BMC Medical Education

RESEARCH

Open Access

Artificial intelligence based resuscitation simulation: a pilot study of a novel approach to team leadership training

Altuğ Kanbakan[1*], Göksu Bozdereli Berikol[2], Buğra İlhan[3], Emel Altıntaş[1] and Fatih Doğanay[4]

Abstract

Introduction Team leadership training is essential alongside with technical training for effective resuscitation management. Addressing this gap, we developed a novel simulation system leveraging Large Language Models (LLMs) to create Artificial Intelligence (AI) agents simulating team members in Advanced Cardiovascular Life Support (ACLS) scenarios. This pilot study aimed to to develop a novel LLM-based ACLS simulation training platform and evaluate its performance in simulated resuscitation scenarios on established protocols.

Method Using the Claude 3.5 Sonnet API, we designed a simulation system with four AI agents assigned specific roles as healthcare staff within an ACLS team. Each agent strictly followed the 2020 American Heart Association (AHA) ACLS guidelines while interacting with an ACLS certified emergency medicine specialist user. The ten patient scenario transcripts were evaluated with three blinded emergency medicine specialists whether all the recommended steps are completed. Inter-rater reliability was assessed using Kendall's W and Krippendorff's Alpha statistics to evaluate agreement both within raters and the model.

Results AI agents consistently adhered to the AHA 2020 ACLS algorithm across scenarios, with a high inter-rater reliability (Kendall's $W > 0.75$). Krippendorff's Alpha values for agreement ranged from substantial (0.84) to almost perfect (0.99), indicating robust compliance with guidelines and effective simulation of resuscitation responses.

Conclusion This study highlights the potential of LLM-powered simulations as an adjunct to traditional resuscitation training. The system effectively supported team leadership training by providing consistent and guideline-compliant responses. While the results are promising, further research with larger participant samples is necessary to evaluate the long-term educational impact and scalability of such systems.

Keywords Resuscitation, Simulation, Artificial intelligence, Emergency medicine, ACLS, Medical education

*Correspondence:
Altuğ Kanbakan
altug.kanbakan@ufuk.edu.tr
[1]Department of Emergency Medicine, Ufuk University, Kızılırmak, Ufuk Ünv. Cd No:1, 06510, Ankara, Turkey
[2]Department of Emergency Medicine, Atılım University, Ankara, Turkey
[3]Department of Emergency Medicine, Kırıkkale University, Kırıkkale, Turkey
[4]Department of Research, Mayo Clinic, Jacksonville, FL 32224, USA

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

Effective team leadership is essential for improving quality of care in Advanced Cardiovascular Life Support (ACLS). Strong leadership enhances team dynamics, communication, and technical performance, leading to better CPR quality, shorter pre-shock pauses, and reduced hands-off time [1]. Leadership in healthcare guides teams through challenges and promotes collaboration and performance [2]. Simulation-based medical education improves teamwork, decision-making, and leadership skills in resuscitation scenarios [3]. Research shows it increases confidence, protocol adherence, and team performance [4, 5]. These benefits improve both patient outcomes and team effectiveness, supporting high-quality emergency care [6].

Reviews highlight the success of leadership-focused training, though more studies are needed to explore its full potential [7]. At the same time, systematic reviews emphasize that leadership training remains heterogeneous, with inconsistent definitions, limited validated assessment tools, and variable integration into curricula. Long-term retention of skills and their transfer to clinical outcomes are also not well established, and scalability across institutions continues to be a challenge [7–10]. Simulations also improve real-life leadership behaviors and team protocol compliance, enhancing patient care [11, 12].

Large Language Models (LLMs) bring new possibilities to simulation-based education. They create realistic scenarios, personalize learning, and support teamwork training [13, 14]. In In clinical education and simulation-based training environments, LLMs have been applied as Virtual Simulated Patients (VSPs), reducing costs while improving the realism of training [15]. Multi-agent simulations powered by LLMs have been shown to foster teamwork skills and collaboration [16]. However, concerns about the accuracy and clinical reliability of artificial intelligence (AI)-based training tools continue to be debated [17].

To provide effective training in team management, it is essential to establish a functioning team. In training programs designed exclusively for physicians or nurses, team members often come from the same professional category. For example, resuscitation training in undergraduate emergency medicine clerkships is conducted with medical students only. However, since real clinical teams are composed of diverse healthcare professionals, training together allows each participant to receive input from the appropriate team member according to their role, leading to more effective learning outcomes. Moreover, the number of team members may vary across different hospital settings. Beyond standardized team structures, physicians in leadership roles must be able to quickly adapt to scenarios with varying team compositions. To address these needs, we propose a game-based platform that enables practice across diverse scenarios and team sizes. Unlike prior single-agent chatbot approaches, our platform introduces multiple role-specific agents with persistent memory mapping, enabling stricter adherence to ACLS algorithms and minimizing inconsistency. In addition, the system allows structured evaluation of both the team leader and team members, directly responding to the lack of standardized assessment frameworks in leadership training. While the current focus is on task sequencing and algorithm adherence, this approach establishes a scalable and standardized foundation upon which broader leadership competencies (such as communication and situational awareness) can be built in future iterations.

Methods

System architecture and technical implementation

The ACLS simulation system was developed with JavaScript on the Cursor platform, a derivative of the Visual Studio 2022 Integrated Development Environment. The user interface was built as a web application with Hyper Text Markup Language and Cascading Style Sheets. The application was deployed to a local server using the live server functionality and tested to observe the results of real-time changes and iterations throughout the dynamic development process. The system has a command-line interface that allows users to communicate with AI agents via text. The simulation integrates a time constraint mechanism to enhance effective decision-making in resuscitation scenarios, emulating the pressing nature encountered in real resuscitation events. The architectural structure of the system is illustrated in Fig. 1 and workflow of web-application and interface is described in Figs. 2 and 3.

AI agent configuration

The simulation employed Anthropic's Claude 3.5 Sonnet model to generate four AI agents, each with distinct roles within the ACLS team. The responsibilities encompassed monitoring, defibrillation, intravenous medication administration, chest compressions, and airway management. To improve team dynamics while maintaining professional behavior, each agent was allocated distinct personality traits e.g., communicative, experienced, methodical, and decisive. The system also employed a memory mapping framework which enables agents to maintain contextual awareness of previous commands, team members' responses, and the progression of the scenario throughout the simulation.

Agents were guided through structured, role-specific prompts that explicitly linked their actions to ACLS 2020 tasks. Prompt design emphasized clear command-response structures, ensuring that agents only act when

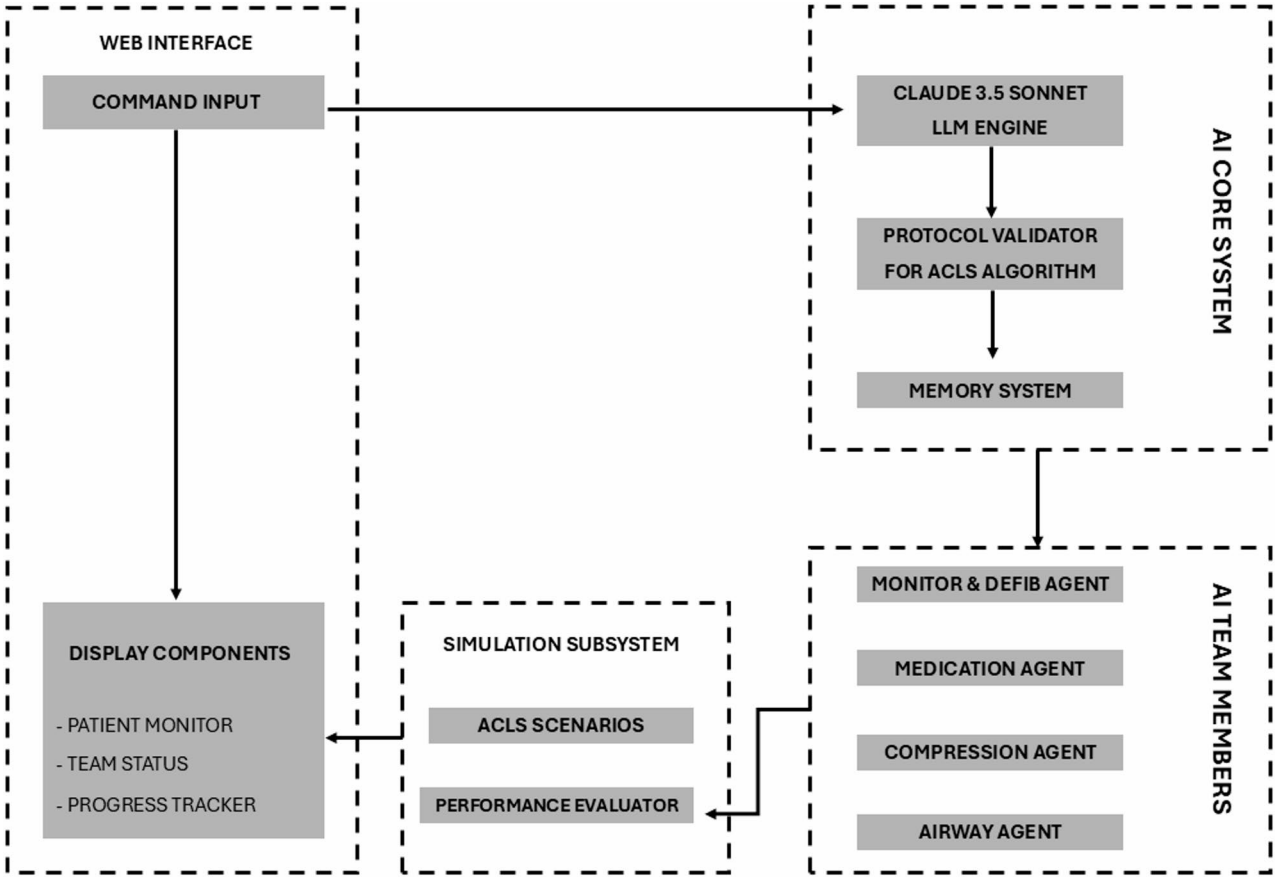


Fig. 1 Architectural structure of the in silico ACLS team leadership simulation system. The system comprises four primary components: (1) A web interface for user interaction and display, (2) An AI core system powered by Claude 3.5 Sonnet LLM for command processing and protocol validation, (3) A simulation subsystem managing ACLS scenarios and performance evaluation, and (4) AI team members representing specific clinical roles

addressed by the team leader. Closed-loop communication was operationalized by requiring agents to acknowledge leader instructions, confirm execution, and provide only immediate, factual feedback.

The AI framework was programmed to ensure strict adherence to the American Heart Association (AHA)'s 2020 ACLS algorithm. Agents were designed to execute role-specific tasks based on team leader directives while implementing closed-loop communication protocols. The system maintained a passive stance regarding subsequent steps unless explicitly queried, ensuring that agents provided patient-focused responses during interventions. Synchronization across agents was maintained through a time-stamped action log that coordinated outputs, preventing overlapping responses and maintaining the flow of realistic team dynamics. To mitigate hallucinations, all outputs were bound by ACLS guideline content, and the refusal behavior of the Claude 3.5 Sonnet model limited unsupported or speculative responses.

Figure 3 demonstrates the web-based interface of the ACLS team leadership simulation. The interface includes (1) a case summary providing patient background, (2) a real-time patient monitoring section displaying ECG rhythms, (3) an AI-driven virtual team panel representing resuscitation team roles, (4) a simulation progress bar tracking scenario completion, (5) a command input area for user interaction with AI agents, and (6) a simulation output log displaying executed commands and system feedback. The user "Team Leader" starts the game, the scenario of the patient is given to the Team Leader and waits for his/her orders. The leader interacts with the AI based agents who obey the orders that leader commands and provide responses as feedbacks which are strictly keen to the ACLS 2020 guidelines. The interactions continue until the patient has ROSC. The system records the duration of the user interactions and response time until the scenario ends.

Scenario design and implementation

Ten patient information were created by two ACLS-certified emergency medicine specialists aligned with ACLS training cases. For each rhythm, a brief clinical vignette of approximately 9–10 words patient information was provided as the case entry point (e.g., "60-year-old male with

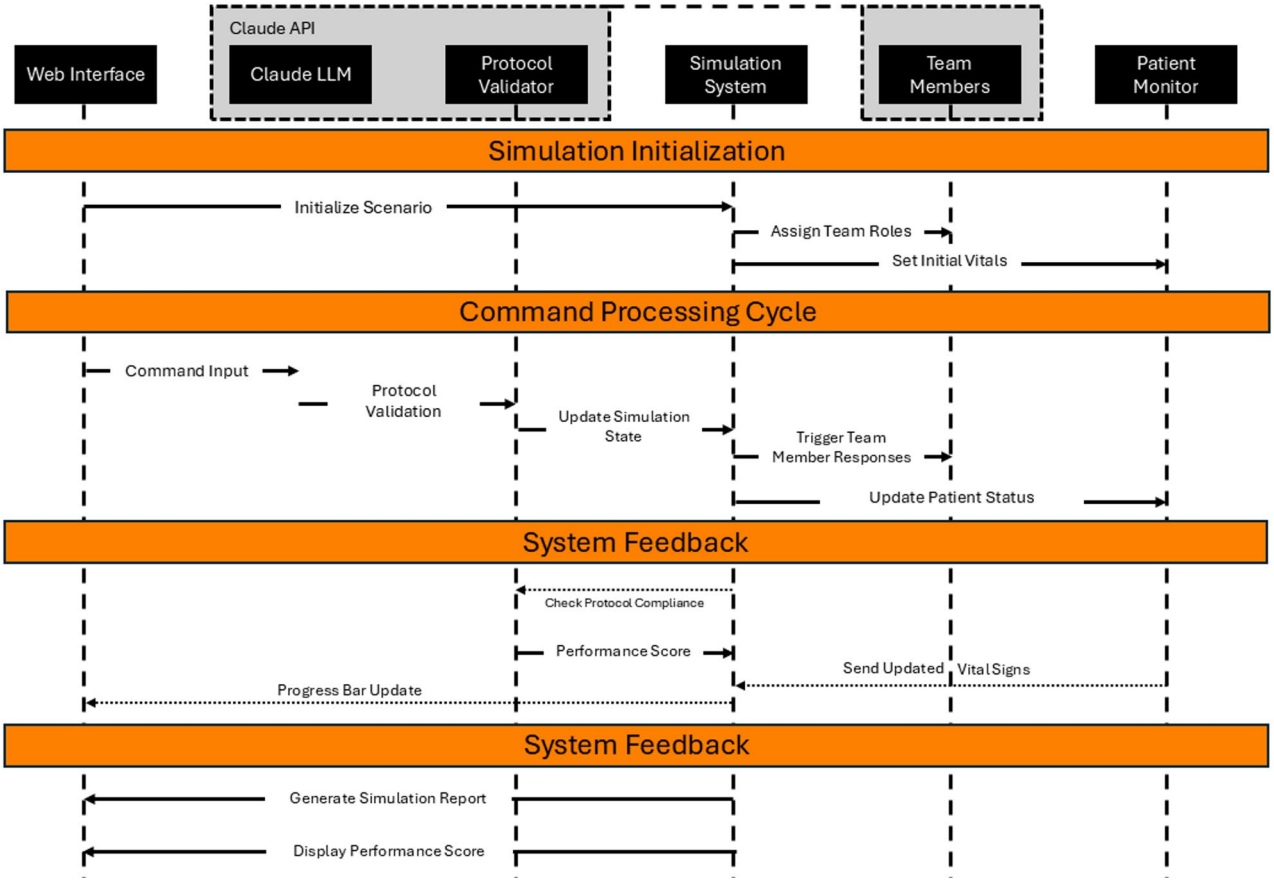


Fig. 2 The procedure is segmented into three primary phases: Simulation Initialization involves the establishment of scenarios, assignment of team roles, and configuration of initial patient vitals; Command Processing Cycle encompasses user input validation, updates to the simulation state, and responsive actions from team members; and System Feedback delivers protocol compliance assessments, performance evaluations, and progress reports

diabetes, found down by family member" or "28-year-old female with history of anorexia nervosa found unresponsive"). These initial prompts were deliberately concise to mirror the limited information often available at the start of real resuscitations. Each case was mapped to one of the four ACLS rhythms (ventricular fibrillation, pulseless ventricular tachycardia, pulseless electrical activity, and asystole). The simulation platform, powered by LLM-based agents, was instructed to (1) instantiate the appropriate resuscitation team roles (e.g., physician, nurse, respiratory therapist), (2) interact with the user in real time according to ACLS 2020 guidelines, and (3) terminate the scenario upon return of spontaneous circulation (ROSC) or completion of the resuscitation algorithm. The entire transcript of team leader commands and AI-agent responses was automatically generated by the system and used for evaluation. Scenario types and clinical contexts are listed in Supplement 1.Each patient scenario includes the initial rhythm (VF/pVT/Asystole/PEA) and return of spontaneous circulation. Rhythm changes were not added to these scenarios during software development for the purpose of testing the concept.

Testing the platform

The system is tested by an ACLS certified emergency medicine specialist ("The user"). The game starts with a short triggering patient information such as "55-year-old male with history of hypertension collapses while jogging". The given scenarios are shown in the screen and the input from the user was awaited. For example, the user reads the history of the patient 55yearold collapsed patient and instructed appropriate person the first to do in the ACLS steps such as "Nurse Davis check the pulse" to an AI agent Nurse Davis. The AI agent "Nurse Davis" responses to the user "No pulse" and awaits the user second input. The process is shown in Fig. 3.

Data collection

The system was designed to evaluate adherence to ACLS algorithms and task sequencing. Other aspects of leadership training, such as closed-loop communication or non-technical skills, were beyond the scope of this study. At the end of each scenario, the simulation results generated by the system were exported as a text document

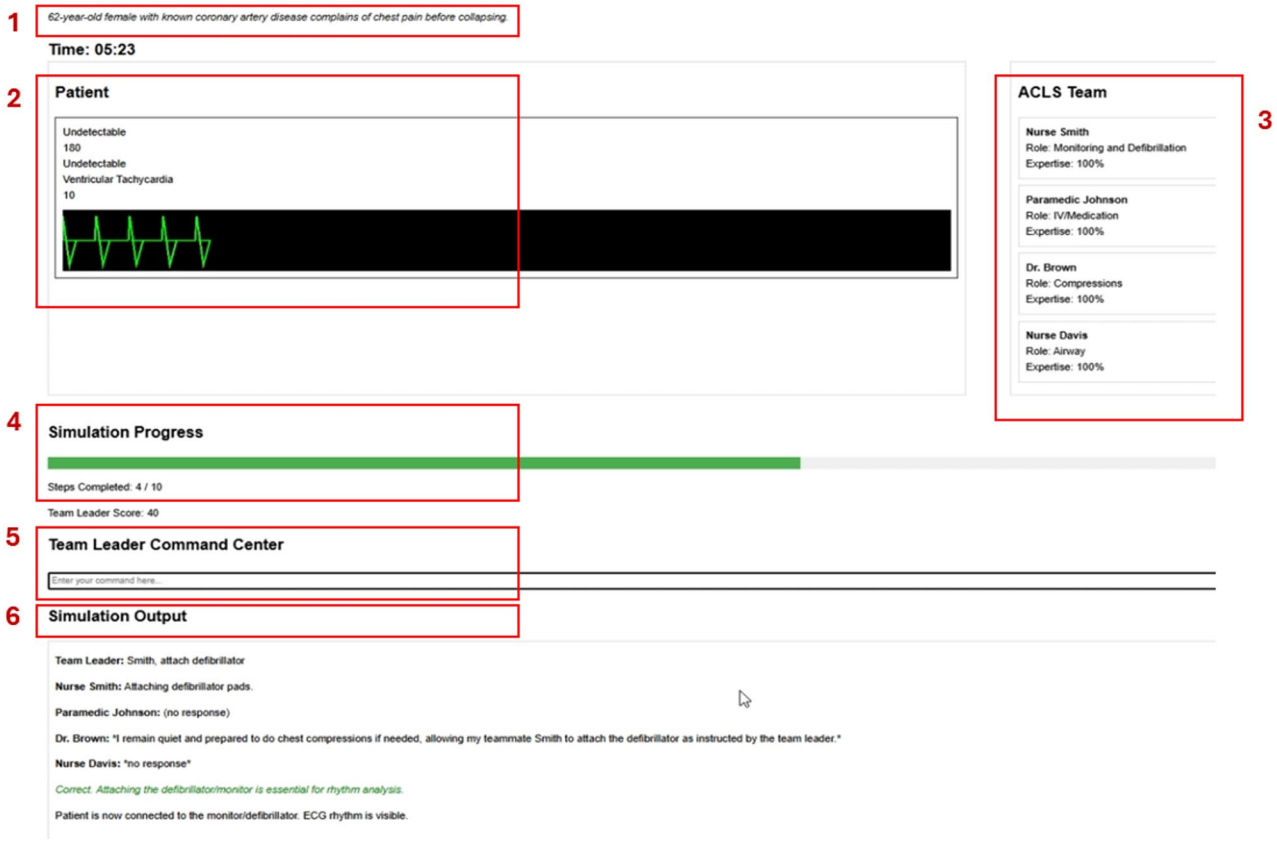


Fig. 3 Web-based interface of the ACLS team leadership simulation. The interface includes (1) a case summary providing patient background, (2) a real-time patient monitoring section displaying ECG rhythms, (3) an AI-driven virtual team panel representing resuscitation team roles, (4) a simulation progress bar tracking scenario completion, (5) a command input area for user interaction with AI agents, and (6) a simulation output log displaying executed commands and system feedback

and scored according to the checklist given in Table 1 and transferred to Microsoft Excel (Microsoft Corp).

Performance assessment

User performance evaluation was conducted through a scoring system that tracked adherence to the ACLS algorithm. The system monitored the accuracy of rhythm interpretation, appropriateness of interventions, and adherence to ACLS protocol sequences.

Expert evaluation and reliability assessment

Three 10-year experienced emergency medicine physicians with certified ACLS training independently evaluated the AI-generated ACLS scenarios to determine the system's reliability in producing clinically appropriate training simulations. The evaluation concentrated on compliance with ACLS protocols through a standardized checklist based on the guidelines established by the American Heart Association. The raters evaluated each critical decision point and intervention utilizing a scoring system as shown in Table 1. The raters completed their assessments independently and they were blinded to one another's scores.

Data analysis

To validate reliability across different statistical paradigms, analysis was performed using Jamovi Package Program version 2.1.6. Krippendorff's alpha (α) coefficient was calculated to assess the overall reliability of ratings across multiple raters, with values closer to 1 indicating stronger agreement. Additionally, inter-rater reliability was assessed using Fleiss kappa coefficients for each scenario type to evaluate pairwise agreement between raters. Statistical significance was set at $p < 0.05$ for all analyses.

Results

The scenarios were tested by an ACLS certified emergency medicine specialist("The user"). and the transcript of the scenarios (Supplement 2) were investigated regarding the criteria in this pilot evaluation (Table 1). The overall duration of each scenario, the steps completed, and the evaluation score given by the simulator was shown in Table 2. Median of the raters for each scenario was shown in Table 3.

The interrater reliability was calculated with Fleiss Kappa and weighted with Kendalls coefficient of

Table 1 Checklist evaluation of the ACLS simulation

	Starting CPR (1pt)	CPR Quality (5pt)	Checking Rhythm (1pt)	Shocking (1pt)	Ordering IV/IO access (1pt)	Checking shockable rhythm, assess, resume CPR (3pt)	Ordering Epinephrine (1pt)	Treating reversible causes (1pt)
pVT1	X	X	X	X	X	X	X	X
pVT2	X	X	X	X	X	X	X	X
PEA1	X	X	X		X	X	X	X
PEA2	X	X	X		X	X	X	X
VF1	X	X	X	X	X	X	X	X
VF2	X	X	X	X	X	X	X	X
VF3	X	X	X	X	X	X	X	X
Asystole1	X	X	X		X	X	X	X
Asystole2	X	X	X		X	X	X	X
Asystole3	X	X	X		X	X	X	X

VF Ventricular fibrillation, *pVT* Pulseless ventricular tachycardia, *PEA* Pulseless electrical activity, CPR Quality has 5 criterias: 2 inches push, rate 100–120/min, 30:2 C: V ratio, hand placement in lower of sternum, avoiding excessive ventilation. The scoring system is based on the ACLS guideline steps. Every step I defined as 1 point each. In patients with pulseless electrical activity doesn't have the step "Shocking" thus NEA/Asystole patients have the total score of 9, VF/Pulseless VT have a total score of 10. After having all 10 patient scenario transcripts, 3 EM Specialist evaluated the texts whether it includes all the steps supposed to be completed in ACLS checklist. The raters are also evaluated for reliability. Afterwards the system is evaluated with its scores. The raters only evaluated the transcripts and blind with system's scoring

Table 2 The overall duration, steps, and score of the LLM assessed

Scenario	Time (Sec)	Steps Completed	LLM self-evaluation score
VF1	263	10/10	100
VF2	338	10/10	100
VF3	349	10/10	100
pVT1	325	10/10	100
pVT2	320	10/10	100
PEA1	351	9/9	100
PEA2	303	9/9	100
Asystole1	367	9/9	100
Asystole2	334	9/9	100
Asystole3	326	9/9	100

VF Ventricular fibrillation, *pVT* Pulseless ventricular tachycardia, *PEA* Pulseless electrical activity

concordance W (Table 4). As shown in Table 4, there is a very high level of agreement among the raters (Kendall $W > 0.754$, $p < 0.05$). This finding suggests that the raters' rankings are statistically consistent in this pilot evaluation.

Krippendorff's Alpha was calculated to measure the scores' agreement with the matching ACLS criterias and we found moderate agreement in one pVT and one VF scenarios and one insufficient agreement in one VT scenario. Seven scenarios are all highly exceeding 0.80 with a strong agreement indicating high reliability.

The agreement between the gold standard and the raters' consensus was assessed using Krippendorff's Alpha, Cohen's Kappa, and Weighted Kappa. Krippendorff's Alpha values indicated substantial to near-perfect reliability in most scenarios, with values ranging from-0.602 (pVT2) to 0.996 (Asystole2, PEA1, PEA2). Cohen's Kappa values showed statistically significant results across all scenarios ($p < 0.05$), with agreement levels ranging from fair (0.390 for pVT2) to excellent (0.919 for Asystole2, PEA1, and PEA2). Weighted Kappa provided additional insight into agreement magnitude, showing substantial agreement in scenarios such as Asystole2, PEA1, and PEA2 (Weighted Kappa = 0.830, $p < 0.001$) but lower agreement in others, such as VT2 (Weighted Kappa = 0.348, $p = 0.035$).

The high level of agreement between the raters and the gold standard checklist, shown by strong Krippendorff's Alpha values (ranging from 0.844 to 0.996). This highlights the system's reliability in providing consistent performances. Similarly, the statistically significant Weighted Kappa and Cohen's Kappa values, especially in VF/pVT scenarios ($\kappa = 0.78$), confirm that the system

Table 3 Median of the raters for each scenario

	Score	Asys1	Asys2	Asys3	PEA1	PEA2	pVT1	pVT2	VF1	VF2	VF3
N	9	9	9	9	9	9	10	10	10	10	10
Median*	15	10	14	13	14	14	10	8	11	12	11
Min**	14	9	13	10	10	10	9	7	11	12	10
Max**	15	10	14	14	14	14	10	9	11	13	12

VF Ventricular fibrillation, *pVT* Pulseless ventricular tachycardia, *PEA* Pulseless electrical activity, N = Number of items, * Median of total score of each rater.** Total score 14 indicates non-shockable rhythms, Total score 15 indicates shockable rhythms

Table 4 Interrater reliability of each scenario

Scenario	Fleiss Kappa	Z	*p*-values for Fleiss' Kappa	95% CI	Kendall W	*p*-values for Kendall W
VF1	0.450	3.48	< 0.01	0.039–1.09	0.860	0.006
VF2	0.639	4.35	< 0.01	0.604–0.709	1	0.001
VF3	0.549	4.33	< 0.01	0.426–0.727	0.884	0.005
pVT1	0.771	5.42	< 0.01	0.388–1.23	0.884	0.005
pVT2	0.583	3.20	0.001	0.077–1.22	0.754	0.016
PEA1	0.487	4.09	< 0.01	0.198–0.899	0.931	0.004
PEA2	0.487	4.09	< 0.01	0.196–0.916	0.931	0.004
Asystole1	0.727	5.01	< 0.01	0.401-1.10	0.886	0.006
Asystole2	0.803	5.97	< 0.01	0.523–1.12	1	0.002
Asystole3	0.540	4.66	< 0.01	0.420–0.711	0.937	0.004

VF Ventricular fibrillation, *pVT* Pulseless ventricular tachycardia, *PEA* Pulseless electrical activity, *Asys* Asystole

Table 5 The gold standard agreement with ACLS criteria and the raters scores

Scenario	Krippendorff α	Cohen κ	*p*-value for Cohen κ	Weighted κ	Z-score	*p* for Weighted κ
Asystole1	0.820	0.431	0.017	0.333	2.12	0.034
Asystole2	0.996	0.919	0.003	0.830	3.49	< 0.001
Asystole3	0.844	0.859	0.007	0.710	3.49	< 0.001
PEA1	0.996	0.919	0.003	0.830	3.49	< 0.001
PEA2	0.996	0.919	0.003	0.830	3.49	< 0.001
pVT1	0.697	0.667	0.018	0.474	3.16	0.002
pVT2	-0.602	0.390	0.035	0.348	2.11	0.035
VF1	0.993	0.783	0.002	0.643	3.16	0.002
VF2	0.995	0.783	0.002	0.643	3.16	0.002
VF3	0.697	0.667	0.018	0.474	3.16	0.002

VF Ventricular fibrillation, *pVT* Pulseless ventricular tachycardia, *PEA* Pulseless electrical activity

effectively simulates realistic and well-structured resuscitation scenarios (Table 5).

Discussion

This pilot study shows that agent based generative models can be used for individual resuscitation training with their ability to provide team members, also assessment of task order adherence to given protocols. The negative Krippendorff's Alpha observed in pVT2 reflects a disagreement with standard score, most likely due to the greater clinical ambiguity and the presence of multiple acceptable management pathways in this scenario, rather than a calculation error. Scenarios involving shockable rhythms such as VF/VT, which follow clear guidelines, demonstrated higher reliability compared to asystole cases that require more complex decision-making. This suggests that the simulation system performs well in structured, protocol-based situations but may need further refinement to effectively manage more variable and challenging cases. In scenarios like pVT2 and VT2, agreement was lower. Although these rhythms follow a clear algorithm, small differences in step sequencing or interpretation appeared to reduce interrater reliability. This highlights a current limitation of the system: while it performs well in strictly structured cases, variability in execution or interpretation can still challenge consistency. Further development of the AI agents is needed to better capture this nuance.

Traditional CPR certification training includes various elements such as video-supported lessons, mannequin-based simulated clinical scenarios, and group practice sessions. However, these methods have few limitations. They require trained instructors and physical training spaces, which increase costs and reduce flexibility. Additionally, the necessity for both participants and instructors to be physically present creates spatial and time constraints [18]. Given these challenges, there is a growing need for scalable, standardized, and flexible training alternatives.

AI-based applications have the potential to address these limitations by providing cost-effective and structured training programs. They can reduce reliance on trained instructors, support self-directed learning, and improve implementation efficiency [19, 20]. In educational contexts, AI-based systems have been reported to enhance communication training by offering instant feedback and guiding students through structured learning experiences [21, 22]. Therefore, these applications have the potential to serve as effective educational tools [23].

Leadership, crisis resource management, communication among team members and situational awareness are key factors that enhance outcomes in both real and

simulated settings according to ERC [24–26]. Despite this evidence, the need for standardized leadership training in resuscitation remains. The 2020 International Liaison Committee on Resuscitation consensus emphasized the importance of developing structured training programs [27]. This suggests that LLM-based simulations could serve as practical and cost-effective solutions. Studies further show that leadership and teamwork are closely related to performance and error rates [28]. Effective leadership in resuscitation requires not only managing technical tasks but also adapting to team dynamics in high-stress scenarios [25]. Therefore, communication-focused training programs can enhance team collaboration and efficiency [29].

To address these challenges, AI-based simulation systems offer a promising solution. By adjusting for team size variations and different clinical scenarios, AI can personalize leadership training to meet individual needs. Additionally, remote access allows participants to engage in training regardless of location, increasing accessibility and scalability.

In traditional ACLS training settings, team roles typically include airway, compressor, monitor/defibrillator, timer, IV/IO medication, and team leader. However, maintaining this structure in both training and real-life scenarios is not always feasible. The recommended number of team members in resuscitation training varies between 2 and 10 [30, 31]. The tool presented in this study enables leadership training and communication practice with teams of different sizes, offering a flexible approach to resuscitation education.

The accuracy of ChatGPT in following ACLS guidelines for cardiac arrest and bradycardia cases has been evaluated in a study that tested twenty different scenarios, where ChatGPT achieved 69% accuracy in arrest scenarios and 42% in bradycardia scenarios [32]. Researchers reported that ChatGPT could provide different responses to identical inputs, skip critical steps, and suggest incorrect drug dosages [32]. Similarly, a ChatGPT-3.5-based simulation module for medical students was evaluated, focusing on ACLS and intensive care case scenarios. This study emphasized the importance of prompt engineering, as variations in prompts and punctuation affected ChatGPT's responses. The researchers also noted that ChatGPT's explanations in ACLS scenarios were sometimes incomplete, suggesting the need for a more robust feedback system [33].

Unlike these studies, our research did not utilize a direct chatbot interface. Instead, we developed four distinct AI agents based on the Claude 3.5 Sonnet model, each assigned specific ACLS team roles. Additionally, our model incorporates a memory mapping framework that allows AI agents to track previous commands, responses from other team members, and scenario progression. This feature ensures that the agents provide patient-centered and consistent responses during interventions.

In one study, 77% of ChatGPT-3.5's responses and 84% of ChatGPT-4's responses aligned with the 2021 European ERC guidelines [34]. However, these guidelines were not explicitly included in the training data of these chatbots, meaning that AI-generated responses relied on statistical patterns rather than true comprehension. Consequently, AI-generated content may sometimes be incomplete, superficial, or incorrect. In particular, missing critical messages and generating 'hallucinations' remain significant risks to clinical accuracy [34]. Compared to general chatbot-style LLMs such as ChatGPT, Claude 3.5 Sonnet was selected for its lower hallucination rate and greater consistency, and the use of multiple role-specific agents with memory mapping further reduced inconsistency and improved adherence to ACLS guidelines. This limitation is consistent with broader evidence that ChatGPT frequently produces fabricated references and factual inaccuracies, raising concerns about its reliability in evidence-dependent settings [35]. In a peer-reviewed evaluation across both common and complex diagnostic vignettes, Claude 3.5 Sonnet achieved a 96.7% overall accuracy rate, outperforming earlier GPT-3.5 models and closely trailing GPT-4o (98.3%) while demonstrating greater stability across case types [36] and Claude 3.5's documented refusal policy adds a safeguard by reducing the likelihood of confidently stated falsehoods [37]. Our multi-agent simulation assigns distinct roles to Claude 3.5 agents and links them through shared memory, which helps reduce factual drift, preserve context, and maintain adherence to ACLS protocols. This design addresses common weaknesses of single agent chatbots and supports consistent, reproducible training in critical care scenarios.

In our study, AI agents with distinct roles were developed. To ensure AI understands clinical context beyond memorizing guideline information, we implemented a memory mapping method, ensuring strict adherence to the AHA's ACLS algorithm. Unlike simple question-answer chatbot models, this approach provides a more structured and reliable system that simulates real resuscitation team dynamics. Therefore, adopting structured and controlled methodologies in AI-assisted clinical applications is crucial for accuracy and reliability. A key limitation of the present study is that the system did not assess or train leadership competencies beyond algorithm adherence. Critical components such as communication, situational awareness, and team coordination remain unaddressed and will require further development in future iterations of the platform. Also, the study did not aim to evaluate the content validation, usability testing, or fidelity assessment of the system. The small sample size and focus on specific cardiac rhythms may

limit how well the findings apply to broader contexts. Although testing the platform with only one user introduces limitations regarding representativeness and generalizability. For this initial pilot, however, we deliberately selected an ACLS-certified emergency physician as the single test user. Our rationale was to minimize variability due to user errors and ensure that any observed inaccuracies could be attributed to system performance rather than user inexperience. We considered this a more standardized way to identify potential system flaws in the early development stage. Nevertheless, we fully acknowledge that this approach limits the generalizability of our findings. In future validation studies, we plan to include a larger and more diverse cohort of participants, including novice learners and multidisciplinary team members, to better assess usability, variability in interaction styles, and educational effectiveness. Another limitation is that the system demonstrated higher reliability in shockable rhythm cases, likely due to their more linear and predictable management protocols. Non-shockable rhythms, which demand greater adaptability and critical thinking, were less consistent and not fully captured by the current design. This finding may also have been influenced by the use of a single experienced test user, whose decision-making style and familiarity with ACLS algorithms could have biased the results. Future studies with a larger and more diverse participant group will be necessary to better evaluate system performance, particularly in non-shockable scenario Although AI agents followed protocols accurately, they may not fully capture non-verbal cues and emotional interactions in real resuscitation situations. The system also lacks the ability to simulate rare but critical complications. Finally, potential biases stemming from the pre-defined AI role traits and prompt engineering may also have influenced system performance.

Conclusion

This pilot study demonstrates the performance of LLM-powered agents simulating ACLS team roles with high guideline compliance. The system showed strong reliability in providing consistent performances except one VT scenario and high agreement with standard checklist. While they are not yet a replacement for traditional methods, they provide a structured and scalable approach to developing and assessing team leadership in resuscitation scenarios. AI-based applications have the potential to offer consistent, cost-effective contributions to ACLS team leadership training. Standardized AI-driven assessments enable objective evaluation of leadership skills, while real-time feedback on performance metrics may accelerate skill development.

Future research should focus on larger validation studies with diverse participant groups and expanded scenarios to improve the performance of the systems. Additionally, it should assess the long-term impact on clinical performance and patient outcomes. Improving AI capabilities to simulate complex team interactions and rare complications could further enhance its educational effectiveness.

Abbreviations

AI	Artificial Intelligence
AHA	American Heart Association
ACLS	Advanced Cardiac Life Support
ERC	European Resuscitation Council
LLM	Large Language Model
pVT	Pulseless Ventricular Tachycardia
PEA	Pulseless Electrical Activity
VF	Ventricular Fibrillation

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-025-08539-z.

Supplementary Material 1.

Acknowledgements

Fatih Doğanay, acknowledges a postdoctoral grant from the Scientific and Technical Research Council of Turkey (TUBITAK, 2219- International Postdoctoral Research Scholarship Programme, 1059B192302217). Additionally, the authors declare that ChatGPT-4o and DeepL were used for translation and language editing during the preparation of this manuscript. The authors reviewed and edited the output as needed and take full responsibility for the content of the publication.

Authors' contributions

AK: Conceptualization, Methodology, Software, Resources, Writing – original draft, Visualization, Project administration . GBB: Conceptualization, Validation, Data curation, Writing – original draft . Bİ: Validation, Formal analysis . EA: Validation. FD: Investigation, Writing - review & editing . All authors read and approved the final manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The datasets generated and analyzed during the current study are available in the supplementary files. Further inquiries can be directed to the corresponding author.

Declarations

Ethics approval and consent to participate

The study protocol was reviewed by the Ufuk University Health Sciences Scientific Research Assessment and Ethics Committee (Ankara, Turkey). It was determined that formal ethical approval and informed consent were not required (Waiver granted), as the study consisted entirely of computer-based simulations and did not involve human subjects or the collection of identifiable personal data, in accordance with the principles of Non-Interventional Clinical Research (Decision Date: 05.12.2025, Decision No: 25.11.21.09/04). However, to ensure adherence to high ethical standards, written informed consent to participate was obtained from the participant involved in the pilot evaluation prior to the study. The study was conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments.

Consent for publication

Not applicable.

Competing interests
The authors declare no competing interests.

Received: 27 May 2025 / Accepted: 26 December 2025
Published online: 05 January 2026

References

1. Yeung JHY, Ong GJ, Davies RP, Gao F, Perkins GD. Factors affecting team leadership skills and their relationship with quality of cardiopulmonary resuscitation. Crit Care Med. 2012;40(9):2617–21. https://doi.org/10.1097/CCM.0b013e3182588a9f.
2. Kello JE, Allen JA. The burned out physician: managing the stress and reducing the errors. 1st ed. Cambridge: Cambridge University Press; 2022. https://doi.org/10.1017/9781009052795.
3. Amatya Y, Shrestha S. Evaluating the impact of ACLS scenario-based resuscitation simulation for emergency team. Med J Shree Birendra Hosp. 2021;20:12–8.
4. Zern SC, Marshall WJ, Shewokis PA, Vest MT. Use of simulation as a needs assessment to develop a focused team leader training curriculum for resuscitation teams. Adv Simul. 2020;5:6. https://doi.org/10.1186/s41077-020-00126-5.
5. Toft LEB, Bottinor W, Cobourn A, et al. A simulation-enhanced, spaced learning, interprofessional code blue curriculum improves ACLS algorithm adherence and trainee resuscitation skill confidence. J Interprof Care. 2023;37(5):623–8. https://doi.org/10.1080/13561820.2022.2122183.
6. Greer A, Selladurai RI, Pfeifle AL, et al. Effective teamwork and healthcare delivery outcomes. In: Selladurai R, Hobson C, editors. Advances in healthcare information systems and administration. Hershey, PA: IGI Global; 2020. pp. 77–99. https://doi.org/10.4018/978-1-7998-2348-5.ch005.
7. Rosenman ED, Shandro JR, Ilgen JS, Harper AL, Fernandez R. Simulation-based healthcare leadership training: a systematic review. Simul Healthc. 2013;8(6):356–72. https://doi.org/10.1097/SIH.0b013e31829e59dc.
8. Ford K, Menchine M, Burner E, Arora S, Inaba K, Demetriades D, Yersin B, Lin S. Leadership and team behaviours are the critical factors in the management of simulated resuscitations by medical students. Resuscitation. 2016;99:62–7. https://doi.org/10.1016/j.resuscitation.2015.11.010.
9. Lapierre V, Fernandez R, Auerbach M, Bhanji F, Cheng A, Grant V, et al. Leadership in pediatric resuscitation: a review and future directions. Pediatr Emerg Care. 2020;36(5):e262–8. https://doi.org/10.1097/PEC.0000000000002001.
10. McLaughlin SA, Doezema D, Sklar DP. Human simulation in emergency medicine training: a systematic review. Acad Emerg Med. 2019;26(8):1002–14. https://doi.org/10.1111/acem.13791.
11. Fernandez R, Rosenman ED, Olenick J, et al. Simulation-based team leadership training improves team leadership during actual trauma resuscitations: a randomized controlled trial. Crit Care Med. 2020;48(1):73–82. https://doi.org/10.1097/CCM.0000000000004049.
12. Fernandez Castelao E, Boos M, Ringer C, Eich C, Russo SG. Effect of CRM team leader training on team performance and leadership behavior in simulated cardiac arrest scenarios: a prospective, randomized, controlled study. BMC Med Educ. 2015;15:116. https://doi.org/10.1186/s12909-015-0414-9.
13. Gao C, Lan X, Li N, Yuan Y, Ding J, Zhou Z, et al. Large Language models empowered agent-based modeling and simulation: a survey and perspectives. Humanit Soc Sci Commun. 2024;11:1259. https://doi.org/10.1057/s41599-024-02521-3.
14. Saiedian H. Leveraging large language models in education: enhancing learning and teaching. In: 2023 ASEE Midwest Section Conference Proceedings. Lincoln, NE: University of Nebraska-Lincoln; 2024. p. 46353.
15. Li Y, Zeng C, Zhong J, Zhang R, Zhang M, Zou L, et al. Leveraging large Language model as simulated patients for clinical education. ArXiv Preprint. arXiv:2404.13066. https://doi.org/10.48550/arXiv.2404.13066. Published April 2024.
16. Sprigler A, Drobek A, Weinstock K, Tapsoba W, Childress G, Dao A, et al. Synergistic simulations: multi-agent problem solving with large language models. ArXiv Preprint. arXiv:2409.13753. https://doi.org/10.48550/arXiv.2409.13753. Published September 2024.
17. Semeraro F, Fijačko N, Gamberini L, Bignami EG, Greif R. The gap between promise and reality: evaluating new ai's role in CPR education. Resuscitation. 2025;208:110540. https://doi.org/10.1016/j.resuscitation.2025.110540.
18. Lee DK, Choi H, Jheon S, Jo YH, Im CW, Il SY. Development of an extended reality simulator for basic life support training. IEEE J Transl Eng Health Med. 2022;10:1–7. https://doi.org/10.1109/JTEHM.2022.3179342.
19. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthc (Basel). 2023;11(6):887. https://doi.org/10.3390/healthcare11060887.
20. Cadario R, Longoni C, Morewedge CK. Understanding, explaining, and utilizing medical artificial intelligence. Nat Hum Behav. 2021;5:1636–42. https://doi.org/10.1038/s41562-021-01146-9.
21. Shorey S, Ang E, Yap J, Ng ED, Lau ST, Chui CK. A virtual counseling application using artificial intelligence for communication skills training in nursing education: development study. J Med Internet Res. 2019;21(10):e14658. https://doi.org/10.2196/14658.
22. Shorey S, Ang ENK, Ng ED, Yap J, Lau LST, Chui CK, et al. Evaluation of a theory-based virtual counseling application in nursing education. CIN Comput Inf Nurs. 2023;41(7):385–93. https://doi.org/10.1097/CIN.0000000000000987.
23. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ. 2019;5(1):e13930. https://doi.org/10.2196/13930.
24. Greif R, Lockey AS, Conaghan P, et al. European resuscitation Council guidelines for resuscitation 2015. Resuscitation. 2015;95:288–301. https://doi.org/10.1016/j.resuscitation.2015.07.017.
25. Greif R, Lockey A, Breckwoldt J, et al. European resuscitation Council guidelines 2021: education for resuscitation. Resuscitation. 2021;161:388–407. https://doi.org/10.1016/j.resuscitation.2021.02.011.
26. Perkins GD, Gräsner JT, Semeraro F, et al. European resuscitation Council guidelines 2021: executive summary. Resuscitation. 2021;161:1–60. https://doi.org/10.1016/j.resuscitation.2021.02.012.
27. Greif R, Bhanji F, Bigham BL, et al. Education, implementation, and teams. Resuscitation. 2020;156:A188–239. https://doi.org/10.1016/j.resuscitation.2020.02.014.
28. Hosseini M, Heydari A, Reihani H, Kareshki H. Elements of teamwork in resuscitation: an integrative review. Bull Emerg Trauma. 2022;10(2):47–55. https://doi.org/10.30476/beat.2021.91963.1291.
29. Farquharson B, Cortegiani A, Lauridsen KG, Yeung J, Greif R, Nabecker S. Teaching team competencies within resuscitation training: a systematic review. Resusc Plus. 2024;19:100687. https://doi.org/10.1016/j.resplu.2024.100687.
30. Siebig S, Kues S, Klebl F, Brünnler T, Rockmann F, Schölmerich J, et al. Cardiac arrest. Dtsch Arztebl Int. 2009;106(6):105–11. https://doi.org/10.3238/arztebl.2009.0065.
31. American Heart Association. Advanced cardiovascular life support provider manual. International english ed. Dallas, TX: American Heart Association; 2020.
32. Pham C, Govender R, Tehami S, Chavez S, Adepoju OE, Liaw W. ChatGPT's performance in cardiac arrest and bradycardia simulations using the American heart association's advanced cardiovascular life support guidelines: exploratory study. J Med Internet Res. 2024;26:e55037. https://doi.org/10.2196/55037.
33. 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. https://doi.org/10.2196/49877.
34. Beck S, Kuhner M, Haar M, Daubmann A, Semmann M, Kluge S. Evaluating the accuracy and reliability of AI chatbots in disseminating the content of current resuscitation guidelines: a comparative analysis between the ERC 2021 guidelines and both ChatGPT 3.5 and 4. Scand J Trauma Resusc Emerg Med. 2024;32:95. https://doi.org/10.1186/s13049-024-01234-x.
35. Chelli M, Descamps J, Lavoué V, Trojani C, et al. Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: comparative analysis. J Med Internet Res. 2024;26:e53164. https://doi.org/10.2196/53164.
36. Dinc MT, Bardak AE, Bahar F, Noronha C. Comparative analysis of large Language models in clinical diagnosis: performance evaluation across common and complex medical cases. JAMIA Open. 2025;8(3):ooaf055. https://doi.org/10.1093/jamiaopen/ooaf055.
37. Anthropic. Claude 3.5 Sonnet model card. 2025. Available from: https://www.anthropic.com/index/claude-3-5-model-card.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.