Prehospital and Disaster Medicine

www.cambridge.org/pdm

Enhancing Emergency Medical Response Education through Generative AI-Powered Game-Based Learning: A Retrospective Comparative Study

Yaning Lai MSc[1,*], Xiaoqin Lai MSc[2,*] and Hai Hu MD[3]

[1]Academic Affairs Department, West China Hospital/West China School of Medicine, Sichuan University, Chengdu, China; [2]Department of Thoracic Surgery, West China Tianfu Hospital/Day Surgery Center, West China Hospital, Sichuan University, Chengdu, Sichuan, China and [3]Emergency Management Office, West China Hospital/West China School of Medicine, Sichuan University, Chengdu, China

Original Research

Cite this article: Lai Y, Lai X, and Hu H. (2026) Enhancing Emergency Medical Response Education through Generative AI-Powered Game-Based Learning: A Retrospective Comparative Study. *Prehospital and Disaster Medicine*. **41**(e14), 1–7. doi: 10.1017/S1049023X26108978

Received: 19 October 2025 UTC
Revised: 1 June 2026 UTC
Accepted: 7 June 2026 UTC

Keywords:
Disaster Medicine; Gamification; Generative Artificial Intelligence; Medical Education; Medical Student

Abbreviations:
AI, Artificial Intelligence; AI-GBL, Generative AI-Powered Game-Based Learning; API, Application Programming Interface; EMR, Emergency Medical Response; GBL, Game-Based Learning; LBL, Lecture-Based Learning; LSD, Least Significant Difference; (M)ANOVA, (Multivariate) Analysis of Variance; RAG, Retrieval-Augmented Generation

Corresponding author:
Hai Hu, MD
Emergency Management Office
West China Hospital/West China School of Medicine
Sichuan University, Chengdu, China
China International Emergency Medical Team (Sichuan)
Chengdu City, Sichuan Province, China
E-mail: huhai@wchscu.cn

*Both Yaning Lai and Xiaoqin Lai contributed equally to this work and should be regarded as co-first authors.

Abstract

Objective: Traditional lecture-based learning (LBL) is often insufficient for cultivating the practical decision-making skills required in high-stakes environments like Emergency Medical Response (EMR). While game-based learning (GBL) offers an immersive alternative, it can lack immediate expert guidance. This study addresses this gap by exploring the integration of generative Artificial Intelligence (AI) as an "intelligent tutor" within GBL. The objective was to evaluate and compare the effectiveness of LBL, GBL, and generative AI-powered game-based learning (AI-GBL) on medical students' knowledge acquisition, retention, learning motivation, and cognitive load in an EMR course.

Methods: A retrospective, comparative study was conducted with 86 medical students from three consecutive cohorts (2022-2024), each exposed to one of the three teaching modalities (n = 29 LBL, n = 28 GBL, n = 29 AI-GBL). Knowledge was assessed via pre-test, post-test, and final-test scores with a maximum score of 10 points. Student feedback was collected for learning motivation, cognitive load, and technology acceptance.

Results: For immediate knowledge acquisition, both GBL (mean difference = 1.124/10 points; 95% CI [0.297, 1.952]; P = 0.008) and AI-GBL (mean difference = 0.897/10 points; 95% CI [0.076, 1.717]; P = 0.033) significantly outperformed LBL. For delayed knowledge retention, the AI-GBL group demonstrated significantly superior retention compared to both the GBL group (mean difference = 0.689 points; unadjusted 95% CI [0.080, 1.299]) and the LBL group (mean difference = 1.310 points; unadjusted 95% CI [0.706, 1.915]). The AI-GBL group also reported significantly lower cognitive load than the GBL group (mean difference = −0.273 points; unadjusted 95% CI [−0.456, −0.090]). Finally, students perceived the AI-powered approach as significantly more useful than the standard game-based approach (mean difference = 0.513 points; unadjusted 95% CI [0.137, 0.889]).

Conclusion: The AI-enhanced GBL model for EMR training improves knowledge acquisition and retention while reducing cognitive load, representing a promising approach for developing proficiency in complex, high-stakes medical competencies.

Introduction

Background and Significance

Disaster events, whether natural or man-made, pose formidable challenges to global public health systems.[1] In these scenarios, a rapid, orderly, and highly effective Emergency Medical Response (EMR) is paramount for reducing mortality and morbidity.[2] Consequently, cultivating medical professionals equipped with robust knowledge, superior decision-making skills, and strong psychological resilience for disaster relief has become a core mission of modern medical education.[3] However, the complexity of EMR – which demands multidisciplinary collaboration, operates under severe resource constraints, and unfolds in highly uncertain environments – presents significant challenges to traditional pedagogical models.

Limitations of Traditional Teaching Models

For decades, the lecture-based format has been the cornerstone of medical education. Yet, its limitations are increasingly evident in highly practical domains like EMR. Passive knowledge transfer struggles to simulate the urgency and chaos of a disaster scene. Students often understand theoretical concepts in isolation but fail to effectively translate this knowledge into real-time decision-making capabilities.[4] Initial teaching observations corroborate this,

© The Author(s), 2026. Published by Cambridge University Press on behalf of World Association for Disaster and Emergency Medicine.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 14:08:09, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1049023X26108978

demonstrating that a lecture-only approach was ineffective in stimulating student interest and fostering comprehensive response skills.

Exploring and Integrating Emerging Pedagogical Methods

To address the shortcomings of traditional methods, educators have explored more immersive and interactive teaching approaches. Game-based learning (GBL), particularly tabletop exercises based on role-playing, significantly enhances learner engagement and practical application by placing students in simulated disaster scenarios.[5] However, this model introduces new challenges: students may encounter knowledge barriers that halt their progress during the exercise, or they may lack immediate, expert guidance to facilitate deep reflection, thereby limiting the depth and breadth of their learning.

Concurrently, the rise of generative Artificial Intelligence (AI) has introduced revolutionary tools to the educational landscape. Generative AI agents, capable of playing the role of an expert, can provide personalized, on-demand question-and-answer support.[6] This capability perfectly addresses the guidance gap often found in GBL. It was postulated that integrating a generative AI agent as an "intelligent tutor" or "on-site advisor" within a tabletop exercise would not only provide real-time answers to student queries, but would also guide them toward deeper critical thinking, thereby creating a synergy between the two pedagogical methods.

Research Gap and Novelty of This Study

While the applications of GBL and generative AI in education have been individually explored, there remains a significant gap in empirical research on their deep integration to form a "Generative AI-Powered Game-Based Learning" (AI-GBL) framework for EMR education. This study directly addresses this gap. The core innovation lies not in simply using AI as an ancillary tool, but in embedding it within the GBL process to create a synergistic and effective educational ecosystem designed to solve the fundamental problem of disconnecting theory from practice in complex scenarios.

Study Objectives and Hypotheses

Based on the aforementioned background, this study aims to systematically evaluate and compare the effectiveness of three distinct teaching models – traditional lecture, role-playing-based GBL, and generative AI-powered GBL – in an EMR course through a retrospective comparative study. It was hypothesized that, compared to traditional lectures and GBL alone, the AI-GBL model would significantly improve students' knowledge acquisition, delayed knowledge retention, and comprehensive problem-solving abilities, while also yielding more positive feedback on their learning experience.

Methods

Study Design

This study employed a retrospective, comparative design to analyze educational data from three consecutive cohorts of medical students enrolled in a mandatory EMR course from 2022 through 2024 at a medical school of a comprehensive university. Each cohort was exposed to a different teaching modality, allowing for a naturalistic comparison of educational outcomes. The study was conducted in accordance with the Declaration of Helsinki and received approval from the Biomedical Ethics Committee of West China Hospital, Sichuan University, Chengdu, China (the institutional review board - Number: 2024[1811]). The requirement for individual informed consent was waived due to the retrospective and anonymized nature of the data analysis.

Participants and Setting

A total of 86 pre-internship medical students participated across the three cohorts (Table 1). This study employed a convenience sampling method, including all students enrolled in the mandatory EMR course across three consecutive academic years. The course was a core component of the Disaster Medicine curriculum, delivered over a single semester. Baseline demographic data, including age and gender, were collected from academic records to ensure comparability between the groups. Given the retrospective design, the sample size was based on the available cohort data rather than an a priori calculation. However, a post-hoc power analysis confirmed that the sample was adequately powered to detect differences across the three arms. Based on the observed effect sizes, the calculated Cohen's *f* values were 0.314 for Change1 (improvement from pre-test to post-test) and 0.473 for Change2 (retention from post-test to final-test). With an alpha of 0.05, three groups, and a total sample size of 86, the achieved statistical power was 0.881 for Change1 and 0.995 for Change2, both exceeding the conventional threshold of 0.80. The detailed calculations for the effect size conversion and power analysis are provided in Supplementary Materials.

Educational Interventions

All three interventions were designed to achieve identical learning objectives, which covered core EMR principles such as the Incident Command System, triage protocols,[7] mass-casualty incident management, and specific injury management in disaster settings. The total contact time for each intervention was standardized to four class hours.

Group 1: Lecture-Based Learning (LBL)—This cohort received traditional didactic instruction. The session was delivered by an experienced faculty member using a slide-based presentation. The content focused on theoretical knowledge and case study descriptions. A 15-minute question-and-answer session was held at the end of the lecture.

Group 2: Game-Based Learning (GBL) with Tabletop Exercise—This cohort participated in a role-playing tabletop exercise simulating a major earthquake scenario. Students were divided into small groups (five-to-six students) and assigned key roles. They were given a map, scenario cards, and resource tokens to manage the incident. An instructor acted as a facilitator, providing the scenario narrative and adjudicating actions, but offered minimal direct instructional feedback to preserve the simulation's fidelity. The introduction to the educational game is provided in the Supplementary Materials.

Group 3: Generative AI-Powered Game-Based Learning (AI-GBL)—This cohort engaged in the same tabletop exercise as Group 2. The key difference was that they were provided with access to a custom-developed generative AI agent, which they could use on their personal laptops or mobile phones. The AI agent was built on the ChatGLM (Zhipu AI; Beijing, China) application programming

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 14:08:09, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1049023X26108978

interface (API). It was specifically prompt-engineered to act as an "EMR Expert Advisor." Crucially, to ensure the accuracy and reliability of the guidance, the agent utilized a Retrieval-Augmented Generation (RAG) architecture. This approach strictly confined the AI's knowledge base to a curated repository of evidence-based Disaster Medicine protocols and the course's core content, preventing hallucinations. Students could interact with the AI in real-time by typing questions related to their situation, such as "What are the triage criteria for a patient with a head injury and altered mental status?" or "How should we prioritize resource allocation between the treatment and transport areas?". The AI provided immediate, concise, and context-aware guidance, thereby serving as an on-demand knowledge resource and decision-support tool. The introduction to the generative AI agent is provided in the Supplementary Materials.

Data Collection and Outcome Measures

The primary outcome was the difference in immediate knowledge retention between the three groups. The secondary outcome was the difference in delayed knowledge retention between the three groups. Exploratory outcomes included cognitive load, learning motivation, and technology acceptance. Data were collected from two sources: knowledge assessments and student feedback questionnaires.

Knowledge Assessment—Student knowledge was evaluated using a 20-question written examination, with each question worth 0.5 points for a total possible score of 10 points. The exam utilized a multiple-choice format designed to test factual recall. To ensure content validity, the initial item pool was developed by five faculty members from the teaching team based on the course learning objectives. Subsequently, the questionnaire underwent a content review by an independent panel of three senior EMR educators who did not participate in the item generation. This panel evaluated the relevance, clarity, and alignment of each item with the curriculum, and their feedback was incorporated to finalize the 20 questions. This format allowed for objective scoring using a standard answer key, thereby eliminating potential inter-rater grading bias. It was administered at three time points:

- Pre-Test: Immediately before the class to establish baseline knowledge.
- Post-Test: Immediately after the class to measure immediate knowledge gain.
- Final-Test: Three weeks after the class, with no intervening exposure to the material, to assess knowledge retention.

Student Feedback Questionnaire—A post-course anonymous questionnaire was administered to gather students' perceptions of their learning experience. The study employed three questionnaires, detailed in Supplementary Materials. For all participants, a learning motivation scale (seven items; adopted from Hwang and Chang[8]) and a cognitive load scale (eight items; adopted from Hwang, et al[9]) were administered. As the technology acceptance questionnaire was only relevant to the novel teaching methods, it was administered exclusively to the game-based groups. This scale (13 items; adopted from Chu, et al[10]) assessed perceived ease-of-use and usefulness. All questionnaires used a six-point Likert scale.

These instruments were selected based on their established validation in prior research. For instance, the cognitive load scale was adapted from Sweller, et al[11] and has demonstrated high reliability in previous studies.[9] Similarly, the perception and attitude questionnaires used by Chu, et al[10] were reviewed by experts to ensure content validity. As the medium of instruction for these medical students is English, the original English versions of the questionnaires were utilized to avoid translation bias. To ensure contextual relevance for the EMR setting in the current study, the items were reviewed by a panel of three senior EMR educators. Minor contextual adaptations were made to the scenario descriptions based on their feedback to ensure clarity and appropriateness. Internal consistency was assessed using Cronbach's alpha. Further details on the questionnaire items, their sources, and validation processes are provided in Supplementary Materials. The total scores for each scale were calculated as the mean of their constituent items. Given the sample sizes and the nature of the scales (averages of multiple items), the data were considered suitable for parametric analysis.

Statistical Analysis

All statistical analyses were performed using IBM SPSS Statistics (version 20.0; IBM Inc.; Armonk, New York USA).

To examine the effects of the three teaching modalities on student performance, a multivariate analysis of variance (MANOVA) was conducted. The MANOVA used the teaching group (LBL, GBL, AI-GBL) as the independent variable and two dependent variables representing performance change: Change1 (post-test score minus pre-test score) and Change 2 (final-test score minus post-test score). Prior to the MANOVA, the assumption of equality of covariance matrices was tested using Box's M test. The assumption of homogeneity of variance was also checked for each dependent variable using Levene's test. If the overall MANOVA yielded a significant result, univariate follow-up ANOVAs were conducted for each dependent variable to determine where the group differences lay. For any significant univariate ANOVA, post-hoc pairwise comparisons were performed using Fisher's Least Significant Difference (LSD) test to identify specific differences between the groups.

For the questionnaire data, a one-way ANOVA was conducted to compare the mean scores across the three groups. For significant findings, Fisher's LSD test was used for pairwise comparisons.

A two-tailed P value of less than 0.05 was considered statistically significant for all analyses. The widths of the 95% confidence intervals (CI) for secondary and exploratory outcomes have not been adjusted for multiplicity, and the intervals should not be used in place of hypothesis testing.

Results

Baseline Characteristics of Study Participants

A total of 86 medical students were included in this retrospective analysis, with 29 students in the LBL group, 28 in the GBL group, and 29 in the AI-GBL group. The baseline characteristics of study participants are shown in Table 1.

Pairwise Comparisons of Group Scores

The descriptive statistics for the knowledge score changes (Change1 and Change2) across the three teaching groups are presented in Table 2.

Subsequently, a MANOVA was conducted to examine the differences among the three teaching groups (LBL, GBL, and AI-GBL) on two dependent variables: Change1 (post-test minus pre-test scores) and Change2 (final-test minus post-test scores); Table 3.

The Box's M test indicated that the assumption of equality of covariance matrices was violated (Box's M = 100.564; F = 16.176;

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 14:08:09, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1049023X26108978

Table 1. Baseline Characteristics of Study Participants

Characteristic	LBL (n = 29)	GBL (n = 28)	AI-GBL (n = 29)
Age (Years), Mean (SD)	23.03 (SD = 0.63)	22.86 (SD = 0.76)	23.10 (SD = 0.67)
Sex, n (%)			
Male	17 (58.62)	15 (53.57)	16 (55.17)
Female	12 (41.38)	13 (46.43)	13 (44.83)
Pre-Test Score, Mean (SD)	5.97 (SD = 1.32)	5.89 (SD = 0.99)	6.21 (SD = 1.04)

Abbreviations: LBL, Lecture-Based Learning; GBL, Game-Based Learning; AI, Artificial Intelligence; AI-GBL, Generative AI-Powered Game-Based Learning; SD, Standard Deviation.

Table 2. Descriptive Statistics for Knowledge Score Changes by Teaching Group

Change	Group	Mean	SD	Unadjusted 95%CI
Change1	LBL	1.483	0.292	0.903 ~ 2.063
	GBL	2.607	0.297	2.017 ~ 3.198
	AI-GBL	2.379	0.292	1.799 ~ 2.959
Change2	LBL	−1.621	0.215	−2.048 ~ −1.194
	GBL	−1.000	0.219	−1.435 ~ −0.565
	AI−GBL	−0.310	0.215	−0.738 ~ 0.117

Note: The width of the 95% confidence intervals has not been adjusted for multiplicity, and they should not be used in place of hypothesis testing.
Abbreviations: LBL, Lecture-Based Learning; GBL, Game-Based Learning; AI, Artificial Intelligence; AI-GBL, Generative AI-Powered Game-Based Learning; SD, Standard Deviation; CI, Confidence Interval.

Table 3. Pairwise Comparisons of Knowledge Score Changes Between Groups

Change	Pairs	Δ Mean	SE	P Value	Unadjusted 95%CI
Change1	LBL~GBL	−1.124	0.416	0.008*	−1.952 ~ −0.297
	LBL~AI-GBL	−0.8966	0.412	0.033*	−1.717 ~ −0.076
	GBL~AI-GBL	0.228	0.416	0.586	−0.599 ~ 1.056
Change2	LBL~GBL	−0.621	0.306	–	−1.230 ~ −0.011
	LBL~AI-GBL	−1.310	0.303	–	−1.915 ~ −0.706
	GBL~AI-GBL	−0.689	0.306	–	−1.299 ~ −0.080

Note: The widths of the 95% Confidence Intervals for secondary outcomes have not been adjusted for multiplicity and should not be used in place of hypothesis testing.
Abbreviations: LBL, Lecture-Based Learning; GBL, Game-Based Learning; AI, Artificial Intelligence; AI-GBL, Generative AI-Powered Game-Based Learning; SE, Standard Error; CI, Confidence Interval.
*P < 0.05.

P < 0.001). However, given the large sample size (N = 86) and relatively equal group sizes (n = 29, 28, and 29, respectively), the MANOVA interpretation was deemed appropriate, as Pillai's Trace is robust to this violation.

The multivariate tests revealed a significant main effect of Group (Pillai's Trace = 0.314; F = 7.730; P < 0.001; partial $\eta^2 = 0.157$), indicating that the teaching methods had differential effects on student performance changes.

Table 4. Student Feedback Questionnaire Scores by Teaching Modality

Questionnaire	Groups	Δ Mean	SE	Unadjusted 95%CI
Learning Motivation	LBL vs GBL	−1.570	−0.081	−1.731 ~ −1.409
	LBL vs AI-GBL	−1.527	−0.080	−1.686 ~ −1.367
	GBL vs AI-GBL	0.043	0.081	−1.182 ~ 2.041
Cognitive Load	LBL vs GBL	1.477	0.092	1.568 ~ 1.932
	LBL vs AI-GBL	1.750	0.091	1.660 ~ 1.294
	GBL vs AI-GBL	0.273	0.092	0.899 ~ 0.456

Note: Scores were calculated as the mean of constituent items and range from 1 to 6. The width of the 95% confidence intervals has not been adjusted for multiplicity, and they should not be used in place of hypothesis testing.
Abbreviations: AI, Artificial Intelligence; AI-GBL, Generative AI-Powered Game-Based Learning; LBL, Lecture-Based Learning; GBL, Game-Based Learning; SE, Standard Error; CI, Confidence Interval.
*P < 0.05.

Univariate follow-up analyses showed significant group differences for both Change1 (F = 4.112; P = 0.020; partial $\eta^2 = 0.090$) and Change2 (F = 9.314; P < 0.001; partial $\eta^2 = 0.183$). The results of the post-hoc pairwise comparisons using the LSD method are shown in Table 3.

For Change1 (Improvement from Pre-Test to Post-Test)—The GBL group showed significantly greater improvement than the LBL group (mean difference = 1.124 points; unadjusted 95% CI [0.297, 1.952]; P = 0.008). The AI-GBL group showed significantly greater improvement than the LBL group (mean difference = 0.897 points; unadjusted 95% CI [0.076, 1.717]; P = 0.033). No significant difference was found between the GBL group and the AI-GBL group (mean difference = −0.228 points; unadjusted 95% CI [−0.599, 1.056]; P = 0.586).

For Change2 (Retention from Post-Test to Final-Test)—The AI-GBL group showed significantly better retention than both the LBL group (mean difference = 1.310 points; unadjusted 95% CI [0.706, 1.915]) and the GBL group (mean difference = 0.689 points; unadjusted 95% CI [0.080, 1.299]). The GBL group showed better retention than the LBL group (mean difference = 0.621 points; unadjusted 95% CI [0.011, 1.230]).

Analysis of Student Feedback Questionnaires

Reliability analyses indicated satisfactory to good internal consistency across the scales used in the current study. Specifically, the Cronbach's α was 0.826 for the learning motivation scale. For the cognitive load scale, the α was 0.880 for the mental effort subscale and 0.747 for the mental load subscale. For the technology acceptance questionnaire, the α was 0.851 for the usefulness subscale and 0.815 for the ease-of-use subscale.

Pairwise comparisons revealed that the GBL and AI-GBL groups scored significantly higher than the LBL group (mean difference = 1.570 points; unadjusted 95% CI [1.409, 1.731] and mean difference = 1.527 points; unadjusted 95% CI [1.367, 1.687], respectively), while no significant difference was observed between the GBL and AI-GBL groups (mean difference = 0.043 points; unadjusted 95% CI [−0.118, 0.204]); Table 4.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 14:08:09, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1049023X26108978

Pairwise comparisons indicated that the LBL group reported the highest cognitive load, followed by the GBL group, with the AI-GBL group reporting the lowest. Specifically, the AI-GBL group had significantly lower cognitive load than both the GBL group (mean difference = −0.273 points; unadjusted 95% CI [−0.456, −0.090]) and the LBL group (mean difference = −1.750 points; unadjusted 95% CI [−1.932, −1.568]). The GBL group also reported significantly lower cognitive load than the LBL group (mean difference = −1.477 points; unadjusted 95% CI [−1.660, −1.294]); Table 4.

The technology acceptance questionnaire, administered exclusively to the GBL and AI-GBL groups, revealed a significant difference in the overall perception between the two modalities. The AI-GBL group reported significantly higher scores than the GBL group (mean difference = 0.338 points; unadjusted 95% CI [0.010, 0.666]). A more detailed analysis of the subscales showed that this difference was driven by the "usefulness" dimension, where the AI-GBL group also reported significantly higher scores (mean difference = 0.513 points; unadjusted 95% CI [0.137, 0.889]). This indicates that students perceived the AI-powered approach as more beneficial to their learning. However, there was no difference between the groups on the "ease-of-use" subscale (mean difference = 0.187 points; unadjusted 95% CI [−0.149, 0.523]).

Discussion

The findings of this study highlight the distinct advantages of both GBL and AI-GBL over traditional LBL. As a critical component of disaster preparedness, EMR education encompasses essential competencies including incident management systems, triage protocols, and immediate medical interventions,[12] ultimately determining survival outcomes in mass-casualty incidents.[13]

For immediate knowledge acquisition (Change1), both the GBL and AI-GBL groups significantly out-performed the LBL group. This aligns with the initial hypothesis of this study and corroborates existing literature that suggests that the immersive nature of the tabletop exercise is critical for bridging theory and practice,[14,15] a known limitation of passive, didactic instruction in high-stakes environments like disaster response.[16,17]

However, the most significant contribution of this study is in delayed knowledge retention (Change2). The AI-GBL group demonstrated significantly better retention than both the GBL and LBL groups, indicating that the AI agent is crucial for solidifying knowledge. The authors propose the AI acted as a cognitive scaffold, providing on-demand support that reduced students' mental effort.[18] This "intelligent tutor" function freed cognitive resources, allowing students to focus on higher-order strategic thinking, which ultimately led to superior knowledge retention.[19]

Furthermore, both GBL and AI-GBL methods dramatically enhanced learning motivation compared to LBL,[20,21] an expected outcome given the inherent engagement of active, role-based participation. The lack of a motivational difference between the GBL and AI-GBL groups, despite the latter's lower cognitive load, suggests the core game-based experience was the primary driver of student interest. The AI's contribution, therefore, appears less about increasing enjoyment and more about enhancing educational efficiency – a finding with important practical implications for instructional design. This efficiency interpretation is corroborated by the technology acceptance results: while both systems were deemed equally easy to use, students in the AI-GBL group perceived their system as significantly more useful. This enhanced perception of usefulness likely stems from the AI agent's role as an on-demand advisor, which provided immediate, context-aware guidance to overcome knowledge barriers.[22] This finding aligns perfectly with the lower cognitive load in the AI-GBL group, indicating that efficient access to support not only reduces mental effort, but also elevates the perceived educational value of the tool.

An important consideration when integrating generative AI into medical education is the risk of AI-generated inaccuracies or "hallucinations" – confident but incorrect responses that could potentially reinforce misconceptions or, worse, promote dangerous clinical practices. In this study, several safeguards were implemented to mitigate this risk, including strict prompt engineering to confine the AI's knowledge base to evidence-based protocols and pre-deployment expert review of AI responses. Nevertheless, the authors acknowledge that no AI system is infallible. It is recommended that future implementations incorporate a human-in-the-loop mechanism, where an instructor periodically reviews AI-student interactions, and that students be explicitly trained to critically evaluate AI-generated advice rather than accepting it uncritically.[23] This aligns with recent calls to frame AI literacy as a core competency in medical education.[24–26]

Ultimately, it is important to acknowledge the fundamental difference in instructional delivery between the LBL and the game-based cohorts. The LBL group received theoretical instruction, whereas the GBL and AI-GBL groups engaged in experiential, practice-driven exercises. The superior performance of the game-based groups likely reflects the intrinsic advantage of experiential learning in bridging the gap between theoretical knowledge and practical application – a gap that traditional didactic methods inherently struggle to close in high-stakes environments like EMR.

Limitations

Despite these promising results, the study has several limitations that must be acknowledged. First, the retrospective design, while allowing for a naturalistic comparison, lacks the randomization of a prospective randomized controlled trial, potentially introducing uncontrolled confounding variables. Second, the use of a convenience sample of students who self-selected to take the Disaster Medicine course may introduce selection bias and limit the generalizability of the findings to the broader medical student population. Third, regarding the measurement instruments, although the questionnaires were adapted from validated instruments and reviewed by an expert panel, a formal pilot study was not conducted prior to their implementation, which represents a methodological limitation.

Fourth, the pre-test and post-test design carries inherent limitations, including the potential for practice effects due to repeated exposure to the assessment items, and the possibility that students acquired relevant knowledge from external sources during the study period. Furthermore, while the three-week follow-up assessed delayed knowledge retention, this interval is insufficient to evaluate true long-term knowledge retention. Fifth, the knowledge assessment focused primarily on factual recall via multiple-choice questions, which may not fully capture the development of complex problem-solving or collaborative skills central to EMR.

Sixth, the study did not adjust for multiplicity in the reporting of 95% confidence intervals for secondary outcomes. Therefore, these intervals should be interpreted with caution and used primarily for hypothesis generation rather than formal hypothesis testing.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 14:08:09, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1049023X26108978

Seventh, the study was conducted at a single institution with a specific cohort of medical students, which may limit the generalizability of the findings to other contexts or learner populations. Another limitation concerns the violation of the MANOVA assumption of homogeneity of covariance matrices. While Pillai's Trace is considered a robust statistic under this condition, particularly with relatively equal group sizes, the violation of this assumption is a methodological limitation that could potentially affect the validity of the statistical inferences. Finally, the AI agent, while effective, was based on a specific model (ChatGLM) and prompt-engineered for this scenario; its performance and applicability in other educational domains remain to be tested.

Research Implications

Given the inherent complexity of EMR education, which requires the seamless integration of theoretical knowledge, practical skills, and rapid decision making under pressure,[1,27] despite these limitations, this study provides a crucial foundation for future research into AI-enhanced medical education. Future studies should aim to validate these findings in prospective, randomized controlled trials to establish stronger causal links and incorporate larger, more diverse samples to improve generalizability. It will also be important to develop more comprehensive assessments that capture complex skills like problem solving and to conduct longitudinal studies that evaluate the delayed retention of knowledge. Furthermore, research should explore the effectiveness of different AI models and prompt-engineering strategies across various medical disciplines, while also carefully considering the ethical implications of wide-spread AI integration in training.

Conclusion

This study suggests that integrating a generative AI agent into GBL has the potential to create a more effective educational environment for EMR training. While the findings show that game-based approaches are superior to traditional lectures, the key contribution of this work is the demonstration that the AI-enhanced GBL model reduces students' cognitive load compared to standard GBL. By alleviating cognitive demand, this approach helps students focus on core learning tasks, potentially mitigating practice effects and significantly improving both knowledge acquisition and retention – outcomes that are critical for developing proficiency in EMR systems.

Supplementary Materials. To view supplementary material for this article, please visit https://doi.org/10.1017/S1049023X26108978

Acknowledgements. The authors would like to thank Xi-Yun Technology Co., Ltd. (Shenzhen, China) for their technical support in the development of the educational game.

Author Contributions. Yaning Lai: Writing; Review and Editing; Methodology; Data Curation; Formal Analysis; Investigation; Resources; Project Administration.
Xiaoqin Lai: Writing; Review and Editing; Methodology; Validation; Formal Analysis; Data Curation.
Hai Hu: Conceptualization; Software; Data Curation; Investigation; Validation; Formal Analysis; Supervision; Resources; Writing Original Draft; Methodology; Funding Acquisition.

Funding Statement. This work was supported by the 1·3·5 Project of Center for High Altitude Medicine, West China Hospital, Sichuan University, Chengdu, China [Grant No. GYYX24010] and the 2025 Medical Education Research Project from the Chinese Society of Medical Education of the Chinese Medical Association and National Center for Health Professions Education Development, Peking University, Beijing, China [Grant No. 2025B277].

Competing Interests. The authors declare that they have no conflicts of interest statement.

Use of AI Technology. To address the need for interactive training in Disaster Medicine, the authors designed a game-based learning environment for medical students, complemented by an external AI-powered support agent. This agent, developed independently using the ChatGLM large language model, functions as a specialized knowledge base and an interactive tutor. While students engage in the role-playing simulation, they can access the agent on-demand to receive context-relevant information – such as triage principles, emergency procedures, and pharmacological guidance – in response to their specific queries. This integration ensures that learners receive immediate, evidence-based support to guide their clinical reasoning and decision-making processes, thereby enhancing the educational experience without disrupting the game's flow.

Reference

1. Hung KKC, Mashino S, Chan EYY, et al. Health Workforce Development in Health Emergency and Disaster Risk Management: The Need for Evidence-Based Recommendations. *Int J Environ Res Public Health.* 2021;18(7):3382. doi: 10.3390/ijerph18073382.
2. Usoro A, Mehmood A, Rapaport S, et al. A Scoping Review of the Essential Components of Emergency Medical Response Systems for Mass Casualty Incidents. *Disaster Med Public Health Prep.* 2023;17:e274. doi: 10.1017/dmp.2022.235.
3. Hu H, Lai Y, Zuo C. Role-playing computer games in disaster medicine education. *Med Educ.* 2024;58(11):1408–1409. doi: 10.1111/medu.15515.
4. Sugaya K, Kakamu T, Iseki K, Hasegawa A. Differences in perceptions and intentions to engage in various disasters by curriculum and social roles of belonging: a cross-sectional study of Japanese medical students. *Fukushima J Med Sci.* 2025;71(3):177–185. doi: 10.5387/fms.24-00049.
5. Hu H, Xiao Y, Li H. The Effectiveness of a Serious Game Versus Online Lectures for Improving Medical Students' Coronavirus Disease 2019 Knowledge. *Games Health J.* 2021;10(2):139–144. doi: 10.1089/g4h.2020.0140.
6. Chow M, Ng O. Beyond Chatbots: Moving Toward Multistep Modular AI Agents in Medical Education. *JMIR Med Educ.* 2025;11:e76661. doi: 10.2196/76661.
7. Peng Y, Hu H. Assessment of earthquake casualties and comparison of accuracy of five injury triage methods: evidence from a retrospective study. *BMJ Open.* 2021;11(10):e051802. doi: 10.1136/bmjopen-2021-051802.
8. Hwang GJ, Chang HF. A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. *Computers & Education.* 2011;56(4):1023–1031. doi: 10.1016/j.compedu.2010.12.002.
9. Gwo-Jen Hwang, Li-Hsueh Yang, Sheng-Yuan Wang. A concept map-embedded educational computer game for improving students' learning performance in natural science courses. *Computers & Education.* 2013;69:121–130. doi: 10.1016/j.compedu.2013.07.008.
10. Chu HC, Hwang GJ, Tsai CC, et al. A two-tier test approach to developing location-aware mobile learning system for natural science course. *Computers & Education.* 2010;55(4):1618–1627. doi: 10.1016/j.compedu.2010.07.004.
11. Sweller J, van Merrienboer JJG, Paas FGWC. Cognitive Architecture and Instructional Design. *Educational Psychology Review.* 1998;10:251–296. doi:10.1023/A:1022193728205
12. Hu H, Lai X, Tan C, Yao N, Yan L. Factors Associated with In-Patient Mortality in the Rapid Assessment of Adult Earthquake Trauma Patients. *Prehosp Disaster Med.* 2022;37(3):299–305. doi: 10.1017/S1049023X22000693.
13. Suda AJ, Franke A, Hertwig M, Gooßen K. Management of mass casualty incidents: a systematic review and clinical practice guideline update. *Eur J Trauma Emerg Surg.* 2025;51(1):5. doi: 10.1007/s00068-024-02727-0.
14. Sena A, Forde F, Yu C, Sule H, Masters MM. Disaster Preparedness Training for Emergency Medicine Residents Using a Tabletop Exercise. *MedEdPORTAL.* 2021;17:11119. doi: 10.15766/mep_2374-8265.11119.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 14:08:09, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1049023X26108978

15. Hu H, Liu Z, Li H. Teaching Disaster Medicine with a Novel Game-Based Computer Application: A Case Study at Sichuan University. *Disaster Med Public Health Prep.* 2022;16(2):548–554. doi: 10.1017/dmp.2020.309.
16. Pandya J, Kharel R, McMahon J, et al. A tabletop exercise approach to global disaster preparedness: insights from Nepal's first international conference on disaster preparedness and management. *Front Public Health.* 2024;12:1400588. doi: 10.3389/fpubh.2024.1400588.
17. Chou WK, Cheng MT, Lin CH. Teaching mass casualty incident management to senior medical students by three-dimensional tabletop exercise without lecture. *BMC Med Educ.* 2025;25(1):846. doi: 10.1186/s12909-025-07434-x.
18. Velupillai J, Waite S. Click, create, critique: Futureproofing critical thinking in the age of generative AI. *Med Teach.* 2026;48(4):735. doi: 10.1080/0142159X.2025.2566267.
19. Zhang Y, Tian H, Lu J. The relationship between higher-order thinking and problem-solving skills development among pre-service teachers using generative AI: an analysis based on moderated mediation. *BMC Psychol.* 2025;13(1):1094. doi: 10.1186/s40359-025-03404-6.
20. Hu H, Lai X, Yan L. Training nurses in an international emergency medical team using a serious role-playing game: a retrospective comparative analysis. *BMC Med Educ.* 2024;24(1):432. doi: 10.1186/s12909-024-05442-x.
21. Hu H, Lai X, Yan L. Improving Nursing Students' COVID-19 Knowledge Using a Serious Game. *Comput Inform Nurs.* 2021;40(4):285–289. doi: 10.1097/CIN.0000000000000857.
22. Alsahafi Z, Baashar A. Knowledge, Attitudes, and Practices Related to Artificial Intelligence Among Medical Students and Academics in Saudi Arabia: A Systematic Review. *Cureus.* 2025;17(5):e83437. doi: 10.7759/cureus.83437.
23. Chow M, Ng O. Beyond Chatbots: Moving Toward Multistep Modular AI Agents in Medical Education. *JMIR Med Educ.* 2025;11:e76661. doi: 10.2196/76661.
24. Ng O, Chow M. Empowering Health Professions Educators: Developing Educational Tools with AI-Assisted Vibe Coding. *Med Sci Educ.* 2025;36(1):47–52. doi: 10.1007/s40670-025-02596-1.
25. Hardie P, Darley A, Derwin R, et al. Applications, attitudes and ethical considerations of Generative Artificial Intelligence (Gen AI) in nursing education: a scoping review. *BMC Nurs.* 2026;25(1):148. doi: 10.1186/s12912-025-04253-9.
26. Holzinger A, Zatloukal K, Müller H. Is human oversight to AI systems still possible? *N Biotechnol.* 2025;85:59–62. doi: 10.1016/j.nbt.2024.12.003.
27. Hai H, Ya-Rong H, Xin-Miao D, et al. Chief complaints associated with mortality involving civilian transport after Wen-Chuan earthquake. *Eur J Emerg Med.* 2014;21(5):364–367. doi: 10.1097/MEJ.0000000000000104.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 14:08:09, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1049023X26108978