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.

Assessing the Learning Gains and User Experience of Large Language Model Simulations for Clinical Medical Education: A Pilot Observational Study

Riley Scherr

rscherr@bidmc.harvard.edu

Beth Israel Deaconess Medical Center

Faris F. Halaseh

University of California Irvine Medical Center

Aidin Spina

Scripps Mercy Hospital

Saman Andalib

University of California Irvine Medical Center

Soheil Saadat

University of California Irvine

Alisa Wray

University of California Irvine

Ronald Rivera

University of California Irvine

Research Article

Keywords: simulation, LLM, user experience, asthma

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10529565/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

Medical students often struggle transitioning from preclinical coursework to clinical environments. Simulation-based training can bridge this gap but can be resource intensive. Large language models (LLMs) provide a scalable resource that can augment traditional training. However, their effects on learning outcomes and student perceptions are currently not understood. This study aims to evaluate medical student knowledge acquisition and perception of the utility of simplified LLM-based clinical simulation.

Methods

Using the RISE framework, a structured simulation was designed to assess medical student knowledge of acute asthma exacerbation management. ZotGPT, an institutional GPT-4 based platform was utilized. Medical students completed pre- and post-simulation assessments assessing knowledge, confidence, and perceptions of simulation utility. Subgroup analysis was stratified by time spent completing the simulation. Statistical analyses included Wilcoxon signal-rank and McNemar tests.

Results

A total of 155 students participated. The simulation significantly improved perceived understanding of key treatment steps, including initial bronchodilator and steroid therapy ($p = 0.001$), magnesium sulfate escalation ($p = 0.002$), refractory management ($p = 0.021$), and discharge planning ($p = 0.005$). Students rated the simulation as more useful than non-simulation learning tools ($p = 0.017$) and noninferior to in-person simulation ($p = 0.226$). Confidence in asthma management increased ($p < 0.001$), but objective knowledge gains were not observed, with balanced correct-to-incorrect and incorrect-to-correct responses for intermediate steps. Longer engagement (> 600 seconds) was associated with greater perceived benefit. Simulations were deemed helpful for learning (94.7%), and most students preferred immediate feedback (68.4%).

Conclusions

LLM-based clinical simulations were well accepted, improved medical student confidence, and were perceived as educationally useful. However, objective knowledge gains were limited. LLMs thus may serve as an adjunctive educational tool but require improved prompt design, structured feedback, and guided implementation to optimize outcomes.

Introduction

Medical students often face challenges during their transition from preclinical education to clinical clerkships(1). This milestone requires students to transition from a mostly classroom-based, conceptual curriculum to application-based, hands-on, patient care in clinical environments. The demands of this transition can be challenging, as students must rapidly adapt to new roles and expectations, navigate unfamiliar healthcare settings, and integrate theoretical knowledge with practical application.

Medical simulation is an increasingly important tool in preclinical medical education, offering standardized experiences in controlled environments without patient risk (2). Various simulation modalities, including partial task trainers, standardized patients, and high-fidelity simulators, are used to enhance learning across different competencies (2). Studies have shown that simulation-based learning can significantly improve student performance, with first-year students demonstrating greater benefits compared to second-years (3). This difference between early and later pre-clinical learners may be explained by the expertise reversal effect and highlights a need for adaptive simulation technology(4). The overall effectiveness of simulation is supported by emerging evidence, and it has already been widely adopted into medical education, though it requires proper integration into the curriculum with appropriate educator oversight to effectively promote skill transfer to clinical practice (5). Simulation-based education has expanded beyond procedural training to include basic science content and early clinical experiences (2). While simulation initiatives in some countries remain fragmented due to barriers like cost, personnel, maintenance, and technical skill, there are efforts to develop more coordinated national approaches to optimize its benefits in medical education (5).

More recently, artificial intelligence (AI) has emerged as a potentially transformative model in medical education, offering innovative solutions to such problems. Among these technologies, large language models (LLMs), such as GPT-based systems, offer personalized and interactive learning experiences that go beyond traditional educational methods (6, 7). Unlike static learning resources like textbooks or question banks, AI-driven tools provide learners with dynamic engagement opportunities that more closely mirror the complexities of real-world patient care. For example, the *CureFun* framework has demonstrated how LLMs can simulate clinical training scenarios with authentic dialogue flows, significantly enhancing the realism of virtual patient interactions compared to less engaging chatbot models (8). These simulations challenge students to think critically, adapt to evolving patient conditions, and refine their diagnostic and therapeutic reasoning. These findings are corroborated by other studies that have shown ChatGPT-based simulations can reliably adapt to user inputs in a way that replicates real-life patient encounters while maintaining high medical accuracy (9, 10). ChatGPT's capacity for medical accuracy has also been demonstrated through passing scores on various standardized national exams in the United States and abroad ((11, 12, 13, 14, 15).

Despite the promising applications of AI and LLMs in medical training, there is a need to assess its real-world applicability, effects on knowledge acquisition, and user experience. Medical student experiences with AI simulations have been relatively unexplored. We have previously demonstrated the feasibility of creating interactive LLM-based simulations for preclinical medical students(9). Then, we developed a nine-point evaluation rubric to assess LLM-based simulations using the Society for Simulation in

Healthcare's guidelines for simulation training and Richard Mayer's multimedia design and education principles(10). When this evaluation rubric was applied to a ChatGPT based simulation, we found that this LLM created interactive, realistic, accurate, and replicable sessions for preclinical learners. Based on that success, this study aimed to evaluate medical student learning acquisition, user experience, and perceptions of simulations' educational utility after using an LLM simulation.

Methods

A decision-based simulation prompt on treating an acute inpatient asthma exacerbation was engineered using the RISE framework(9, 16). This simulation scenario was selected because it was validated in prior work. Medically, an asthma exacerbation has a relatively simplistic presentation for early learners that has a well-established stepwise treatment algorithm that we have simplified to Fig. 1 (17, 18).

The simulation was instructed to dynamically change patient conditions based on user interventions, allowing participants to engage in real-time decision-making. Key features, as determined by our prior AI-simulation metric, included providing free-text input instead of multiple-choice options and focusing on treatment strategies without requiring drug dosages or conditional logic. The prompt also requested overall feedback on the clinical appropriateness of the user's choices be provided by the LLM after completion of the simulation. The simulation prompt is included below:

Please create a clinical scenario on a patient presenting to the hospital with an acute asthma exacerbation and quiz me on what the proper next step of management is. Please make it free response and interactive, meaning you ask me what the next step is one question at a time, and then I write out what I would do, and then you ask me another question based on how my answer would affect the patient. Please update / change the patient's condition based on my actions, and do not tell me the right answers until the end of the scenario.

Simulations were run on ZotGPT, a free-to-student LLM built on GPT-4 Turbo but configured at the principal investigators' institution with additional data-sharing safety features. Third-year medical students were invited to participate in the study through email invitations and classroom announcements. Participants began the study by completing a pre-simulation survey hosted on the Qualtrics platform. The survey included baseline questions assessing their confidence and knowledge of acute asthma exacerbation management. Participants then accessed ZotGPT through a secure institutional web interface and followed structured prompts to complete the simulation. After completing the simulation, participants then answered the same questions assessing confidence and knowledge of acute asthma exacerbation management to evaluate knowledge acquisition, user experience, and the perceived educational value of the simulation. Both surveys also assessed participants' perceived utility of ChatGPT in comparison to traditional educational methods, such as textbooks and in-person simulation labs. Feedback was collected regarding preferences for immediate versus delayed simulation feedback and the overall feasibility of incorporating ChatGPT into medical education.

For subgroup analysis, survey completion times were stratified into three categories based on the authors' experience completing the simulations in a timely fashion: ≤180 seconds, 181–600 seconds, and > 600 seconds. This stratification aimed to identify potential patterns in survey engagement and learning outcomes. The user responses were evaluated by author S.S. using SPSS statistical software version 28 (IBM Corporation, Armonk, NY, USA). Statistical analysis was conducted using a one-sample Wilcoxon signed rank test, with significance determined at an alpha level of 0.05. This study was carried out in accordance with the Institutional Review Board (Protocol #4986) at the University of California, Irvine.

Results

A total of 155 students participated in the study, 54 of which did so in structured class time and 101 of which participated via an email invitation. Response size varied per question, as each question was optional, and ranged from 20 to 102. All data collected and reviewed for this manuscript are available in the related files section as a Microsoft Excel file for review.

Student Perception and Time Sub-Groups

Students rated the simulation as significantly helpful in clarifying the proper first step in acute asthma management i.e. short-acting beta-agonist and glucocorticoid administration ($p = 0.001$), identifying appropriate second-line treatments for unresolved respiratory distress i.e. intravenous magnesium sulfate ($p = 0.002$), understanding treatment for severe exacerbations refractory to other interventions i.e. endotracheal intubation ($p = 0.021$), and identifying the best next step in recovered severe exacerbations i.e. inhaler refill and use education with oral steroids and outpatient follow-up ($p = 0.005$). The simulation was found to be more useful than non-simulation learning methods (e.g., textbooks, lectures, flashcards) ($p = 0.017$) and as useful as in-person simulation labs (no statistical difference, $p = 0.226$). The simulation also significantly improved student confidence in their ability to manage acute asthma exacerbations compared to before the intervention ($p < 0.001$).

Subgroup analysis examined student perception of the simulations' helpfulness in clarifying correct treatment steps and student perception of the simulations' utility compared to other learning modalities. Statistical significance was maintained (p-values: 0.004–0.037) for users who spent more than 600 seconds. However, this effect went away for participants who spent < 600 seconds on their surveys (Fig. 2).

Among students who received feedback during the simulation, 94.7% found it helpful, indicating that real-time reinforcement may have contributed to these learning gains. 68.4% of students also expressed a preference for immediate feedback rather than delayed feedback. 51.9% of students were "somewhat familiar" with AI and LLMs before the simulation. 82.4% said the simulations were somewhat or extremely easy to use. 70.6% said it was pleasant to use. 56% said they were somewhat or extremely likely to use these simulations to prepare for standardized patient encounters or OSCEs. 51% were

somewhat or extremely confident in the accuracy of ChatGPT's information, while 28.6% were either somewhat or extremely unconfident in the accuracy. Please see Table 1.

Table 1
Student perceptions of ChatGPT simulation user experience, usefulness, confidence, and feedback preferences. Values represent percentages.

User Experience	**Extremely/Somewhat Positive**		**Extremely/Somewhat Negative**
How familiar are you with artificial intelligence (AI) / large-language models (LLMs)?	32.7		57.7
How easy was the ChatGPT simulation to use?	82.4		5.9
Relative to other learning tools, how pleasant or unpleasant was the ChatGPT simulation experience?	70.6		2.0
How likely would you be to use ChatGPT simulations to prepare for standardized patient encounters, OSCEs, clinical encounters, etc.?	56.0		16.0
Usefulness	**Positive/More Usefulness**	**Neutral**	**Negative/Less Usefulness**
How useful was the ChatGPT simulation for learning?	66.0	0	34.0
How useful was the ChatGPT simulation for learning compared to in-person simulation (e.g. Sim Labs)?	22.0	40.0	38.0
How useful was the ChatGPT simulation for learning compared to non-simulation learning (e.g. textbooks, flashcards, lectures, videos)?	43.0	43.0	14.0
Confidence	**Extremely/Somewhat Confident**		**Extremely/Somewhat Unconfident**
How confident were you in the accuracy of ChatGPT's feedback in your simulation?	28.6		51.0
Feedback	**Yes**		**No**
If ChatGPT gave you feedback before the simulation ended, was this helpful for your learning?	94.7		5.3
Would you prefer immediate feedback after each step of the simulation, or prefer delayed feedback at the end of the simulation?	68.4		31.6

Potential Uses

43.8% of students said they would use the simulation to identify weak topics and then use other resources like lectures and online libraries to learn these topics, 35.4% said they would only use ChatGPT simulations for practice but not to guide learning, 10.4% said they would not use ChatGPT simulations at all, and 10.4% would use ChatGPT simulations as a primary learning resource. When responses were separated based on the setting of simulation completion (in-person administration versus email recruitment) there were no significant differences in their responses to these questions.

Qualitative Comments

Samples of student comments include the following:

I enjoyed how immediate the responses were via the AI, it made the simulation flow well.

The level of feedback I received at the end of the simulation was better than I thought it would be!

Discussion of contraindicated medications never came up, also am [sic] slightly biased because I am at the end of MS3 so this was less helpful than it would be, presumably, for an MS1.

Reliability is a prevalent concern if I do not know the information myself.

Knowledge Acquisition

Pre- and post-simulation knowledge was also assessed at each step in the asthma exacerbation algorithm. Respondents were categorized for each question as having answered correctly both pre- and post-simulation, changed from correct to incorrect, changed from incorrect to correct, or stayed incorrect both times (Table 2). Despite the increase in student confidence presented in Table 1, the data did not show corresponding significant knowledge acquisition. For the intermediate treatment steps where students had the greatest knowledge gaps (magnesium sulfate and intubation), equal numbers of students improved as declined (7% vs 7% for magnesium, 5% vs 5% for intubation), representing a net learning effect of zero. Nearly half of students answered these questions incorrectly pre-simulation and remained incorrect post-simulation.

Table 2

Student pre/post-intervention knowledge assessment. Each column indicates the sequential correctness of student answers. For example, a student who initially answered Step 1 incorrectly before the simulation but answered Step 1 correctly after the simulation is represented under "Incorrect/Correct." P values are based on McNemar test.

Algorithm Step	Correct / Correct	Correct / Incorrect	Incorrect / Correct	Incorrect / Incorrect	Sample (n)	p-value
1: Nebulizer + Steroids	40 (82%)	2 (4%)	3 (6%)	4 (8%)	49	.65
2: Magnesium Sulfate	12 (40%)	2 (7%)	2 (7%)	14 (47%)	30	> .99
3: Endotracheal Intubation	22 (51%)	2 (5%)	2 (5%)	17 (40%)	43	> .99
4: Patient Education	38 (72%)	3 (6%)	7 (13%)	5 (9%)	53	.21
	154	12	18	46		.27

Discussion

Conventional educational tools like textbooks and lectures, while effective in building foundational knowledge, may fall short in replicating the dynamic nature of clinical environments. Simulation labs and standardized patient interactions are designed to address some of these gaps but are resource-intensive and not always accessible on a large scale. AI-driven simulations, particularly those utilizing large language models (LLMs), can offer a promising augmentation to existing training structures as shown in prior work. Integrating AI-based simulation sessions into preclinical medical education could accordingly assist with knowledge retention and confidence-building through an evidence-based and well-received interface. These interventions are particularly important for improving medical training in low resourced settings where access to high fidelity simulation is limited or not feasible.

This well-received interface and the subsequent student use experience is an important part of any learning tool's engagement and adoptability. This LLM-based simulation session was highly ranked for ease of use, perceived educational value, and showed moderate levels of engagement (as measured by student willingness to use the tool again or in favor of other learning modalities). Interestingly, the student's answers regarding their confidence in the accuracy of feedback mirrored their levels of experience using LLM's. This could imply that greater familiarity and experience with these models would improve confidence. These responses exemplify principles of the Technology Acceptance Model, which describes new technologies' likelihood of acceptance based on perceived usefulness and ease of use (citation(19)). Notably, GPT simulations were perceived as noninferior to in-person simulation, which may be due to shared key characteristics of multimedia learning technology that are more impactful to learners than their differences. This is supported by numerous recent studies showing virtual simulations are generally noninferior to in-person simulations despite lacking the physicality of things

like IV placement or mannequins (20, 21). Both GPT and in-person simulations require deeper engagement, rapid cognitive processing, and answer generation; both are thus consistent with the generative learning principle of self-explanation, in that the increased cognitive demand of creating answers by integrating previously known information will be more effective for learning than selecting an answer from a list of multiple choice options. This has long been known to be effective in medical education, although some data suggests well-constructed multiple choice questions are noninferior to open-ended assessment and correlate with clinical skill (22, 23, 24). Live simulation sessions often have the added benefit of debriefs to call attention to individual and team performance, reflect on what went well and areas for improvement, and to review the important takeaways for the session. The studied prompt generated feedback at the end of the sessions, but it could be adapted to include these elements as a form of self-reflection, which may deepen learning. This also calls to attention the lack of team dynamics and interprofessional knowledge sharing that can be gained from in-person simulations, especially if combined with students from schools of nursing and pharmacy. Future studies could look at differences in learning based on these features.

The discordance observed between students' high confidence and poor knowledge acquisition (Tables 1 and 2), may be representative of the Dunning-Kruger effect. This is a well-documented phenomenon within medical education, especially in studies on preclinical medical students (25, 26). High fidelity simulation has, in particular, been linked to this type of confidence-competence gap (27). Advanced clinical medical students using ChatGPT as a standardized patient (rather than as a simulation scenario) have also demonstrated a confidence-competence gap (28). Literature on measurable learning outcomes in LLM simulations is scarce overall though; less than a quarter of studies on LLM-based virtual patients use objective learning measurements, which highlights a serious gap in our understanding of these tools (29). Knowledge acquisition in these simulations also relied solely on ChatGPT feedback. One study showed ChatGPT feedback to be inferior to expert human feedback for improving clinical reasoning skills in complicated urinary tract infection management (a difference not seen in less complex cases) (30). This raises the possibility that ChatGPT is not capable of the expert teaching required for knowledge acquisition, but future studies should directly compare knowledge acquisition between ChatGPT and in-person simulations. Clearer direction for ChatGPT on desired feedback and learning goals should also be explored. For example, an instructor could upload a pre-made simulation session with specific learning points alongside the prompt. A simple line added to the prompt could ensure these learning points are reviewed at the end of the simulation. Lastly, widely available LLMs (Claude, Gemini, ChatGPT) perform differently on scenario generation, which could also present different learning experience, feedback, and subsequent learning acquisition, further studies should be done to evaluate their engagement, efficacy, and utility using a rubric like the one created in our prior study (31).

Regarding simulation utility, subgroup analysis by time spent showed increased time spent on the simulation increased its perceived utility in clarifying the learning points and overall utility compared to non-simulation learning methods. Though we did not observe each participant during the simulations, time may be a proxy for thoughtful engagement and effort with the simulation. This is both a potential

limitation of the data–if all participants had a standard, enforced timeframe to complete the simulation, their perceived utility may have changed–and an insight; simulations may need to be used in structured learning environments with clear expectations, appropriate time, and sustained focus to be a positive learning experience and provide maximal educational benefit. Though data are limited, this is supported by one study showing improved critical thinking skills in a tutored-form of ChatGPT case-based learning compared to an independent-form of ChatGPT (32). This points to facilitation being key. Simulation sessions are an opportunity for teacher facilitation to establish learning objectives, guide students towards these objectives, and provide guardrails when learners get stuck. Instructors can also monitor engagement by asking students to submit their simulation narratives at the end of the activity, allowing the instructor to review the depth of student responses for appropriate engagement. Future work could explore the impact of timed and/or facilitated AI-based simulation sessions on user experience and learning compared to unfacilitated AI-based simulations. With large repositories of simulation training sessions available from free, open-access, medical education (FOAMEd) sites like JetEM and MedEd Portal, future studies could evaluate an LLM's ability to run these pre-made simulation sessions. This could reduce the burden on medical educators to come up with new materials, instead focusing on how to best improve learning acquisition. We also see the applicability of this technology for broadening the scope of learning beyond assessment and treatment of patients. Simple changes to the prompt can have the LLM include modifiers to the case surrounding health equity, social determinants of health, and whole person care through the bio-pscyho-social-spiritual model of learning. The potential for adaptation and modification to learning environments is exciting.

Limitations

While we acknowledge the limitations inherent in single-scenario studies, our methodical approach to data collection and analysis seeks to mitigate these concerns. Still, this study is limited by its small sample size. Response number per question varied widely between questions because no survey question was mandatory. This was a single site study with only one scenario tested. Limited responses to demographic questions prevent subgroup hypothesis generation or assessments of learning based on student demographics. Our initial intent was to see if students who self-identify as belonging to groups underrepresented in medicine have different experiences with this technology, or if other demographic factors such as gender, age, and neurodivergence modify the outcomes measured. Future studies should include these factors to determine if this technology can bridge equitable education gaps, or if it contributes to them. Our prompt did not specify which learning points should be covered, so feedback (the key educational mechanism here) was dependent on learner inputs. There was no control group of in-person simulation. Subgroup analysis by time spent is inherently hypothesis-generating rather than evidence. Knowledge assessment was not repeated to assess long-term retention. Future studies should examine knowledge acquisition from LLM prompts with specified learning objectives. More broadly, integrating LLMs into medical education still presents challenges, including concerns about content accuracy, potential overreliance on AI, and risks of academic misconduct (7, 33). Future iterations may benefit from structured post-simulation debriefs or automated feedback summaries that consolidate key

teaching points and address common misconceptions. More investigation is needed to assess learning gains more robustly, adaptability across diverse clinical specialties, and reliability when attempting complex vignettes before widespread implementation.

Conclusions

Medical students rated GPT-based clinical simulations as significantly helpful for clarifying treatment steps for acute asthma exacerbation, and easy to use. GPT simulations were also rated more useful than non-simulation learning materials and noninferior to in-person simulation. This was driven by participants who spent more time on the simulation, suggesting that simulation utility may depend on student engagement and instructor facilitation. Knowledge acquisition was poor and students showed a competence-confidence mismatch, but this may depend on better prompt engineering. GPT simulations may be a viable learning tool to diversify educational methods, augment existing simulation curricula, or promote practical training where in-person simulation is unavailable or not feasible.

Abbreviations

AI
artificial intelligence
LLM
large language model
OSCE
observed standardized clinical encounter

Declarations

Ethics Statement:

This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and its subsequent amendments. The study protocol was reviewed and approved by the Institutional Review Board (IRB) of The University of California, Irvine (Approval Number: 4986). All participants provided written informed consent prior to enrollment. All methods were performed in accordance with relevant institutional and national guidelines and regulations.

Consent for publication:

Not applicable

Competing Interests:

All authors declare no financial or non-financial competing interests.

Funding:

This study received no funding

Author Contribution

All authors contributed equally: RS created the initial research plan, developed the online survey, delivered the classroom simulation session to students, wrote parts of the manuscript, edited the manuscript, and made tables 1 & 2, and figure 1. FH developed the initial research plan, wrote parts of the manuscript, edited, and refined the programmatic prompt. AS developed the original research plan, wrote parts of the results section, edited, and created figure 2. SA developed the research plan, wrote the abstract, and edited. SS evaluated the research plan for statistical feasibility, analyzed the data, wrote the results section, and edited. AW delivered the classroom simulation session to students, wrote parts of the discussion, added educational theory portions of the manuscript, and edited. RR developed the initial research plan, created the online survey, wrote parts of the methods, results, and discussion, and added educational theory portions of the manuscript.

Acknowledgements:

Not applicable.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

References

1. Malau-Aduli BS, Roche P, Adu M, Jones K, Alele F, Drovandi A. Perceptions and processes influencing the transition of medical students from pre-clinical to clinical training. BMC Med Educ. 2020;20:292.
2. Lipps JA, et al. The expanding use of simulation for undergraduate preclinical medical education. Int J Acad Med. 2017;3:59–65.
3. Jabaay MJ, et al. Medical simulation-based learning outcomes in pre-clinical medical education. Cureus. 2020;12:e11875.
4. Kalyuga S, Ayres P, Chandler P, Sweller J. The expertise reversal effect. Educ Psychol. 2003;38:23–31.
5. Weller JM. Simulation in undergraduate medical education: bridging the gap between theory and practice. Med J Aust. 2012;196:597–8.

6. Abd-Alrazaq A, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. 2023;9:e48291.
7. Benítez TM, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. J Am Med Inf Assoc. 2024;31:776–83.
8. Li Y, arXiv et al. https://arxiv.org/abs/2404.13066
9. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. JMIR Med Educ. 2023; 9, e49877.
10. Scherr R et al. Novel evaluation metric and quantified performance of ChatGPT-4 patient management simulations for early clinical education: experimental study. JMIR Form Res. 2025; 9, e66478.
11. Gilson A et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023; 9, e45312.
12. Nakao T, et al. Capability of GPT-4V(ision) in the Japanese National Medical Licensing Examination: evaluation study. JMIR Med Educ. 2024;10:e54393.
13. Kung TH, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198.
14. Ali R, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery. 2023;93:1353–65.
15. Brin D, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep. 2023;13:16492.
16. Heston TF, Khun C. Prompt engineering in medical education. Int Med Educ. 2023;2:198–205.
17. Reddel HK, et al. Global Initiative for Asthma strategy 2021: executive summary and rationale for key changes. Am J Respir Crit Care Med. 2022;205:17–35.
18. Chung KF, et al. International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma. Eur Respir J. 2013;43:343–73.
19. Davis F. A technology acceptance model for empirically testing new end-user information systems: theory and results. Doctoral dissertation, MIT Sloan School of Management. 1986.
20. Liaw SY, et al. Desktop virtual reality versus face-to-face simulation for team-training on stress levels and performance in clinical deterioration: a randomised controlled trial. J Gen Intern Med. 2023;38:67–73.
21. Jiang N et al. Effectiveness of virtual simulations versus mannequins and real persons in medical and nursing education: meta-analysis and trial sequential analysis of randomized controlled trials. J Med Internet Res. 2024; 26, e56195.
22. Larsen DP, Butler AC, Roediger HL.. 3rd. Test-enhanced learning in medical education. Med Educ. 2008;42:959–66.

23. McConnell MM, St-Onge C, Young ME. The benefits of testing for learning on later performance. Adv Health Sci Educ Theory Pract. 2015;20:305–20.
24. Hift RJ. Should essays and other open-ended-type questions retain a place in written summative assessment in clinical medicine? BMC Med Educ. 2014;14:249.
25. Knof H, Berndt M, Shiozawa T. Prevalence of Dunning-Kruger effect in first semester medical students: a correlational study of self-assessment and actual academic performance. BMC Med Educ. 2024;24:1210.
26. von Hoyer J, et al. False certainty in the acquisition of anatomical and physiotherapeutic knowledge. BMC Med Educ. 2022;22:765.
27. Massoth C, et al. High-fidelity is not superior to low-fidelity simulation but leads to overconfidence in medical students. BMC Med Educ. 2019;19:29.
28. Öncü S, Torun F, Ülkü HH. AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. BMC Med Educ. 2025;25:278.
29. Zeng J et al. Embracing the future of medical education with large language model-based virtual patients: scoping review. J Med Internet Res. 2025; 27, e79091.
30. Çiçek FE, Ülker M, Özer M, Kıyak. Y.S. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. Postgrad Med J. 2025;101:458–63.
31. Öncü S, Torun F, Ülkü HH. Comparison of large language models for clinical scenario generation in medical education: a mixed-methods study. BMC Med Educ. 2026;26:412.
32. Wang R, Liu C, Yi F. Reconfiguring tutor-peer roles to foster critical thinking in medical PBL assisted by tutor-managed chatGPT. Med Educ. 2025. https://doi.org/10.1111/medu.70100.
33. Lucas M, Zhang Y, Bem-haja P, Vicente P. The interplay between teachers' trust in artificial intelligence and digital competence. Educ Inf Technol. 2024;29:22991–3010.

Figures

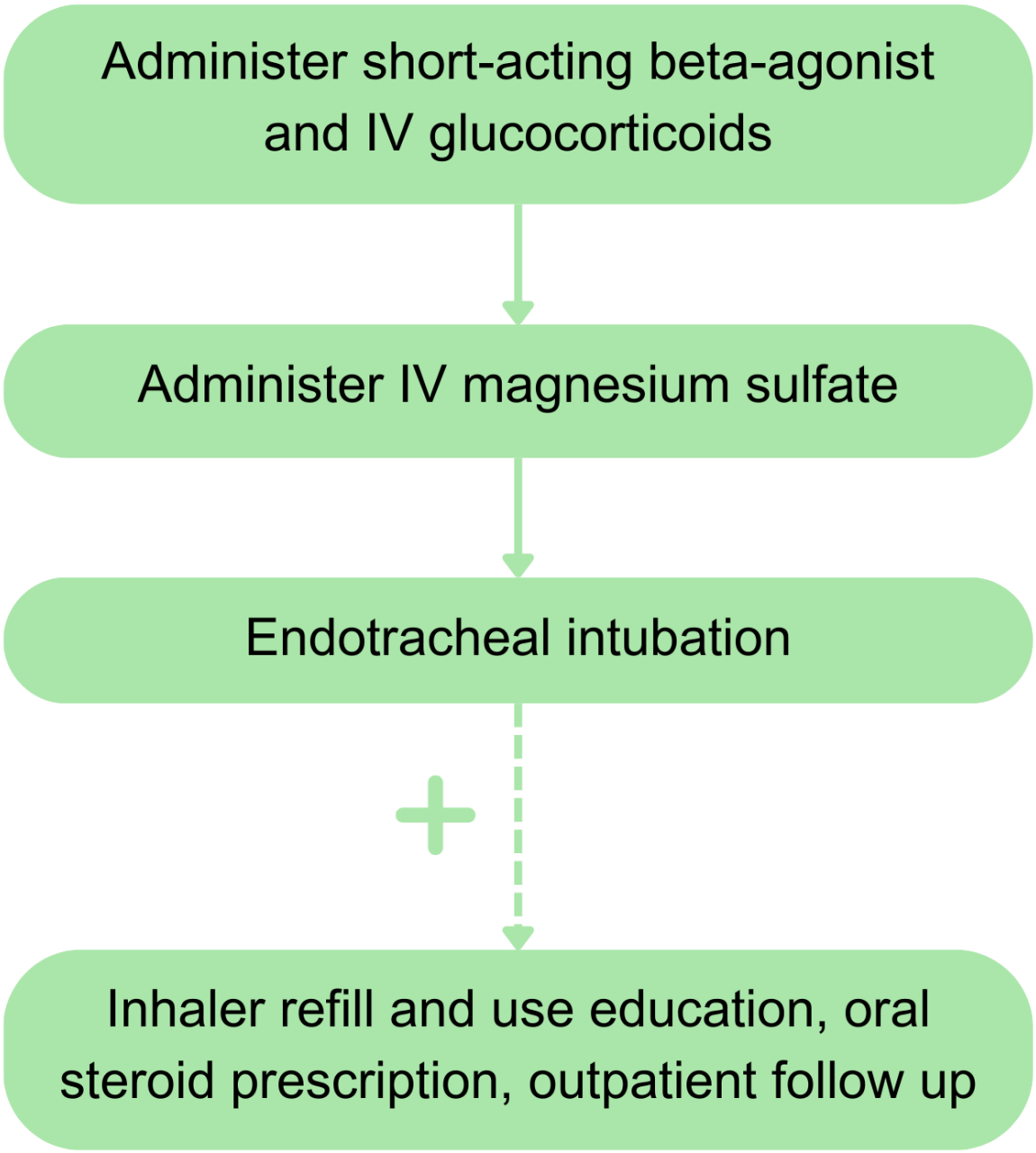


Figure 1

Acute asthma exacerbation treatment algorithm. Movement through the algorithm is dependent on a patient's failure to respond to prior treatment. The + indicates proper discharge actions that follow patient stabilization and improvement and does not necessarily follow endotracheal intubation.

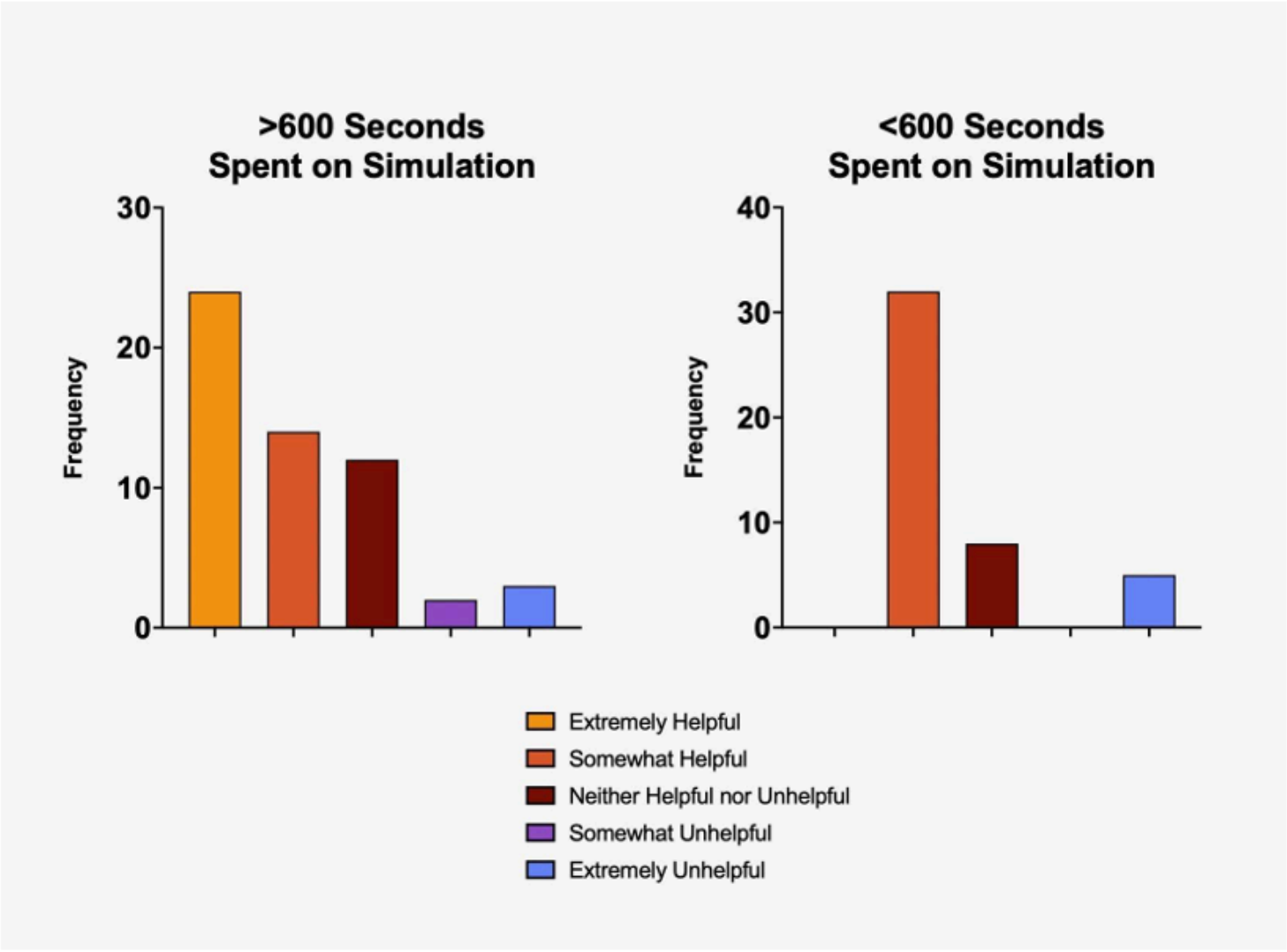


Figure 2

Distribution of student ratings of simulation helpfulness by time spent performing the simulation. Students spending >600 seconds on the simulation most frequently rated the simulations as very helpful. Students spending <600 seconds on the simulation most frequently rated the simulation as somewhat helpful.

Supplementary Files

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

- ChatGPTSimulationsPostTestNovember21202418.052.xlsx