

<https://doi.org/10.1038/s44484-026-00009-3>

Integrating virtual reality and large language models for team-based non-technical skills training and evaluation in the operating room



Jacob Barker¹, Doga Demirel¹✉, Cullen Jackson², Anna Johansson³, Robbin Miraglia³, Darian Hoagland³, Stephanie B. Jones⁴, John Mitchell⁵, Daniel B. Jones⁶ & Suvaranu De⁷

Although effective teamwork and communication are critical to surgical safety, structured training for non-technical skills (NTS) remains limited compared with technical simulation. The ACS/APDS Phase III Team-Based Skills Curriculum calls for scalable tools that both teach and objectively assess these competencies during laparoscopic emergencies. We introduce the Virtual Operating Room Team Experience (VORTeX), a multi-user virtual reality (VR) platform that integrates immersive team simulation with large language model (LLM) analytics to train and evaluate communication, decision-making, teamwork, and leadership. Team dialogue is analyzed using structured prompts derived from the Non-Technical Skills for Surgeons (NOTSS) framework, enabling automated classification of behaviors and generation of directed interaction graphs that quantify communication structure and hierarchy. Two laparoscopic emergency scenarios, pneumothorax and intra-abdominal bleeding, were implemented to elicit realistic stress and collaboration. Twelve surgical professionals completed pilot sessions at the 2024 SAGES conference, rating VORTeX as intuitive, immersive, and valuable for developing teamwork and communication. The LLM consistently produced interpretable communication networks reflecting expected operative hierarchies, with surgeons as central integrators, nurses as initiators, and anesthesiologists as balanced intermediaries. By integrating immersive VR with LLM-driven behavioral analytics, VORTeX provides a scalable, privacy-compliant framework for objective assessment and automated, data-informed debriefing across distributed training environments.

Effective teamwork and communication are critical to surgical safety, yet structured training for non-technical skills (NTS) remains limited compared with technical simulation. The American College of Surgeons (ACS)/Association of Program Directors in Surgery (APDS) Phase III Team-Based Skills Curriculum explicitly calls for scalable tools that both teach and objectively assess these competencies during laparoscopic emergencies¹. To address this need, we introduce the Virtual Operating Room Team Experience (VORTeX), a multi-user virtual reality (VR) platform that integrates immersive team simulation with large language model (LLM)

analytics to train and evaluate communication, decision-making, teamwork, and leadership. By using consumer-grade VR hardware and an LLM-driven feedback engine, VORTeX aims to overcome the logistical and resource barriers of traditional facility-based simulation while providing objective, data-informed debriefing.

Surgical teams represent complex socio-technical systems composed of interdependent professionals, surgeons, anesthesiologists, and nurses, who must coordinate seamlessly under variable clinical, temporal, and cognitive demands². Elevated stress during these crises may impair performance and

¹School of Computer Science, University of Oklahoma, Norman, OK, USA. ²Department of Industrial Engineering, Clemson University, Clemson, SC, USA. ³Beth Israel Deaconess Medical Center, Boston, MA, USA. ⁴Department of Anesthesiology, Northwell, New York City, NY, USA. ⁵Department of Anesthesiology, Henry Ford Health, Detroit, MI, USA. ⁶Department of Surgery, Rutgers New Jersey Medical School, Newark, NJ, USA. ⁷College of Engineering, Florida Agricultural and Mechanical University and Florida State University, Tallahassee, FL, USA. ✉e-mail: doga@ou.edu

jeopardize patient outcomes³, particularly when compounded by breakdowns in teamwork and communication⁴. While simulation-based training has substantially advanced technical skill acquisition, structured training for NTS remains limited, and few simulators replicate the acute cognitive load and stress inherent to intraoperative emergencies. Emerging evidence suggests that incorporating realistic stressors into simulations enhances transfer of training to the intraoperative environment⁵.

Laparoscopic cholecystectomy (LC) serves as one of the procedural foundations for the ACS/APDS Phase III curriculum¹ and provides a clinically relevant context for practicing NTS during emergencies. The validation study by Powers et al.⁶ demonstrated that structured team training for LC emergency scenarios significantly improved communication and error mitigation, yet widespread implementation of such training remains challenging. Traditional facility-based simulation that employs high-fidelity mannequins and mock operating rooms has proven effective for team training⁷, but requires co-located teams, dedicated simulation staff, and substantial time commitments, creating significant logistical and cost barriers⁸. As a result, adoption of the ACS/APDS Phase III curriculum has been limited, leaving most surgical trainees without structured opportunities to practice non-technical skills in high-fidelity crisis scenarios.

Virtual reality offers a geographically flexible alternative that can overcome these barriers while preserving psychological fidelity. Although VR has proven effective for individual laparoscopic psychomotor training^{9,10}, its application to team-based non-technical skills remains nascent. Recent work demonstrates that multi-user VR environments can replicate the spatial and communication demands of clinical settings^{11,12}, enabling distributed teams to train together without physical colocation. Consumer-grade VR systems further reduce cost barriers, making frequent, team-based practice feasible. Existing VR training systems have largely focused on procedural skills and often depend on expensive hardware or physical simulation centers. In contrast, VORTeX leverages consumer-grade, multi-user VR to target the behavioral and cognitive dimensions of surgical teamwork. The platform's novel integration of immersive crisis simulation with a large language model (LLM)-based analytics engine enables automated, objective assessment of communication, leadership, and decision-making behaviors. To our knowledge, this is among the first applications of LLM-driven behavioral analysis to surgical team training, transforming unstructured dialogue into structured performance metrics.

The primary aim of this study is to evaluate the feasibility and perceived educational value of VORTeX for training and assessing non-technical skills within the ACS/APDS Phase III curriculum framework. The secondary aim is to validate the LLM-based analysis pipeline's capability to generate communication interaction networks that reflect expected role-dependent hierarchies during simulated emergencies. We hypothesize that (1) participants will perceive VORTeX as an effective and usable tool for enhancing teamwork under stress, and (2) LLM-generated interaction graphs will quantitatively mirror known intraoperative communication structures (surgeons as central integrators, nurses as initiators, anesthesiologists as intermediaries), supporting their use for objective assessment.

The following sections describe the design, implementation, and pilot validation of VORTeX. Results section presents results from a pilot study conducted at the 2024 Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) annual meeting, including usability metrics, participant perceptions, and analysis of LLM-generated communication graphs. Discussion section discusses implications for competency-based surgical education under the Phase III framework and limitations of this study. Methods section details the platform architecture, professional roles, emergency scenarios, and LLM-based feedback methodology.

Results

The results address the two central objectives of this study: evaluating the usability, feasibility, and perceived educational value of the VORTeX multi-user virtual reality simulator and examining the validity of the LLM-based analytics pipeline for objectively characterizing team communication and leadership behaviors during simulated operative crises. All twelve

participants completed both the simulation and questionnaires with no missing data. Findings related to the usability and educational value of the simulator are presented first, including analysis of post-questionnaire responses, relationships between participant characteristics and performance metrics, and system performance outcomes. The subsequent section focuses on the LLM-derived communication networks, providing evidence of the model's ability to capture authentic team dynamics and leadership patterns during the simulations.

User experience and perceived educational value

The post-questionnaire results suggest that participants generally found the simulator easy to use and perceived value in it, particularly in improving non-technical skills. The highest-rated item (Q11: $\bar{x} = 4.42$, $\sigma = 0.67$) reflected the belief that continued use of the simulation would help enhance teamwork, communications, and decision-making. Participants also indicated that they did not need to learn a lot before getting started (Q10: $\bar{x} = 2.58$, $\sigma = 1.16$) and did not find the system cumbersome (Q2: $\bar{x} = 2.50$, $\sigma = 0.90$). However, confidence in using the system was relatively low (Q9: $\bar{x} = 2.67$, $\sigma = 1.56$), pointing to a wide range of comfort levels among users and suggesting that the technology may be somewhat intimidating. Interestingly, only one-third of participants reported regular VR use, which could have contributed to this lower confidence. Nonetheless, this added challenge may have benefited the study's aim of simulating stressful conditions to foster communication and collaboration.

Despite some hesitancy in operating the system, participants' responses were otherwise positive. Many felt that the system was not unnecessarily complex (Q2: $\bar{x} = 2.08$, $\sigma = 0.90$) and believed its various functions were well-integrated (Q5: $\bar{x} = 3.67$, $\sigma = 1.15$). There was also a strong indication that most people would learn to use the system quickly (Q7: $\bar{x} = 3.75$, $\sigma = 0.97$). Team performance during the laparoscopic emergency scenario received a high average rating of 4.08 (Q13: $\bar{x} = 4.08$, $\sigma = 0.90$), suggesting that the system supported effective collaboration under pressure. Overall, while some variability existed in individual comfort and confidence levels, highlighted by higher standard deviations in those areas, the results demonstrate that participants found the simulator valuable, especially in cultivating non-technical skills.

The full distribution of post-questionnaire responses is shown in Fig. 1. Median values above the midscale (2.5) for the odd-numbered items (and Q12), and below the midscale for the even-numbered items, indicate generally favorable usability and educational value.

Relationships between user characteristics and perceived usability

Given the small sample size ($n = 12$) and the ordinal structure of Likert-scale responses, Spearman rank correlation coefficients (ρ) were computed to examine monotonic associations among demographic variables, perceived usability, and self-reported non-technical skills confidence. All correlations are interpreted descriptively due to the exploratory nature of this pilot study.

Years in clinical practice showed a strong monotonic inverse association with perceived ease of use ($\rho = -1.00$) and a strong positive association with perceived need for technical support ($\rho = 1.00$) and perceived need for additional learning prior to effective use ($\rho = 1.00$). These results suggest that more experienced clinicians tended to report lower perceived usability and greater need for assistance when interacting with the VR system.

Years in practice was also negatively associated with self-reported team performance during the simulated emergency ($\rho = -0.87$), indicating that more experienced clinicians tended to rate team performance more critically. Similarly, age demonstrated a negative monotonic association with post-simulation confidence in non-technical skills ($\rho = -0.87$), suggesting that older participants reported more modest perceived gains.

Positive associations were observed between perceived ease of use and confidence using the system ($\rho = 0.87$), and between confidence using the system and confidence in non-technical skills ($\rho = 0.87$). Conversely, perceived system complexity was negatively associated with confidence using

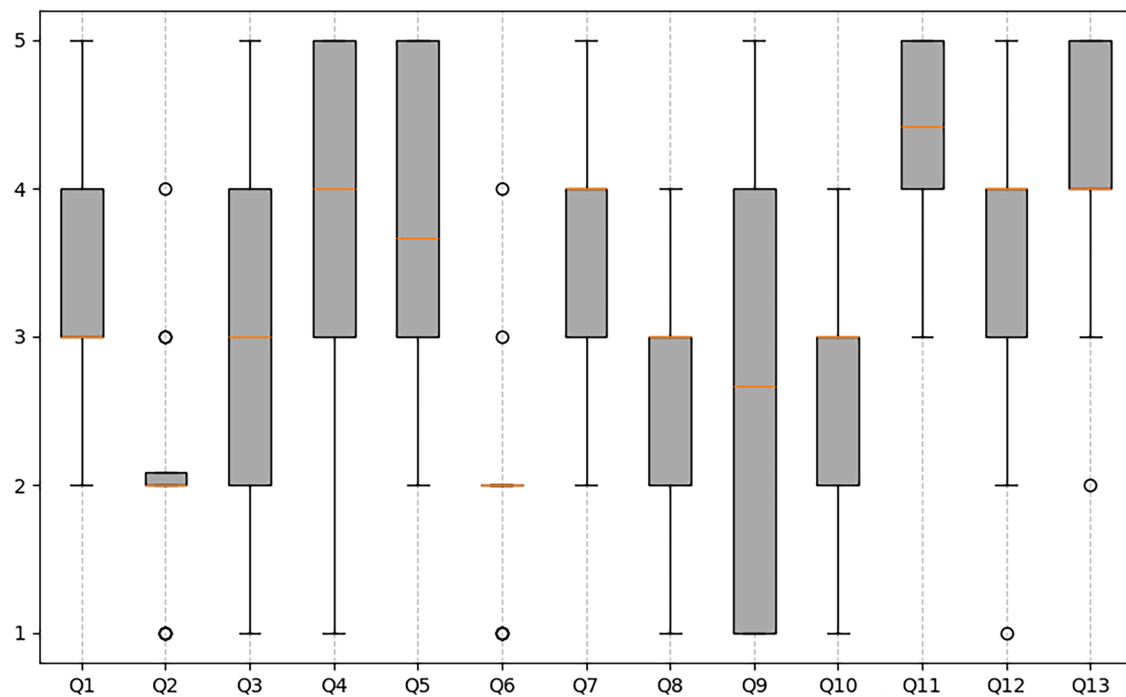


Fig. 1 | Distribution of post-questionnaire responses. Box-and-whisker plots summarize participant responses ($n = 12$) to each questionnaire item (Q1–Q13).

the system ($\rho = -0.87$). Together, these patterns suggest that participants who found the system intuitive also reported greater comfort operating it and greater perceived educational benefit.

System performance and communication fidelity

Consistent system performance and low communication latency are critical for maintaining immersion and the realism of team interaction in multi-user VR environments. Across all participant sessions, the simulator achieved an average frame rate of 73 fps, which maintained the 60 fps threshold for mitigating incidents of cybersickness¹³. The high refresh rate likely contributed to overall comfort and the favorable usability ratings (Q1: “I would like to use this system frequently,” $M = 3.5$).

Equally important for preserving immersion is a low and stable latency between the server and clients. Higher latencies can result in interactions taking noticeable time to complete and more importantly, gestures and other nonverbal communication may not synchronize with spoken words. The average round-trip latency between clients and the central server was 75 ms, with a range from 20 ms during low network load to 174 ms under heavier traffic. These values fall within accepted parameters for real-time multi-user VR systems¹⁴ and did not produce noticeable delays in motion or speech synchronization.

LLM-based analysis of team communication dynamics

The LLM analysis produced valid and interpretable results across all sessions, successfully translating voice transcripts into structured communication networks. Each graph conformed to the requested schema, with nodes representing the surgeon, anesthesiologist, and nurse, and directed edges denoting initiator–receiver exchanges.

Figure 2 presents a representative LLM-generated interaction network. The directed edges capture not only who communicates with whom, but also the *function* of each interaction - situational awareness, communication and teamwork, leadership, or decision making - along with precise temporal context. In this case, the network reveals a dense, bidirectional flow of information that evolves with the clinical situation. Early in the sequence, the surgeon initiates a timeout (“Communication and Teamwork, 1:34:13”), which is immediately reinforced by the nurse’s confirmation of procedure and site, reflecting shared responsibility rather

than unilateral command. As the case progresses, situational awareness propagates laterally across roles: the nurse discloses a critical comorbidity (“1:35:38”), the surgeon requests and integrates status updates (“1:37:09”), and the anesthesiologist both receives and acts on this information through blood product ordering and escalation. Leadership behaviors are distributed rather than centralized, with both the surgeon and nurse independently confirming that help is en route, illustrating redundancy and resilience under stress. Overall, the structure of the network reflects anticipatory coordination, role-appropriate initiative, and timely escalation.

Across all sessions, the LLM consistently captured the directionality of team interactions, distinguishing between initiating and receiving behaviors during the simulated intraoperative events. To interpret the resulting graphs, we characterized each participant’s communication behavior using three standard network metrics. The **out-degree** reflects the number of distinct communication events initiated by a participant, indicating a more directive or leadership-oriented role. The **in-degree** represents the number of communications received, denoting information integration or confirmatory behavior. The **clustering coefficient** measures how interconnected a participant’s communication partners were, thereby reflecting the cohesion and redundancy of local information exchange. Together, these features capture the balance between leadership, coordination, and information flow within the team.

As detailed, we calculated in-degree, out-degree, and clustering coefficients for each role. The aggregated results are presented in Table 1. The surgeon node had the highest mean in-degree (3.84), indicating that the surgeon was the most frequent recipient of information and confirmations from other team members, consistent with the surgeon’s role as the primary decision-maker in operative settings. On the other hand, the nurse node exhibited the highest mean out-degree (3.62), reflecting a tendency to initiate communication regarding patient status and operating-room coordination. The anesthesiologist had a more balanced interaction pattern (3.46 out-degree, 2.62 in-degree), serving as both an initiator and recipient of critical exchanges. The surgeon also showed the highest clustering coefficient (0.92), indicating participation in the most tightly interconnected communication subnetwork within the team.

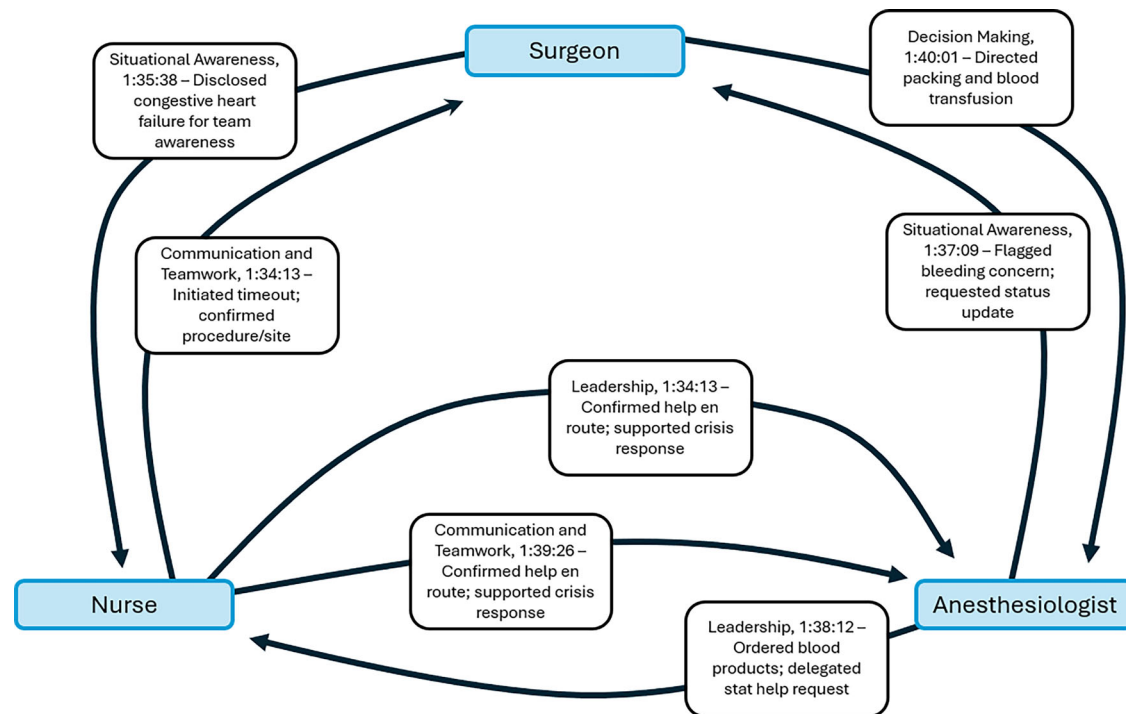


Fig. 2 | Example LLM-generated communication interaction network during a simulated intraoperative crisis. Nodes represent team roles (surgeon, anesthesiologist, nurse), and directed edges indicate communication events from initiator to receiver. Edge attributes show behavioral category, timestamp, and rationale.

Table 1 | Mean $\pm$ standard deviation of in-degree, out-degree, and clustering coefficients for each team role in the LLM-generated interaction graphs

	In-Degree (Mean, SD)	Out-Degree (Mean, SD)	Clustering Coefficient (Mean, SD)
Surgeon	3.84 $\pm$ 1.77	2.23 $\pm$ 1.30	0.92 $\pm$ 0.19
Anesthesiologist	2.61 $\pm$ 1.45	3.46 $\pm$ 1.13	0.85 $\pm$ 0.23
Nurse	2.77 $\pm$ 1.30	3.62 $\pm$ 1.56	0.88 $\pm$ 0.24

Discussion

This work presents an integrated framework that combines immersive VR with LLM analytics to support training and evaluation of team-based NTS in the operating room. The results from the SAGES 2024 pilot study establish technical feasibility, educational plausibility, and analytic potential of this approach. Specifically, the findings demonstrate that (i) a multi-user, networked VR environment can elicit authentic intraoperative team interactions under simulated stress; and (ii) a locally hosted LLM can extract and structure these interactions into interpretable communication networks aligned with recognized NTS frameworks. Together, these components illustrate how immersive simulation and AI-driven analysis can converge to form a scalable, data-centric paradigm for NTS education.

The ACS/APDS Phase III curriculum emphasizes laparoscopic emergencies such as bleeding (occurring in approximately 10% of LC cases¹⁵) and pneumothorax (0.01–0.4% incidence¹⁶) as critical training scenarios. However, a 2013 survey of general surgery program directors found that only 16% implemented Phase III, citing cost, coordinator availability, and limited personnel as primary barriers⁷. VORTeX addresses these challenges by providing a scalable, VR-based alternative that can be deployed across distributed settings without the need for physical simulation centers.

VORTeX bridges two traditionally separate domains: high-fidelity simulation and automated behavioral analytics. The VR component

provides experiential realism and psychological engagement, conditions known to enhance learning transfer, while the LLM introduces scalable, post-hoc interpretation of team communication. Unlike prior VR simulators focused primarily on psychomotor training or scripted role-play, VORTeX enables real-time, geographically distributed collaboration through consumer-grade hardware. Its integration with an LLM-based analytic backend transforms the rich but unstructured communication data into structured, machine-interpretable graphs that quantify leadership, coordination, and communication flow. This fusion demonstrates a credible pathway toward AI-augmented debriefing, addressing long-standing barriers in simulation-based education: limited facilitator bandwidth and the subjectivity of human observation.

Participants perceived the system as a valuable tool for developing teamwork, communication, and decision-making skills (Q11, $\bar{x}$ = 4.42). The VR environment replicated the stress, and coordination demands of real operative crises (e.g., pneumothorax, intra-abdominal bleeding) within a psychologically safe setting. Meanwhile, the LLM-derived communication graphs captured role-specific patterns consistent with recognized intraoperative hierarchies. Surgeons as central information integrators, nurses as coordinators and initiators, and anesthesiologists as balanced intermediaries. This correspondence provides preliminary evidence of construct coherence, suggesting that the automated analytic pipeline can extract recognizable team behaviors without manual annotation. However, the evidence should be viewed as early and exploratory; without expert-coded benchmarks or inter-run reliability data, definitive claims of validity cannot yet be made.

Beyond the behavioral insights, the integration of VR and LLMs represents a methodological advance. The VR component supports distributed, synchronous simulation using accessible, consumer-grade hardware, while the analytic backend produces machine-readable summaries that could standardize debriefing and reduce facilitator dependence. By maintaining all processing locally through a privacy-compliant, open-source model (DeepSeek-R1-0528), the framework meets data governance requirements critical for educational and clinical settings. The resulting directed communication graphs introduce a reproducible quantitative language for team behavior-transforming qualitative debriefing into a

structured, scalable process that can support longitudinal tracking and adaptive scenario design.

Several limitations warrant discussion. The small sample size ($n = 12$) and single-site, event-based recruitment limit statistical inference and generalizability, and the findings should be interpreted as feasibility and construct-coherence evidence rather than confirmatory validation. As detailed in the results section, multiple non-perfect associations demonstrated strong effect sizes ($|\rho| \approx 0.87$), including years in practice versus team performance ($\rho = -0.87$), age versus confidence in non-technical skills ($\rho = -0.87$), and ease of use versus confidence using the system ($\rho = 0.87$). However, none reached statistical significance, likely reflecting the limited sample size ($n = 12$). Accordingly, these findings should be interpreted as descriptive monotonic trends rather than inferential evidence. Larger, multi-institutional studies will be required to determine whether these patterns persist across broader clinician populations. Participants were drawn from a self-selected group of surgical professionals at a national conference, individuals who are likely more technologically receptive and possibly more experienced in simulation than the general clinical population. The simulation time per participant was relatively short (10 minutes), constraining behavioral diversity and limiting the observation of adaptive team processes such as leadership transitions or strategy evolution over time. Additionally, six of the twelve participants were PGY-1 residents, resulting in a relatively junior and homogeneous cohort. While this limits direct generalization to fully established multidisciplinary surgical teams, the primary objective of this pilot was to evaluate platform feasibility, usability, and analytic functionality rather than to measure expert-level team performance. The communication structures analyzed by the LLM are role-based rather than expertise-based, and thus the underlying analytic framework is not inherently dependent on participant seniority.

In this pilot study only the surgeon role was played by study participants, while the nurse and anesthesiologist roles were performed by experienced system users. This configuration was intentionally selected to maintain scenario stability during live conference deployment and to stress-test the platform's capacity for geographically distributed, multi-site collaboration. As a result, the observed interaction networks reflect structured surgeon-team coupling within a controlled role scaffold rather than fully emergent dynamics of intact clinical teams with shared institutional culture and baseline familiarity. Accordingly, conclusions regarding team behavior, leadership distribution, or relative team functionality should be interpreted as preliminary indicators of analytic coherence rather than definitive measurements of real-world team effectiveness. The analytic pipeline relies on automatic speech recognition (ASR) and speaker diarization, both of which can introduce transcription and segmentation errors. Misattribution of utterances or timing drift may affect the accuracy of edge assignments in the communication graph. Additionally, the mapping of dialogue segments to Non-Technical Skills for Surgeons (NOTSS) categories remains heuristic and context-dependent; the LLM's interpretation of leadership, decision-making, or teamwork behaviors has not yet been benchmarked against expert human raters. Without such cross-validation, the current outputs represent plausible patterns rather than verified classifications.

Although the LLM-generated behavioral classifications showed structural plausibility and alignment with expected role-based communication patterns, formal analysis with expert NOTSS raters was beyond the scope of this pilot feasibility study. An expert-LLM agreement analysis with an inter-rater reliability assessment will be conducted to establish criterion validity.

Regarding the LLM, while inference parameters were held constant across sessions, the use of a nonzero temperature introduces potential sampling variability that we did not quantify in this pilot. We assessed LLM output from single-pass analysis only. We prioritized evaluating the model's capability to produce interpretable structures from single instances rather than measuring inter-run stability. Also, while this feasibility study utilized a single locally hosted LLM to ensure data privacy and consistent prompt engineering, we recognize that single-model dependency introduces

potential biases specific to the model due to differences in training sets or architecture.

Hardware tradeoffs also affected ecological fidelity. The absence of facial expression capture, eye-tracking, and fine-grained hand gestures limits the range of non-verbal cues available to the model, potentially underrepresenting affective and leadership behaviors. While these omissions were deliberate to preserve portability and cost-efficiency, they reduce the multimodal realism necessary for comprehensive NTS evaluation.

Importantly, this pilot was not intended as a full implementation of the ACS/APDS Phase III Team-Based Skills Curriculum but as an early demonstration of how immersive, AI-driven analytics could support its objectives. Future studies should evaluate the platform within the Phase III framework, with larger, institutionally diverse cohorts, extended scenarios, and expert-LLM concordance analysis to establish criterion validity and educational impact.

Despite these constraints, the study establishes a credible foundation for expanding the integration of immersive simulation and AI-based analytics in surgical education. Future work should include multi-institutional cohorts, extended scenarios, and human-LLM concordance analysis to quantify reliability and validity. Linking network-derived features to objective performance outcomes, such as expert NTS ratings or error mitigation, will be critical to establish criterion validity. Incorporating multimodal signals (gaze, physiology, and high-fidelity gestures) could further refine behavioral inference. Ultimately, the combination of VR-based experiential learning and LLM-driven analysis offers a scalable, privacy-conscious, and data-rich framework for the objective assessment and development of non-technical skills in operative teams.

Methods

System overview

VORTEX is a multi-user VR platform designed to train and objectively assess non-technical skills such as communication, teamwork, and decision-making in surgical teams. The platform enables participants to collaborate in shared operating room scenarios through networked VR environments, allowing geographically distributed teams to engage in realistic simulations of both routine and emergent intraoperative events.

As illustrated in Fig. 3, the system comprises three core components: A head-mounted consumer-grade VR system to gather user input, the simulation system, and the feedback engine. The simulation system integrates the headset, hand controllers, and voice communication from participants with a network synchronization layer that ensures real-time multi-user interactions. A physiology engine governs patient dynamics and event triggers, while an interaction module interprets controller inputs to coordinate interactions with simulation elements between clients. All verbal exchanges and simulation interactions are recorded and time stamped.

Captured data are processed by the feedback engine, which employs an LLM guided by structured prompts derived from the NOTSS framework¹⁷. The LLM analyzes communication transcripts in conjunction with interaction timestamps to identify key behavioral patterns and generate directed interaction graphs that quantify team structure and information flow.

Unlike traditional simulators that rely on co-located participation and costly facilitators, VORTEX emphasizes accessibility and ease of deployment. By shifting focus from procedural fidelity to behavioral coordination, it leverages affordable consumer VR hardware to provide an immersive, scalable platform for data-driven NTS training. Figure 4 illustrates the virtual operating room environment used in the simulations, featuring a fully equipped laparoscopic setup with adjustable lighting, surgical displays, and interactable instruments. This environment reproduces the spatial layout, visual cues, and workflow of a real operating suite, promoting psychological fidelity and natural team communication while allowing distributed participants to engage synchronously from any networked location.

Simulation design

The VORTEX simulation environment was designed to recreate the spatial layout, visual context, and cognitive demands of a real operating room while

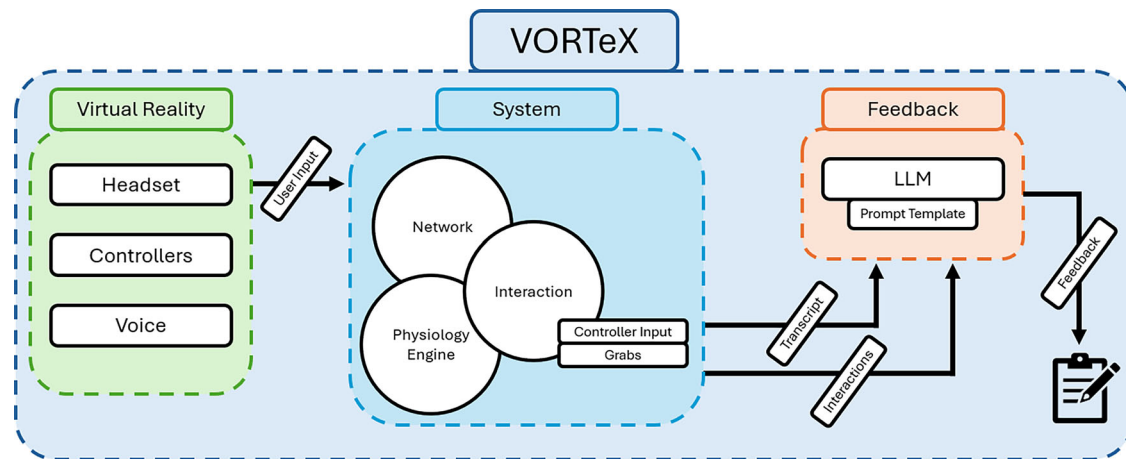


Fig. 3 | System architecture of the VORTeX platform integrating VR simulation and LLM-based analytics. The framework consists of three main components: (i) the VR input layer, (ii) the simulation system, and (iii) the feedback engine. Arrows

indicate data flow between components, including real-time interaction capture and post-session analysis.

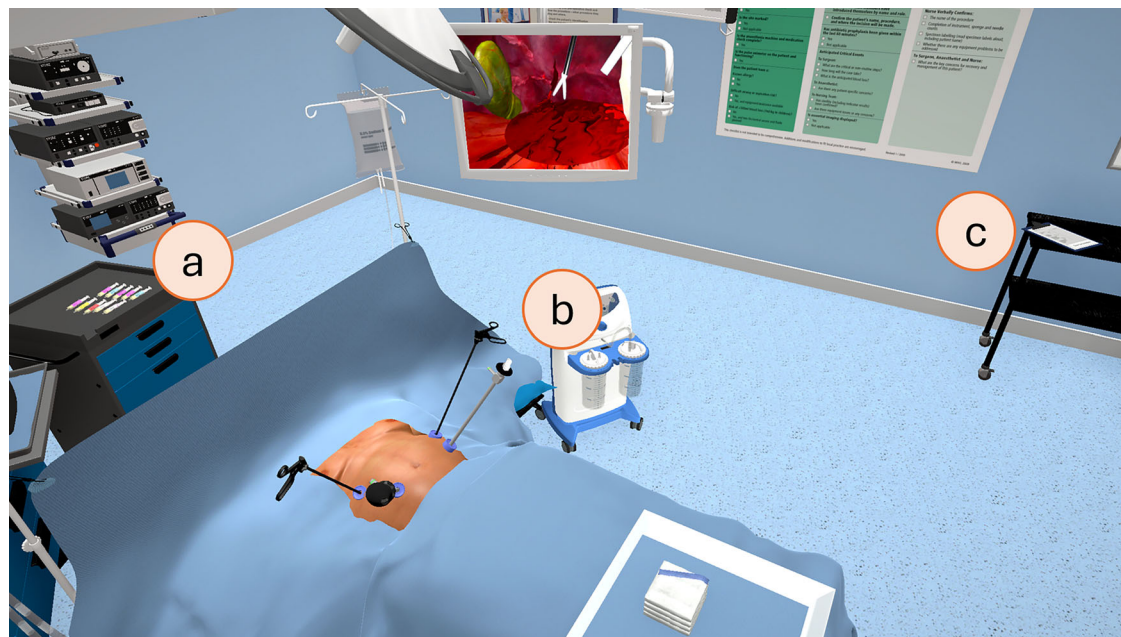


Fig. 4 | Virtual operating room environment for team-based simulation of laparoscopic cholecystectomy. **a** Anesthesiologist's station with drug cart and medication interface. **b** Surgeon's station with laparoscopic instruments and

adjustable display monitors. **c** Nurse's station with control interface for simulation flow, including timeout initiation and resource requests.

emphasizing the behavioral and communication aspects of team performance. The design goal was to elicit authentic team interactions under conditions of time pressure, uncertainty, and interdependence—key elements of non-technical skill expression. Each scenario was developed in consultation with surgeons, anesthesiologists, and perioperative nurses to ensure clinical realism, appropriate task sequencing, and accurate role responsibilities.

Two emergency scenarios were embedded within an LC procedure, reflecting the priorities of the ACS/APDS Phase III National Skills Curriculum. The **pneumothorax scenario**, as seen in Fig. 5, manifests as sudden hypoxia requiring rapid cross-disciplinary recognition and communication to diagnose and manage. The **intra-abdominal bleeding scenario**, as seen in Fig. 6, presents as progressive hemorrhage with escalating vital sign deterioration, eventually requiring the team to decide on conversion to open surgery. Multi-user intra-abdominal bleeding emergency simulation can be seen in the supplementary video. Both scenarios were developed with

clinical experts to ensure realistic progression: Bleeding evolves through discrete severity stages revealed by laparoscopic exploration, while pneumothorax demands immediate coordination to stabilize ventilation. These scenarios introduce contrasting temporal pressures which elicit distinct patterns of decision making and communication under stress.

The virtual operating room replicates the real-life counterpart with adjustable lighting, movable monitors, and spatial arrangements promoting natural team communication (Fig. 4). Stations within the virtual operating room have role-specific interfaces and interactables that are also designed with these principles of realism and immersion. The anesthesiologist administers pharmacologically accurate medications through an IV line that dynamically alters patient vitals (Fig. 4—marked as “a”), the surgeon manipulates laparoscopic instruments via simplified tool interactions (Fig. 4—marked as “b”), and the nurse controls simulation flow via a clipboard interface, initiating the timeout, requesting equipment (blood, argon, major kit), and calling for help (Fig. 4—marked as “c”). All roles wear appropriate



Fig. 5 | Virtual operating room during pneumothorax emergency scenario. The scene illustrates the onset of pneumothorax, including changes in patient status indicators and team positioning.



Fig. 6 | Virtual operating room during intra-abdominal bleeding emergency scenario. The scene shows progressive hemorrhage with corresponding deterioration in patient vital signs.

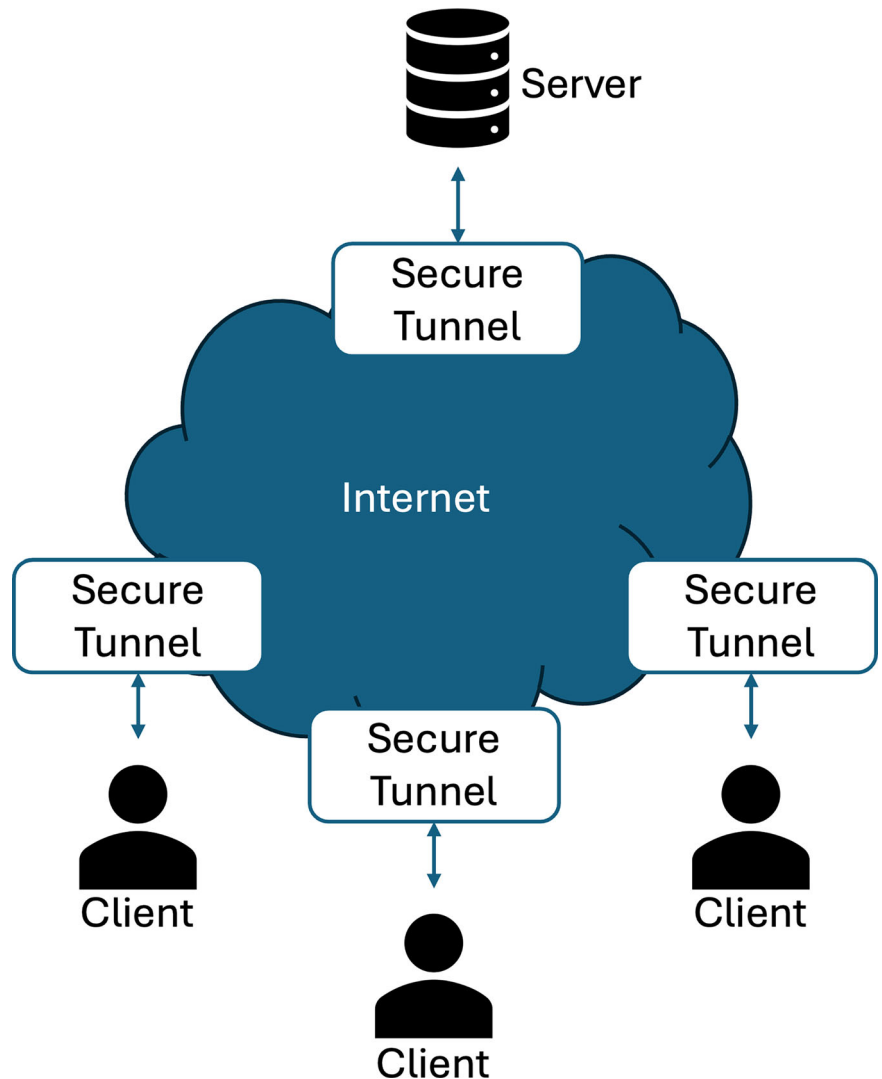
PPE, and most environmental objects are interactable to preserve immersion and provide contextual affordances for non-verbal communication.

Patient vital signs are driven by a physiology engine augmented to respond to trauma and drug onsets, ensuring that hemodynamic and respiratory responses authentically reflect clinical actions and temporal progression. Numerical values and waveform displays update in real-time based on drug administration, blood loss, and procedural interventions, creating realistic triggers for team communication. Event sequencing (e.g., bleeding severity escalation, pneumothorax onset) is pre-scripted but responsive to team performance, prompting adaptive decision-making without breaking clinical plausibility or immersion.

Performance optimization prioritized stable network synchronization and visual continuity across participants that may be

geographically distributed. This was achieved through asset optimization until the 24-fps threshold for VR continuity was exceeded. Adopting a hybrid Transmission Control Protocol (TCP)/ User Datagram Protocol (UDP) network stack ensured critical state data such as patient vitals and simulation events were reliably synchronized without overburdening the network. Interaction design focused on high-frequency tracking of head movements, hand positions, and gaze vectors to support gestural communication (pointing, nodding) without requiring costly facial or finger tracking; this deliberate limitation preserved the portability of consumer-grade hardware while making most objects interactable, as participants consistently attempted environmental manipulation upon entry and non-responsiveness broke immersion.

Fig. 7 | Client–server communication architecture supporting distributed multi-user VR simulation. Each participant node connects to a central server via encrypted VPN tunnels. Solid lines with arrows represent secure bidirectional communication channels (IPSec) and synchronization of simulation state and user interactions.



Network architecture

The multi-user aspect of VORTEX is implemented using a client-server architecture (Fig. 7) that supports distributed collaboration while maintaining secure, low-latency communication. The server functions as an independent coordination node responsible for synchronizing simulation states and managing bidirectional data exchange among connected clients. To emulate a local-area network and ensure data security, a virtual private network (VPN) was used to create encrypted tunnels between clients and the central server. This configuration enables geographically dispersed users to participate in shared training sessions while preserving data integrity and minimizing latency.

The separation of client and server processes provides greater deployment flexibility, enabling the server to run on a local host or a remote machine, depending on resource availability. Separation of client and server also allows each side of the application to focus on data collection relevant to its side of the network. Ensuring low network latency is crucial for maintaining the realism of team communications in multi-user environments¹⁸. To achieve this, VORTEX employs a hybrid transport strategy, combining the TCP for reliable synchronization of critical information (e.g., simulation state, patient vitals, and avatar identifiers) with the UDP for high-frequency updates of avatar movements and gestures. This design prioritizes essential state synchronization while allowing lightweight, low-latency streaming of motion data, thereby ensuring smooth performance under variable network conditions.

Large language model

During each VR training session, client systems continuously record multiple digital streams: Participant interactions with virtual objects, gaze vectors, and audio transcriptions of spoken dialogue. At the end of a session, these data streams are transmitted to a central server, where they are time-synchronized, anonymized and merged into a unified transcript annotated with participant identifiers and event timestamps.

To evaluate team performance within the simulated surgical environment, this consolidated transcript is analyzed by an LLM. Behavioral performance is operationalized in terms of behavioral indicators defined by the NOTSS framework¹⁷, which categorizes observable behaviors into four domains: situational awareness, decision making, communication and teamwork, and leadership. The objective of the analysis is to detect instances of these non-technical skills, attribute them to the relevant participants, and generate a directed interaction graph representing communication flow and team dynamics.

Due to privacy and data governance concerns, the analysis was performed using a locally hosted, open-source reasoning model rather than a cloud-based API. For this work, DeepSeek-R1-0528¹⁹ was chosen due to its strong performance on reasoning and human-preference benchmarks and its support for detailed control over prompt structure and inference parameters. This reasoning model performs relatively well compared to other open-source and closed-source models on a number of human preference and reasoning benchmarks.

The model receives as input a composite prompt consisting of three components: (i) a prefix, which demarcates the beginning of the transcript; (ii) the transcript itself; and (iii) a postfix containing task instructions and the NOTSS behavioral rubric. The postfix instructs the model to (1) identify demonstrations of NOTSS-aligned behaviors by any participant, (2) self-verify and refine the extracted results, and (3) construct a directional interaction network in JSON format. In this network, each node corresponds to a participant, and each edge encodes an observed interaction, where the source node denotes the individual demonstrating a NOTSS skill, the target node represents the recipient of that behavior, and the edge attributes include the skill category, a timestamp, and a brief descriptive rationale. Full prompt wording and the JSON output schema are provided in Table S1 of Supplementary Materials to ensure methodological transparency and reproducibility.

For this feasibility study, each transcript underwent a single analysis pass to assess the model's capability to generate interpretable interaction networks. Inference parameters were configured to match those recommended by the model manufacturer: temperature = 0.6, and top-p = 0.95 to balance reasoning exploration with coherence. While these settings permit sampling variability, we maintained fixed random seeds and identical prompt structures across sessions to maximize reproducibility. The resulting JSON output provides a machine-readable representation of inter-participant behavioral dynamics that can be visualized or quantitatively analyzed to assess team performance.

We implement three safeguards to reduce model error and improve auditability: (i) prompt-regularized outputs (strict JSON schema with required fields), (ii) temporal consistency checks (edge timestamps must map to existing transcript spans), and (iii) internal coherence checks (role labels must match session metadata). This pilot study focused on evaluating the feasibility and structural coherence of automated communication network generation rather than formal criterion validation of behavioral classification accuracy.

To interpret the resulting graphs, we characterized each participant's communication behavior using three standard network metrics. The out-degree reflects the number of distinct communication events initiated by a participant, indicating a more directive or leadership-oriented role. The in-degree represents the number of communications received, denoting information integration or confirmatory behavior. The clustering coefficient measures how interconnected a participant's communication partners were, thereby reflecting the cohesion and redundancy of local information exchange. Together, these features capture the balance between leadership, coordination, and information flow within the team.

User design study

An Institutional Review Board–approved study was conducted during the 2024 Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) conference to evaluate the usability and perceived educational value of the VORTEX simulator. Each participant completed a short pre-questionnaire to collect demographic and experiential data, followed by a ten-minute simulation session and a post-questionnaire assessing system usability and NTS perception. Informed consent was obtained from all participants in this study, and all human participants research was performed in accordance with the Declaration of Helsinki. This project was approved by Beth Israel Deaconess Medical Center IRB, IRB # 2024P000306.

Twelve surgical professionals participated (eight male, four female; ages 20–59 years). Nine were surgical residents (six PGY-1, two PGY-3, one PGY-4) and three were attending surgeons. Five participants had prior experience with team-based medical simulation training, and eight had previous exposure to virtual-reality systems.

Each participant operated as the surgeon within the simulation. Two additional medical professionals, experienced users of the system, remotely played the nurse and anesthesiologist roles from Boston, MA, while the simulation server was hosted in Cleveland, OH. This configuration demonstrated the system's ability to support geographically distributed multi-user participation.

The post-questionnaire (Tables S2 and S3 in Supplementary Materials) consisted of 13 items. The first ten items followed the System Usability Scale (SUS)²⁰, assessing ease of use, learnability, and integration. The final three items measured perceived improvement in teamwork, communication, and decision-making based on the NOTSS framework¹⁷. Participants rated all items on a five-point Likert scale (1 = strongly disagree to 5 = strongly agree).

For interpretability, mean scores above 2.5 were considered directionally positive for odd-numbered questions and mean scored below 2.5 were considered directionally positive for even-numbered questions. Although SUS yields a single composite score, individual item responses were analyzed descriptively to explore usability trends and their correlation with self-reported confidence in non-technical skills (Q12). Correlation analysis referenced specific question numbers (Q#) to maintain reproducibility.

Statistical analysis

Descriptive statistics were calculated for all questionnaire items to summarize usability and perceived NTS outcomes. Item responses were reported as means $\pm$ standard deviations on a five-point Likert scale. For exploratory analysis, Spearman rank correlation coefficients were computed to examine associations among key variables, including prior clinical experience, age, system usability (SUS items), and self-reported confidence in NTS (Q12). Correlation strength and direction were interpreted in the context of the small sample size ($n = 12$), and results were considered indicative rather than inferential. Statistical computations were performed in Python 3.11 using the *pandas* and *scipy.stats* libraries.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to data privacy concerns but are available from the corresponding author on reasonable request.

Code availability

The underlying code for this study is not publicly available for proprietary reasons.

Received: 22 January 2026; Accepted: 6 March 2026;

Published online: 04 June 2026

References

1. ACS/APDS Surgery Resident Skills Curriculum - Phase 3 | American College of Surgeons | Online Learning. <https://learning.facs.org/content/acsapds-surgery-resident-skills-curriculum-phase-3#group-tabs-node-course-default1>.
2. Moreal, C. et al. The impact of operating room teamwork on patients outcomes: a systematic review. *Curr. Probl. Surg.* **74**, 101946 (2025).
3. Arora, S. et al. The impact of stress on surgical performance: a systematic review of the literature. *Surgery* **147**, 318–330 (2010).
4. Sami, A. et al. Real-time observations of stressful events in the operating room. *Saudi J. Anaesth.* **6**, 136–139 (2012).
5. Prabhu, A., Smith, W., Yurko, Y., Acker, C. & Stefanidis, D. Increased stress levels may explain the incomplete transfer of simulator-acquired skill to the operating room. *Surgery* **147**, 640–645 (2010).
6. Powers, K. A. et al. Simulated laparoscopic operating room crisis: an approach to enhance the surgical team performance. *Surg. Endosc.* **22**, 885–900 (2008).
7. McGaghie, W. C., Issenberg, S. B., Petrusa, E. R. & Scalese, R. J. A critical review of simulation-based medical education research: 2003–2009. *Med. Educ.* **44**, 50–63 (2010).
8. Baker, D. P., Gustafson, S., Beaubien, J. M., Salas, E. & Barach, P. Medical team training programs in health care. *Adv. Patient Saf. Res. Implement.* **4**, 253–267 (2005).
9. Portelli, M., Bianco, S., Bezzina, T. & Abela, J. Virtual reality training compared with apprenticeship training in laparoscopic surgery: a meta-analysis. *Ann. R. Coll. Surg. Engl.* **102**, 672–684 (2020).

10. Linsk, A. M. et al. Validation of the VBLaST pattern cutting task: a learning curve study. *Surg. Endosc.* **32**, 1990–2002 (2018).
11. Creutzfeldt, J., Hedman, L. & Felländer-Tsai, L. Cardiopulmonary resuscitation training by avatars: a qualitative study of medical students' experiences using a multiplayer virtual world. *JMIR Serious Games* **4**, e6448 (2016).
12. Kaufman, M., Dev, P. & Youngblood, P. Application of multiplayer game technology to team based training of medical first responders. In *Interservice/Industry Training, Simulation & Education Conference (IIITSEC)* (2005).
13. Wang, J. et al. Effect of frame rate on user experience, performance, and simulator sickness in virtual reality. *IEEE Trans. Vis. Comput. Graph.* **29**, 2478–2488 (2023).
14. Van Damme, S. et al. Impact of latency on QoE, performance, and collaboration in interactive multi-user virtual reality. *Appl. Sci.* **14**, 2290 (2024).
15. Kaushik, R. Bleeding complications in laparoscopic cholecystectomy: Incidence, mechanisms, prevention and management. *J. Minimal Access Surg.* **6**, 59 (2010).
16. Kumar, G. & Singh, A. Pneumothorax during laparoscopic cholecystectomy. *Med. J. Armed Forces India* **63**, 277–278 (2007).
17. Yule, S. & Paterson-Brown, S. Surgeons' non-technical skills. *Surg. Clin.* **92**, 37–50 (2012).
18. Milef, N. et al. Disruptions to shared mental models from poor quality of service in collaborative virtual environments. *Sci. Rep.* **11**, 23556 (2021).
19. DeepSeek-AI et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Preprint at <https://doi.org/10.48550/arXiv.2501.12948> (2025).
20. Brooke, J. SUS-A quick and dirty usability scale. *Usability Eval. Ind.* **189**, 4–7 (1996).

Acknowledgements

This project was supported by grants from the National Institutes of Health (NIH)/NIBIB R01EB005807, R01EB033674, and R01EB032820.

Author contributions

JB and DD constructed the simulator. JB, DD, CJ, AJ, and RM conducted the user study. JB, DD, and CJ worked on the formal analysis of the results. DH,

SJ, JM, and DJ ensured medical accuracy within the simulator. JB, DD, SD, CJ, and DJ worked on the original draft of the manuscript. Everyone contributed to reviewing and editing the manuscript. SD, DD, CJ, and DJ supervised and contributed to the funding acquisition.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s44484-026-00009-3>.

Correspondence and requests for materials should be addressed to Doga Demirel.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

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/>.

© The Author(s) 2026