction, AI case deduction, AI skill analysis, and AI immediate feedback, whereas the control group received traditional teaching. Theoretical examination, objective structured clinical examination (OSCE), clinical case analysis, and a learning satisfaction questionnaire were compared between the two groups after the clerkship. Results: Baseline data were comparable between the two groups (all P > 0.05). Scores of theoretical knowledge, OSCE, clinical case analysis, and learning satisfaction in the experimental group were significantly higher than those in the control group (all P < 0.05). Conclusions: The four-step GenAI-assisted surgical clerkship teaching model can effectively improve students' theoretical knowledge, operational skills, clinical thinking, and their learning experience. This model provides empirical evidence for the standardized application of GenAI in surgical clerkship teaching.***

Keywords: Generative artificial intelligence, General surgery, Clinical clerkship, Teaching model, Teaching effectiveness.

1. Introduction

Surgery is an indispensable component of clinical medicine, and clinical clerkship represents a crucial stage for medical students to bridge theoretical knowledge with clinical practice and develop clinical competence. However, traditional surgical clerkship teaching is commonly constrained by limited case resources, insufficient teaching staff capacity, passive student engagement, and unequal opportunities for skills training [1]. This problem is particularly pronounced in primary teaching hospitals, where clerkship students have limited exposure to typical cases and classroom teaching may be disconnected from clinical practice, compromising educational outcomes. In recent years, generative artificial intelligence (GenAI), represented by large language models (LLMs), has rapidly emerged in medical education and offers new opportunities to address these challenges [2-4]. GenAI can generate realistic clinical scenarios, simulate case-based reasoning, and provide immediate interactive feedback, demonstrating broad application prospects in medical knowledge inquiry, clinical reasoning training, virtual patient simulation, and skills training [5–7].

International studies have shown that ChatGPT and other GenAI tools can answer medical questions, with some models achieving passing scores on the United States Medical Licensing Examination (USMLE) [7]. Qu et al. discussed the implications of ChatGPT for transforming medical education models [8], and Zhao et al. combined AI with blended teaching in surgery education with favorable outcomes [9]. Huang et al. integrated ChatGPT into a surgical clerkship teaching model and significantly improved students' learning compliance and satisfaction [10]; Gan et al. conducted a randomized controlled trial of ChatGPT in orthopedic education and reported improved student engagement and course organization satisfaction [11]; and Ba et al. compared ChatGPT-assisted teaching with traditional methods in clinical skills training [12]. Chinese scholars have also systematically explored the current status, potential, and challenges of GenAI in medical education [2,3,13], as well as its ethical risks [14]. A meta-analysis of 17 randomized controlled trials showed that AI-assisted teaching significantly improved theoretical scores (SMD = 1.35, 95% CI 0.78–1.92) and practical/operational skills (SMD = 2.45, 95% CI 0.89–4.02) in medical education [15].

Despite the growing body of research on GenAI in medical education, most studies focused on single teaching component or single disease entity, and systematic evaluations of its overall effects across knowledge, skills, and clinical thinking in the specific context of surgical clerkship remain limited. So, this study developed a four-step GenAI (DeepSeek)-assisted surgical clerkship teaching model and used a randomized controlled design to evaluate its teaching effectiveness from four dimensions—theoretical knowledge, operational skills (OSCE), clinical thinking, and learning satisfaction—aiming to provide empirical evidence for the standardized application of GenAI in surgical clerkship teaching.

2. Methods

2.1 Participants

Undergraduate clinical medicine students who underwent surgical clerkship in the Department of General Surgery of

Qinghai University Affiliated Hospital between July 2025 and February 2026 (two clerkship periods, summer and winter) were recruited. Inclusion criteria were: (1) full-time undergraduate clinical medicine students; (2) first participation in surgical clerkship; and (3) understanding of the study and provision of written informed consent. Exclusion criteria were: (1) previous experience with GenAI-assisted teaching; (2) failure to complete the full clerkship curriculum; and (3) absence from assessments. A total of 53 students were initially enrolled; six were excluded for incomplete curriculum or assessment absence, leaving 47 students for analysis (experimental group, n = 24; control group, n = 23). The study was approved by the Ethics Committee of Qinghai University Affiliated Hospital (Approval No. QHDFY-2025-05-10).

2.2 Teaching Models

The experimental group adopted a four-step teaching model — "AI scenario introduction, AI case deduction, AI skill analysis, and AI immediate feedback"—using DeepSeek, a Chinese large language model, as the primary teaching assistance tool (Figure 1).

1) AI scenario introduction (pre-class): Teachers used DeepSeek to generate clinical scenario cases related to the day's clerkship theme (e.g., acute appendicitis, cholelithiasis, fractures) as pre-class preparation tasks.

2) AI case deduction (in-class): A combined "bedside observation plus AI case deduction" approach was adopted. During bedside sessions, students observed patients' signs and practiced history-taking and physical examination; during AI sessions, students reasoned through diagnostic pathways using virtual cases generated by DeepSeek, which posed progressive questions and prompted differential diagnoses, guiding students through the complete clinical reasoning chain of "symptoms and signs—diagnosis—management plan."

3) AI skill analysis (skills training): For core skills including aseptic technique, suturing and knot tying, dressing changes, and physical examination, AI assisted in generating structured analyses of standard operating points and reminders of common errors, allowing students to repeatedly observe and compare their practice.

4) AI immediate feedback (post-class): Students used DeepSeek for knowledge consolidation, self-directed questioning, and mock tests, with AI providing immediate feedback on answers and generating personalized analysis of errors and review suggestions; supervising teachers reviewed students' AI usage records and provided targeted guidance on weak areas.

The control group received traditional teaching consisting primarily of classroom lectures, bedside teaching, and case discussions, without the use of GenAI tools. The two groups had identical clerkship content, class hours, and teaching staff qualifications, differing only in the teaching model.

2.3 Outcome Measures

(1) Theoretical knowledge: A unified closed-book examination was administered after the clerkship (full score 100 points), covering general surgical fundamentals and diseases involved in the clerkship. (2) Operational skills (OSCE): Four stations—aseptic technique, suturing and knot tying, dressing changes, and physical examination—were set, each scored out of 100 points; the mean of the four stations constituted the OSCE total score. Assessments were scored blindly by teachers not involved in the teaching, using a unified scoring rubric. (3) Clinical thinking: A typical case analysis question was used to evaluate the completeness and logical coherence of diagnostic reasoning, differential diagnosis, and management plan (full score 100 points), scored independently by two teachers and averaged. (4) Learning satisfaction: A self-designed questionnaire (Likert 5-point scale, 1–5) covering four items—teaching method, learning interest, self-directed learning ability, and overall satisfaction—was used, with the mean score taken as the satisfaction score. The questionnaire demonstrated good reliability and validity in a pilot test (Cronbach's $\alpha = 0.87$).

2.4 Statistical Analysis

Data were analyzed using SPSS version 26.0 (IBM Corp., Armonk, NY, USA). Continuous variables are presented as mean ± standard deviation (SD) and compared between groups using the independent samples t test. Categorical variables are presented as frequencies (percentages) and compared using the chi-square test. A two-tailed $P < 0.05$ was considered statistically significant.

3. Results

3.1 Baseline Characteristics

No significant differences were found between the two groups in sex, age, or admission examination scores (all $P > 0.05$), indicating comparability between groups (Table 1).

Table 1: Baseline characteristics of clerkship students in the two groups

Characteristic	Experimental (n = 24)	Control (n = 23)	t/χ^2	P
Age (years, mean ± SD)	22.3 ± 1.1	22.1 ± 1.2	0.61	0.55
Male, n (%)	13 (54.2)	12 (52.2)	0.03	0.87
Admission score (mean ± SD)	84.6 ± 4.2	85.1 ± 4.0	0.42	0.68

3.2 Teaching Effectiveness

The experimental group achieved significantly higher scores in theoretical knowledge, OSCE, and clinical case analysis than the control group (Table 2), and also reported significantly higher learning satisfaction (Table 3).

Table 2: Comparison of assessment scores between the two groups (points, mean ± SD)

Dimension	Experimental (n = 24)	Control (n = 23)	t	P
Theoretical knowledge	86.4 ± 6.3	81.5 ± 6.9	2.56	< 0.05
Operational skills (OSCE)	88.1 ± 5.5	82.7 ± 6.2	3.14	< 0.01
Clinical thinking (case analysis)	84.6 ± 6.6	79.5 ± 7.3	2.48	< 0.05

Table 3: Comparison of learning satisfaction between the two groups (points, mean ± SD)

Measure	Experimental (n = 24)	Control (n = 23)	t	P
Learning satisfaction	4.6 ± 0.4	4.1 ± 0.5	3.75	< 0.01

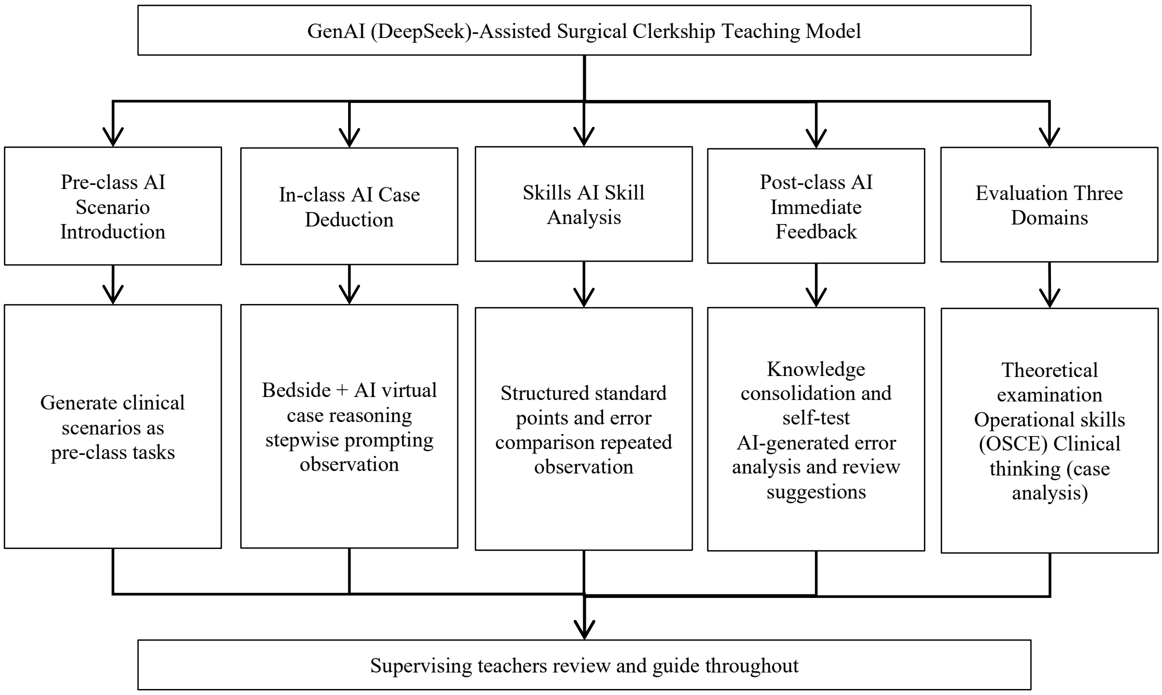

Figure 1: Flowchart of the GenAI-assisted surgical clerkship teaching model

4. Discussion

This study demonstrated that a four-step GenAI (DeepSeek)-assisted surgical clerkship teaching model significantly improved students' theoretical knowledge, operational skills, and clinical thinking and enhanced their learning experience, consistent with the conclusions of systematic reviews and overviews of AI-assisted medical education [2–4,15,17].

Regarding theoretical knowledge, the experimental group achieved significantly higher scores than the control group. From an educational theory perspective, the "AI scenario introduction" component presents abstract surgical concepts within realistic clinical contexts, aligning with the core tenets of situated learning theory, which facilitates activation of students' prior knowledge and promotes deep processing and transfer of new knowledge. Furthermore, the "AI immediate feedback" component generates personalized error analysis and review suggestions tailored to individual students' weaknesses, consistent with differentiated instruction and addressing the practical limitation of finite teaching staff capacity [13]. These findings are consistent with previous research suggesting that AI-assisted teaching, including ChatGPT-assisted problem-based learning, can improve theoretical scores [15,20].

Regarding operational skills, the experimental group achieved superior OSCE total scores, with particularly pronounced advantages at the aseptic technique and suturing/knot-tying stations. The acquisition of procedural skills follows the cognitive–associative–autonomous three-stage model. The "AI skill analysis" component deconstructs standard procedures into key steps with comparisons of common errors, providing students with low-cost, repeatable, and immediately feedback-driven opportunities for observation and practice. The combination of AI with bedside teaching also enables timely validation of theoretical points in real operations, facilitating progression from the cognitive to the autonomous stage, consistent with research on ChatGPT-assisted clinical skills training and surgical education [6,12,19].

Regarding clinical thinking, the experimental group outperformed the control group in case analysis scores. The "AI case deduction" component simulates the progressive cognitive process of clinical decision-making: through stepwise questioning and prompts for differential diagnosis, students are guided to develop a complete reasoning pathway of "symptoms—signs—diagnosis—management," thereby training critical thinking and integrative analysis. Notably, previous research indicates that AI remains inferior to human experts in some reasoning tasks [5]; therefore, in this study, AI case deduction was always conducted under the guidance of supervising teachers, positioning AI as a "thinking partner" rather than a "source of answers," consistent with approaches to GenAI-assisted clinical reasoning training [16].

Regarding learning experience, the experimental group reported significantly higher satisfaction than the control group, consistent with the findings of Huang et al. [10], Gan et al. [11], and Ba et al. [12], suggesting that GenAI-assisted teaching enhances students' learning initiative and sense of achievement, and that its immediacy, interactivity, and personalization align with the learning preferences of contemporary medical students. Eysenbach et al. also early called attention to the role and impact of generative AI in medical education [18].

Compared with previous studies, the innovations of this study are threefold. First, it constructed a systematic four-component teaching model covering the entire "pre-class–in-class–post-class" process, which is more comprehensive and operational than previous applications focused on a single component or single disease. Second, it adopted a randomized controlled design and performed multidimensional evaluation across three core domains — knowledge, skills, and clinical thinking — combined with learning satisfaction, providing a more comprehensive reflection of teaching effectiveness than satisfaction-based studies. Third, by targeting the specific context of surgical clerkship in a teaching hospital and addressing its practical pain points, the findings offer direct reference value for comparable teaching hospitals.

Regarding tool selection, DeepSeek was chosen primarily for its domestic accessibility, free open access, good comprehension of Chinese-language contexts, and alignment with the Chinese medical textbook system. It should be noted that general-purpose LLMs carry risks such as generating inaccurate or outdated content and fostering overreliance that may weaken independent thinking [8,14]. Accordingly, risk-control measures were implemented: supervising teachers reviewed AI-generated content before use; students were encouraged to critically appraise rather than blindly accept AI output; and AI's role was limited to teaching assistance, not replacing teachers' professional judgment, thereby mitigating risks to educational equity and academic integrity [14].

This study has several limitations. First, it was a single-center study with a limited sample size, cautioning against broad generalization. Second, a general-purpose LLM was used, and differences in performance among tools may affect comparability of results. Third, the evaluation focused on short-term outcomes at the end of the clerkship, without long-term follow-up of clinical competence or knowledge retention. Fourth, although baseline characteristics were comparable and teaching content and staff were consistent, confounding factors such as learning motivation cannot be entirely excluded. Fifth, the learning satisfaction questionnaire was self-designed and subject to some subjectivity. Future research should conduct multicenter, larger-scale randomized controlled trials with extended follow-up and explore specialty-specific AI tools or retrieval-augmented generation (RAG)-based specialty knowledge bases to further strengthen the level of evidence [17].

In conclusion, the four-step GenAI-assisted surgical clerkship teaching model can systematically improve students' knowledge, skills, and clinical thinking and enhance their learning experience, providing valuable empirical evidence for the standardized application of GenAI in medical education and clerkship teaching in teaching hospitals. The model warrants further promotion and application.

5. Conclusion

The four-step GenAI (DeepSeek)-assisted surgical clerkship teaching model effectively improves undergraduate medical students' theoretical knowledge, operational skills, and clinical thinking, and enhances their learning satisfaction. It offers a practical and replicable framework for integrating generative artificial intelligence into surgical clerkship teaching, and warrants broader application in teaching hospital settings.

Fund Project

Education and Teaching Research Reform Project of Qinghai University School of Medicine (qyjg-202519); Kunlun Talent Action Plan of Qinghai Province; Supported by the "Kunlun Talents · Science and Technology Leading Talents" project in Qinghai Province.

References

[1] Wang GD. Current situation and reform thinking of clerkship teaching for clinical medicine majors [J]. Medical Education Research and Practice, 2021, 29(4): 580-583.
[2] Fan DR, Wang QQ, Shen A, et al. From ChatGPT to Sora: potential and challenges of artificial intelligence application in medical education [J]. China Medical Education Technology, 2025, 39(1): 33-40.
[3] Huang MF, Hou QH, Zhang W. Application status and future trends of generative artificial intelligence in medical education [J]. Medicine and Society, 2025, 38(1): 29-34.
[4] Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84 [J]. Medical Teacher, 2024, 46(4): 446-470.
[5] Zhou ZH, Song JL, Liu JC, et al. Construction and clinical teaching application of virtual patient system based on artificial intelligence large language model technology [J]. Medical, 2024, 34(7): 833-842.
[6] Zhang S, Meng LS, Liu JJ, et al. DeepSeek in the classroom: application exploration and reflection of AI agents in organ-system-based integrated curriculum teaching [J]. China Medical Education Technology, 2025, 39(4): 432-436.
[7] Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models [J]. PLOS Digital Health, 2023, 2(2): e0000198.
[8] Qu X, Yang JM, Chen T, et al. Reflections on the transformation of medical education models by ChatGPT [J]. Journal of Sichuan University (Medical Sciences), 2023, 54(5): 937-940.
[9] Zhao Y, Qu HM, Li ZG, et al. Application of AI combined with blended teaching mode in surgery teaching [J]. China Continuing Medical Education, 2024, 16(6): 88-92.
[10] Huang Y, Wang L, Teng W. Implementation and evaluation of an optimized surgical clerkship teaching model utilizing ChatGPT [J]. BMC Medical Education, 2024, 24: 1540.
[11] Gan W, Ouyang J, Li H, et al. Integrating ChatGPT in orthopedic education for medical undergraduates: randomized controlled trial [J]. Journal of Medical Internet Research, 2024, 26: e57037.
[12] Ba HJ, Zhang LJ, Yi ZZ. Enhancing clinical skills in pediatric trainees: a comparative study of ChatGPT-assisted and traditional teaching methods [J]. BMC Medical Education, 2024, 24(1): 558.
[13] Zhang W, Chen X. Application prospect of generative artificial intelligence in medical education [J]. Chinese Journal of Medical Education Exploration, 2024, 23(1): 12-16.
[14] Wang RJ, Wang D. Ethical risks of ChatGPT in medical education and coping strategies [J]. Medicine and Philosophy, 2024, 45(2): 76-81.

[15] Wei W, Luo GM, Liu RM, et al. A meta-analysis of the application effect of artificial intelligence-assisted teaching in medical education [J]. Medical, 2025, 35(10): 1242-1248.
[16] 16. Chen G, He RQ, Qin DY, et al. Application of artificial intelligence in clinical practice, scientific research and education of medicine [J]. Journal of Guangxi Medical University, 2025, 42(4): 489-499.
[17] Vrdoljak J, et al. A review of large language models in medical education: applications, challenges, and future prospects [J]. Healthcare, 2025, 13(6): 603.
[18] Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers [J]. JMIR Medical Education, 2023, 9: e46885.
[19] Rao L, et al. The use of generative artificial intelligence in surgical training [J]. Smart Medicine, 2024, 3(1): e20240050.
[20] Zeng H, Zhu ZW, Hu J, et al. Application of ChatGPT-assisted problem-based learning teaching method in clinical medical education [J]. BMC Medical Education, 2025, 25(1): 50.