

AutoCLC: Towards Automated Assessment and Feedback on Closed-Loop Communication in Team-based Healthcare Simulation Training

KAYLEE YAXUAN LI and KUNPENG HUANG, Computer Science and Engineering, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA

BRYAN HARMER, CASEY DUCHARME, BENJAMIN HEASMAN, and ELEANOR FALAHEE, Medical School, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA

JAMES COOKE, Department of Learning Health Sciences, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA

MICHAEL COLE, Department of Emergency Medicine, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA

ALANSON SAMPLE, Computer Science and Engineering, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA

VITALIY POPOV, Medical School, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA

This study introduces AutoCLC, an AI-powered system designed to assess and provide feedback on Closed-Loop Communication (CLC) in professional learning environments. CLC, where a sender's Call-Out statement is acknowledged by the receiver's Check-Back statement, is a critical safety protocol in high-reliability domains, including emergency medicine resuscitation teams. Existing methods for evaluating CLC lack quantifiable metrics and depend heavily on human observation. AutoCLC addresses these limitations by leveraging natural language processing and large language models to analyze audio recordings from Advanced Cardiovascular Life Support (ACLS) simulation training. The system identifies CLC instances, measures their frequency and rate per minute, and categorizes communications as effective, incomplete, or missed. Technical evaluations demonstrate that AutoCLC achieves 78.9% precision for identifying Call-Outs and 74.3% for Check-Backs, with a performance gap of only 5% compared to human annotations. A user study involving 11 cardiac arrest instructors across three training sites supported the need for automated CLC assessment. Instructors found AutoCLC reports valuable for quantifying CLC frequency

Kunpeng Huang and Kaylee Yaxuan Li contributed equally to this research.

This work was supported by US National Science Foundation (Grant Nos. IIS-2202451, DRL-2506865).

Authors' Contact Information: Kaylee Yaxuan Li (corresponding author), Computer Science and Engineering, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA; e-mail: yaxuanli@umich.edu; Kunpeng Huang, Computer Science and Engineering, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA; e-mail: kunpeng@umich.edu; Bryan Harmer, Medical School, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA; e-mail: bryharme@med.umich.edu; Casey Ducharme, Medical School, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA; e-mail: ducharca@umich.edu; Benjamin Heasman, Medical School, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA; e-mail: bheasman@umich.edu; Eleanor Falahee, Medical School, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA; e-mail: efalahee@umich.edu; James Cooke, Department of Learning Health Sciences, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA; e-mail: cookej@med.umich.edu; Michael Cole, Department of Emergency Medicine, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA; e-mail: mcolemd@umich.edu; Alanson Sample, Computer Science and Engineering, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA; e-mail: apsample@umich.edu; Vitaliy Popov, Medical School, University of Michigan-Ann Arbor, Ann Arbor, Michigan, USA; e-mail: vipopov@umich.edu.



This work is licensed under Creative Commons Attribution-NonCommercial International 4.0.

© 2026 Copyright held by the owner/author(s).

ACM 2637-8051/2026/1-ART2

<https://doi.org/10.1145/3768164>

and quality, as well as for providing actionable, example-based feedback. Participants rated AutoCLC highly, with a System Usability Scale score of 76.4%, reflecting above-average usability. This work represents a significant step toward developing scalable, data-driven feedback systems that enhance individual skills and team performance in high-reliability settings.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing systems and tools**;

Additional Key Words and Phrases: Closed-Loop Communication, Large-Language Model, Team-based Communication, Healthcare Training, Medical Education

ACM Reference format:

Kaylee Yaxuan Li, Kunpeng Huang, Bryan Harmer, Casey Ducharme, Benjamin Heasman, Eleanor Falahee, James Cooke, Michael Cole, Alanson Sample, and Vitaliy Popov. 2026. AutoCLC: Towards Automated Assessment and Feedback on Closed-Loop Communication in Team-based Healthcare Simulation Training. *ACM Trans. Comput. Healthcare* 7, 1, Article 2 (January 2026), 31 pages.

<https://doi.org/10.1145/3768164>

1 Introduction

Closed-Loop Communication (CLC), originating in military and aeronautical contexts, ensures clear and unambiguous information exchange by requiring acknowledgment and confirmation of conveyed verbal messages [4, 17, 81]. This structured communication method is fundamental for accurate and reliable performance in high-stakes environments, including nuclear power generation, air traffic control, and maritime operations [10, 17, 31, 60]. In healthcare, CLC has become a best practice for acute care teams. It is now widely taught through standardized training programs used across surgical, critical care, and outpatient settings to ensure seamless communication during high-acuity situations [71].

In resuscitation settings, where chaotic environments and time-sensitive decisions prevail, effective communication is essential to reduce errors and improve patient outcomes [62, 80]. Teams must collaborate to assess, diagnose, and perform advanced life support interventions under significant cognitive, psycho-emotional, and physical demands [9, 41, 85]. Breakdowns in communication, which are common in these settings, contribute to adverse events and medical errors [34, 43, 50]. To address these challenges, CLC is employed as a critical communication strategy to standardize verbal exchanges as follows:

- Call-Out (CO): The transmitter issues a message, typically addressing the receiver by name.
- Check-Back (CB): The receiver acknowledges and repeats back the message for confirmation.
- Loop Closure: The transmitter confirms the acknowledgment, ensuring the message was understood.

For example, Figure 1(a) depicts an **Advanced Cardiovascular Life Support (ACLS)** scenario where a team leader might initiate a Call-Out: “John, please check the patient’s pulse.” John, the receiver, responds with a Check-Back: “Checking pulse. The patient does not have a pulse.” The leader then closes the loop with: “Got it, no pulse.” Such structured communication reduces ambiguity, ensures clear information exchange, and promotes task efficiency.

Adherence to CLC best practices has been linked to decreased adverse events and is endorsed by national organizations such as the **American Heart Association (AHA)** and the **Agency for Healthcare Research and Quality (AHRQ)** [20, 49]. However, its adoption during patient care remains inconsistent, likely due to limited standardization in training, evaluation challenges, and competing demands during critical moments [6, 38]. While team-based training in healthcare often emphasizes improved communication, including CLC, it typically lacks clear benchmarks for evaluating communication performance. Current methods for assessing CLC in medical training rely heavily on human observations, manual video review, or self-reported surveys, which are labor-intensive and prone to variability [25, 38, 40]. This highlights the need for a standardized, data-driven tool to assess and provide feedback on the frequency and quality of CLC during training in near real time.

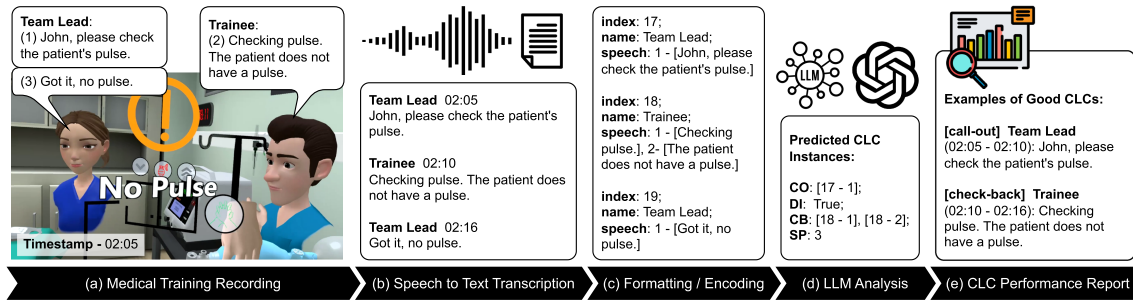


Fig. 1. Overview of AutoCLC: Starting with the left panel, a team is shown conducting a VR simulation-based cardiac arrest training. During the training, trainees must work collaboratively to identify cardiac rhythms, administer treatments, and communicate dynamically changing information regarding the patient's clinical status. Care team conversations are recorded and transcribed into text. A custom-designed algorithm, supported by a large language model, then analyzes the text to identify instances of effective and ineffective Closed-Loop Communication (CLC). This tool allows instructors, for the first time, to provide immediate and quantitative feedback on CLC performance to trainees with specific examples of effective and ineffective CLCs.

To address this critical gap, we propose AutoCLC, an automated tool designed to analyze team communication. AutoCLC identifies key elements of CLC, such as Call-Outs and Check-Backs as well as evaluates their quality and generates detailed quantitative reports to support post-training debriefings. As shown in Figure 1(a), AutoCLC leverages a **Virtual Reality (VR)** simulation-based medical training for ACLS training, part of an independent initiative at our institution. This multi-user “open-world” training simulation enables four clinicians to interact with each other, non-player characters, and a virtual patient, creating dynamic and non-deterministic scenarios [46]. Recordings from these simulations serve as input for AutoCLC. Speech data are transcribed, labeled, and processed through a custom pipeline (Figure 1(b) and (c)) and then analyzed using GPT-4 (Figure 1(d)). The system produces comprehensive reports (Figure 1(e)), offering qualitative and quantitative insights into CLC proficiency for targeted feedback. While AutoCLC is designed and evaluated within the VR environment, its mechanisms and analytical pipeline are broadly applicable and can be extended to non-VR settings, such as manikin-based simulations or real-world clinical environments. To generate ground-truth data for model training, evaluation, and assessment, we recruited a team of annotators with medical training in ACLS and expertise in CLC. This team analyzed a total of 361 minutes of training recordings, including VR video and audio data. Their efforts resulted in the annotation of 4,293 lines of dialogue, yielding 1,297 labeled CLC events.

Our evaluation showed that AutoCLC achieves 78.9% precision for identifying Call-Outs and 74.3% for Check-Backs. In comparison, results from three trained human annotators, used as ground truth for our evaluation, achieved 83.4% for Call-Outs and 79.6% for Check-Backs. The small performance gap, around 5%, highlights the efficacy of the AutoCLC approach. We fully assessed all system components and elicited instructors' perceptions of the educational value and utility of the CLC reports generated from the AutoCLC system through an evaluation study with 11 AHA instructors. Instructors appreciated the reports offering examples of good, poor, and unresponded CLC instances and quantifying CLC frequency. Finally, instructors suggested improvements for AutoCLC, including viewing transcripts of all CLC incidents, adding visual charts describing CLC frequencies, and enabling real-time feedback. Ultimately, this foundational work is the first to enable automatic detection and assessment of CLC and has applications across a wide range of professional domains that train teams to communicate effectively in mission-critical and life-saving scenarios. The core elements of AutoCLC—including its modular analysis framework, configurable timing thresholds, and adjustable metrics for quality communication—enable customization and scaling across diverse professional domains, e.g., remote team collaboration platforms, emergency response coordination, industrial safety operations. Importantly, AutoCLC demonstrates

how technology can augment human abilities while preserving autonomy, offering automated analysis that enhances rather than replaces human decision-making in complex team environments.

2 Related Work

This section reviews prior research on techniques for training and evaluating CLC in high-reliability teams. It also examines the literature on CLC in simulation-based medical training and studies on **Natural Language Processing (NLP)** for dialogue analysis.

2.1 CLC in High-Reliability Teams

First introduced in the military radio transmissions and aviation industry to ensure effective, clear, and unambiguous information exchange, CLC has been widely adopted across high-reliability teams as a fundamental communication principle [4, 13, 52], particularly where timely, precise, and error-intolerant communication is critical. It also provides a structured framework for analyzing interactions between team members in high-intensity fields such as aviation, energy production, and maritime, facilitating effective debriefing on team functioning [14].

Healthcare represents one of these high-intensity fields where CLC plays a crucial role by enhancing communication quality, improving team collaboration and performance, preventing adverse events, and safeguarding patient outcomes [20, 27, 49, 83]. Ineffective communication, whether delayed, misdirected, incomplete, or inaccurate, can result in communication-based errors, potentially leading to patient harm [74]. CLC has been shown to reduce the risk of such preventable errors by fostering transparent communication and coordinated teamwork toward shared objectives [21, 67]. Structured communication should be practiced in a safe environment, like high-fidelity simulation training.

2.2 Simulation-based Medical Training for CLC

National and international organizations, such as the AHA, the AHRQ, the World Health Organization, and the European Resuscitation Council, promote CLC due to its association with decreased adverse events [20, 27, 49, 83]. Most CLC training for healthcare teams is simulation-based [72], providing a controlled and safe environment to practice without real-world consequences. These simulations mimic high-pressure situations where CLC is crucial [35], although education strategies and assessment methods vary widely. The standard instructional practice emphasizes team debriefing immediately after the simulation, allowing for direct feedback while the events are still fresh in participants' minds. Most evaluation methods rely on instructor observations, manual video reviews, or self-reported surveys, which limit the timeliness, precision, and scalability of feedback during debriefs. For example, video-based reviews allow educators to observe communication skills in realistic scenarios. However, they require substantial resources and face challenges in quickly identifying teachable moments (i.e., opportunities to recognize learner needs and provide timely, targeted instruction or feedback), which limits their scalability. While some existing tools incorporate quantitative assessments, these approaches are typically labor-intensive and conducted post-hoc, requiring extensive manual coding of communication interactions (e.g., [6, 38]). Thus, while simulation training effectively allows teams to practice CLC, existing methods depend heavily on subjective observations and manual reviews, lacking efficient, quantitative evaluation tools. This creates a critical need for an automated tool capable of quickly analyzing team speech for CLCs and generating a readily available report for instructors to reference during debriefing.

In recent years, VR simulations are gaining popularity in healthcare training, including resuscitation education [64, 76]. Immersive technology has been used to teach both clinical techniques and nontechnical skills such as teamwork, communication, and situation awareness [15], with studies demonstrating improved perceptions of the learning experience and increased engagement and self-efficacy from the students [70]. In addition, VR environments facilitate unobtrusive data collection, directly capturing participants' speech, video, and key actions, which is adopted for AutoCLC.

2.3 NLP for Dialogue Analysis

CLC has been explored using machine learning techniques for classifying and annotating dialogue acts in speech [23, 30, 39, 59, 77]. For instance, previous research segmented dialogues into coherent units based on topic shifts, employing human annotations and case-based reasoning classifiers [58]. Another study introduced LCSeg, a lexical chain-based method for identifying changes in phonetic units to segment discourse [39]. The approach utilized features from word transcripts and audio inputs to train a classifier and explored real-time unsupervised segmentation. Robinson et al. investigated methods to enhance the accuracy and efficiency of detecting CLC instances in **Critical Care Air Transport (CCAT)** team simulations [68]. They integrated NLP techniques into the Team Dynamics Measurement System to evaluate team coordination. The study analyzed CCAT simulation data using flow-based and content-based methods, with the latter showing better accuracy but still prone to false positives.

Besides adherence to CLC, other factors and antecedents that influence effective communication in the healthcare environment include cognitive overload caused by distractions, inter-professional communication breakdowns during handoffs, interruptions caused by defective devices or equipment, social dynamics between colleagues, and the emotional status of team members [26, 33, 37, 73]. For instance, Coro et al. applied a cluster analysis to syllabic-scale spoken dialogue using the K-means algorithm to quantitatively evaluate how emotional factors such as agitation, stress, fear, or misunderstandings affect team communications [22]. LeBaron et al. assessed CommSense, an NLP tool for clinician communication analysis. Participants highlighted its potential to enhance compassionate communication and care quality, emphasizing interest in metrics such as medical jargon, interruptions, and speech dominance [53]. Another study developed an automated pipeline for analyzing teamwork communication in manikin-based nursing simulations [86]. The pipeline uses spatial tracking and f-formation theory to segment audio recordings into spatially distributed dialogue segments. OpenAI Whisper [66] automatically transcribed these segments. Then, BERT-based language model [24] classified each utterance as specific communication behavior, e.g., task allocation, handover, escalation, information sharing, responding to requests. The results show that this automated approach can achieve coding quality similar to manual methods.

While numerous factors contribute to reliable verbal communication in healthcare, CLC serves as a prominent marker of the communication quality for acute care teams, where interprofessional collaboration and the high risk of miscommunication can make both patients and clinicians especially vulnerable to errors [12, 42]. Advancements in algorithms and modern developments in NLP have introduced an increasing number of tools and techniques to support the analysis of CLC in the medical domain. In recent years, **Large Language Models (LLMs)** have demonstrated remarkable capabilities in understanding and reasoning about the world through language [88]. AutoCLC leverages the reasoning, contextualization, and comprehension abilities of LLMs to develop an automated tool for assessing care team communication.

3 Formative Study

A formative study was performed with four medical professionals and educators to understand their natural workflows in teaching CLC. This section describes the procedure and findings of the formative study. The key objectives of the formative study are to:

- Understand the challenges of assessing CLC in clinical training to enable precision-guided feedback and deliberate practice;
- Gather user requirements and expectations for new tools to facilitate CLC evaluation;
- Collect guidance and insights to inform the subsequent design process for solution development.

3.1 Participants and Procedure

The formative study consisted of several interview sessions. Four medical professionals from a large academic healthcare system in the United States participated in the formative study. These individuals, three clinician educators and one resident physician, regularly engage in hands-on simulation-based training. The interviews

were held in person at the hospital, lasting between 60 and 90 minutes. The study was conducted in accordance with our institution's Institutional Review Board guidelines.

3.1.1 Interview Session. During the interview, participants answered structured questions about their experiences with acute care team-based simulation training to identify challenges for evaluating CLC in the current medical training process. They detailed their facilitation of training sessions, including ACLS training procedures, necessary communication skills, and mechanics of post-simulation debriefing.

To gain comprehensive insights into both instructors' and trainees' perspectives on CLC in medical simulation-based training, the interview included distinct sets of questions for each viewpoint. The first part focused on the instructor's perspective. Participants described their methods for evaluating trainees' communication skills, particularly CLC, and the time spent analyzing performance and discussing communication skills with trainees. They also discussed challenges in providing feedback on CLC during debriefs. The second part focused on the trainee's perspective, as all ACLS providers and instructors are required to repeat training or recertify every 2 years. Participants reflected on how their CLC skills were assessed and how feedback was communicated. They also described their self-assessment of CLC skills in clinical situations and the challenges of understanding and receiving actionable feedback on CLC during debriefs.

3.1.2 VR Simulation-based Training Recordings. Although CLC is widely implemented across diverse scenarios, this research selects VR simulation-based medical training recording videos as the context for evaluating CLC. The Immersive Virtual Reality Environment for Training Acute Care Teams system (Figure 2) [47] is a multi-user "open-world" VR simulation-based training for medical education. This VR training immerses a team of four trainees in various in-hospital rapid-response scenarios, requiring collaborative efforts to deliver life-saving care to a virtual patient experiencing critical heart rhythm disturbances. Over the 10–15-minutes session, the team must assign specific roles, including team leader, chest compressions, airway management, defibrillator operation, and medication administration. The team's unique decisions, reactions, and actions lead to non-deterministic, dynamically evolving scenarios and outcomes. VR simulation enables unobtrusive data collection, directly capturing participants' speech, video, and key action metadata (e.g., ventilation, drug delivery, heart rate monitoring) through the VR headset and Unity engine. This setup enhances analysis granularity and ensures precise speaker identification, a critical feature for analyzing CLC, which is challenging to achieve in physical training environments.

3.1.3 CLC Trial with VR Recordings. After each interview, participants were shown a video recording of a VR simulation training (Figure 2) and asked to watch it without interruption. Following the video review, participants evaluated team performance with a focus on CLC. They identified specific instances of CLC from memory and proposed areas for improvement. After completing the video assessment, participants were provided with a hardcopy transcript of the recording, documenting each speaker's speech. They were then instructed to label or mark the transcript to evaluate CLC usage. Subsequently, participants were asked the same set of questions, including evaluations of overall team performance, identification of specific CLC instances, and suggestions for improvement.

This dual approach aimed to capture both immediate impressions and reflective insights. Additional questions gathered feedback on developing an intelligent interface for evaluating and providing effective feedback on CLC. Participants suggested features for optimizing CLC use and facilitating learning, identified key metrics for assessing CLC quality, and distinguished characteristics between high- and low-quality CLC. Finally, participants were asked about their expectations for features that would provide useful data summaries or visualizations as outcomes of an automated evaluation process.

3.2 Results of Formative Study

3.2.1 Limitations in Assessing CLC in Current Simulation-based Medical Training. According to participants, a typical ACLS training session lasts 10–15 minutes, followed by a 15–20 minutes debriefing. The debriefing



Fig. 2. Multi-user Immersive Virtual Reality Environment for Training Acute Care Teams (iREACT) simulation environment used in this study to collect, annotate, and evaluate closed-loop communication data.

provides time for reflection, discussion, feedback, and scenario repetition. Existing CLC assessments primarily rely on instructors' observations during sessions and their recollections during debriefing. These assessments are often presented in a checkbox format, asking questions such as, "Did this team use CLC?" or "How was the team communication?" Technical skills were often described as being the focus of the evaluation of team performance, leaving nontechnical skills, such as CLC, as secondary considerations. Participants attributed this to: (i) existing checklists and learning outcomes being primarily focused on technical performance, (ii) the difficulty of evaluating both technical and nontechnical skills simultaneously, and (iii) very frequent communication among team members, making it difficult to pay attention to all incidents of CLC. The assessment of CLC has traditionally been subjective without formal rating tools. Most evaluation approaches rely heavily on subjective instructor observations, manual video reviews, or self-reported surveys, which limit the timeliness, precision of feedback, consistency, and scalability of these approaches. Existing CLC evaluation tools incorporating quantitative assessments are typically labor-intensive and performed post hoc, requiring extensive manual coding of communication interactions [6, 38].

3.2.2 Video and Transcript Analysis Reveals Gaps and Affordances in CLC Assessment. After watching the example video, participants identified excerpts of effective communication and rated the team performance as good, noting that the team leader gave clear instructions and communicated goals. Despite satisfactory overall communication, many CLC instances were noted as needing improvement.

Participants were then given a written transcript of the simulation video to identify and count high- and low-quality CLCs and ineffective communication events. This step aimed at understanding the criteria used to identify CLC instances and how these were explained to trainees. In the marking system for the transcript, each communication instance was labeled with a number and a letter (e.g., 1A, 1B). A check mark indicated correct communication, while a square symbol represented incorrect communication. This method, later integrated into the annotation system for technical evaluation (see Section 5.1), proved more effective than video alone but was more labor-intensive and time-consuming.

Participants suggested focusing on specific areas in the transcript where CLC should be assessed to identify errors and opportunities for improvement. They highlighted the value of quantitative metrics, such as the number of Call-Outs, the percentage of completed Check-Backs, and the frequency of unresponded communications. They recommended showing specific examples of good and poor CLC from the trainees' dialogues. They noted that examining contextual factors affecting communication is crucial for understanding and addressing team performance issues.

3.3 Key Design Goals

Based on the formative study, we formulated the following design goals for developing an AI-generated report for evaluating CLC that can be used for providing data-informed feedback to trainees:

- *Automated CLC Instance Identification*: The system should be capable of automatically identifying and classifying instances of CLC within the team member dialogues, including Call-Outs, Check-Backs, and unresponded Call-Out statements.
- *Quantitative CLC Analysis*: The system should provide quantitative metrics on the CLC performance during the training session, such as the total number of Call-Outs, percentage of Call-Outs with Check-Backs, average frequency of Check-Backs and unresponded Call-Outs per minute, and overall CLC quality score.
- *Contextual CLC Highlighting*: The system should highlight and present relevant examples of good, poor, and unresponded CLC instances from the training session dialogues, along with the contextual information (speaker, timestamp, and dialogue content) to facilitate targeted feedback and enable trainees to see verbatim communication segments.
- *Visualization and Reporting*: The system should generate clear and concise visual reports or dashboards that summarize the CLC performance. To maximize the pedagogical value of the debriefing session, quantitative metrics and highlighted excerpts should enable instructors to quickly identify areas for improvement, track progress in case of multiple attempts by the same trainee team, and tailor their teaching strategies accordingly.

3.4 Additional Key Concept Definitions

Several key concepts of CLC, Call-Out, Check-Back, and Closed-Loop, as outlined in Section 1, were derived from existing definitions. Beyond these foundational definitions, a refined analysis was performed to identify precise steps within sentence structures, evaluate the necessity of each component in Call-Outs and Check-Backs, and extract high-level common elements essential for these interactions. Extensive consultations with medical educators and researchers guided this iterative design, definition, and quantification process. In the VR-based medical training videos, conversations were often observed that consisted primarily of a Call-Out followed by a Check-Back, with the Check-Back frequently lacking acknowledgment by the original sender.

After consulting experts and educators on the fidelity of real-world acute care team practices, we concluded that it has become standard practice for team members in ACLS scenarios to forgo verbally “closing the loop” by acknowledging the Check-Back. This is likely done for efficiency, as all participants are present in the same room and can rely on visual cues to confirm task completion. Additionally, team members typically have a shared understanding of the situation, reducing the perceived need for explicit verbal confirmation.

This contrasts with professions like aviation, where the tower controller and pilot operate in separate physical spaces and do not share the same situational context. In such settings, verbal loop closures are critical to ensure mutual understanding. Consequently, in the ACLS training scenarios we observed, formal verbal loop closures were virtually nonexistent.

- *Non-Specific Check-Back*: Response without explicitly repeated medical instructions or dosage in the “Check-Back.”

- *Inadequate Response*: Ambiguous responses such as “uh-huh,” “okay,” or “I got it,” etc.
- *Unresponded*: A “Call-Out” sentence that lacks an associated “Check-Back.”
- *Directed Call-Out*: Direct identification of an individual team member, intervention requested, and dosage or voltage if appropriate.
- *Uncategorized (None)*: All other forms of communication that do not belong to the categories listed above.

4 AutoCLC: CLC Evaluation Tool

With the design goals outlined in Section 3.3, we created *AutoCLC*, an automatic CLC evaluation tool that generates performance reports on healthcare teams’ use of CLC from recordings of simulation training. Figure 1 shows an overview of the AutoCLC pipeline. In the leftmost panel, a cardiac arrest training simulation is being recorded with conversations from all team members. Then, the audio file of the simulation session recording is transcribed using otter.ai [63], a speech-to-text transcription software, with speaker identifications from each headset and timestamps for each line of dialogue. Next, a Python script encodes the transcript into the input format for LLM analysis with the line index, speaker identity, and transcribed speech content. Then, the formatted transcript is analyzed using the *GPT-4* model from OpenAI ChatGPT with a customized prompt, which identifies and outputs instances of CLC specifying the indices of Call-Out (CO) and Check-Back (CB) statements, whether the Call-Out is *directed* (DI), and how *specific* (SP) the Check-Back is. The ChatGPT prompt engineering process details are explained in Section 4.1. Finally, a CLC report is generated from the predicted CLC instances with team performance statistics and examples of good, poor, and unresponded CLCs. The features of the AI-generated report and an example CLC report are provided in Section 4.2.

4.1 LLM Transcript Analysis with Customized Prompt

To automatically evaluate a resuscitation team’s use of CLC and generate a performance report to be presented to the medical instructors and trainees, the *GPT-4* LLM from the OpenAI ChatGPT is used to analyze the transcripts of VR simulation session recordings [2]. To clearly explain the tasks to be performed and generate examples of CLC instances that conform to the definition provided by medical professionals, a prompt engineering process is required to fine-tune the ChatGPT model. Figure 3 visualizes the prompt engineering pipeline, including the *Initial Prompt* and the *Execution Loop* through which sections of the transcripts are analyzed.

The *Initial Prompt* (Figure 3(1)), which contains around 1,100 words, consists of five components. The first two introduce the definition of CLC in a medical context (Figure 3(1a)) and provide examples of CLC and non-CLC from the training transcripts (Figure 3(1b)), both of which are constructed from the results of the formative study (Section 3.2). The examples emphasize the existence of actionable items (e.g., task delegation) in the Call-Out and the repetition of the medication and its dosage or other quantitative information in the Check-Back. The third component provides the input format of the transcript to be analyzed (Figure 3(1c)). The components of each line of the transcript are delimited by semicolons. Each line contains three components. The number after the phrase “index:” is the line index of the dialogue. The phrases after “name:” indicate the speaker. The words after “speech:” are the actual dialogue, which is separated into individual sentences. Each sentence is indexed by a number followed by a dash, with the speech content delimited by a pair of square brackets. The *Input Transcripts* component in Figure 3(2) demonstrates the input format through an example transcript.

Then, the fourth part explains the four analysis tasks to be performed (Figure 3(1d)):

- (1) Identify the occurrences of any questions or requests (Call-Outs) in the dialogue.
- (2) For each instance in Task 1, identify whether the question or request is directed, i.e., whether the requester mentions the name or identity of the receiver in the Call-Out.
- (3) Identify whether the Call-Out has a corresponding Check-Back within the next *five* lines of dialogue and return its line and sentence index.

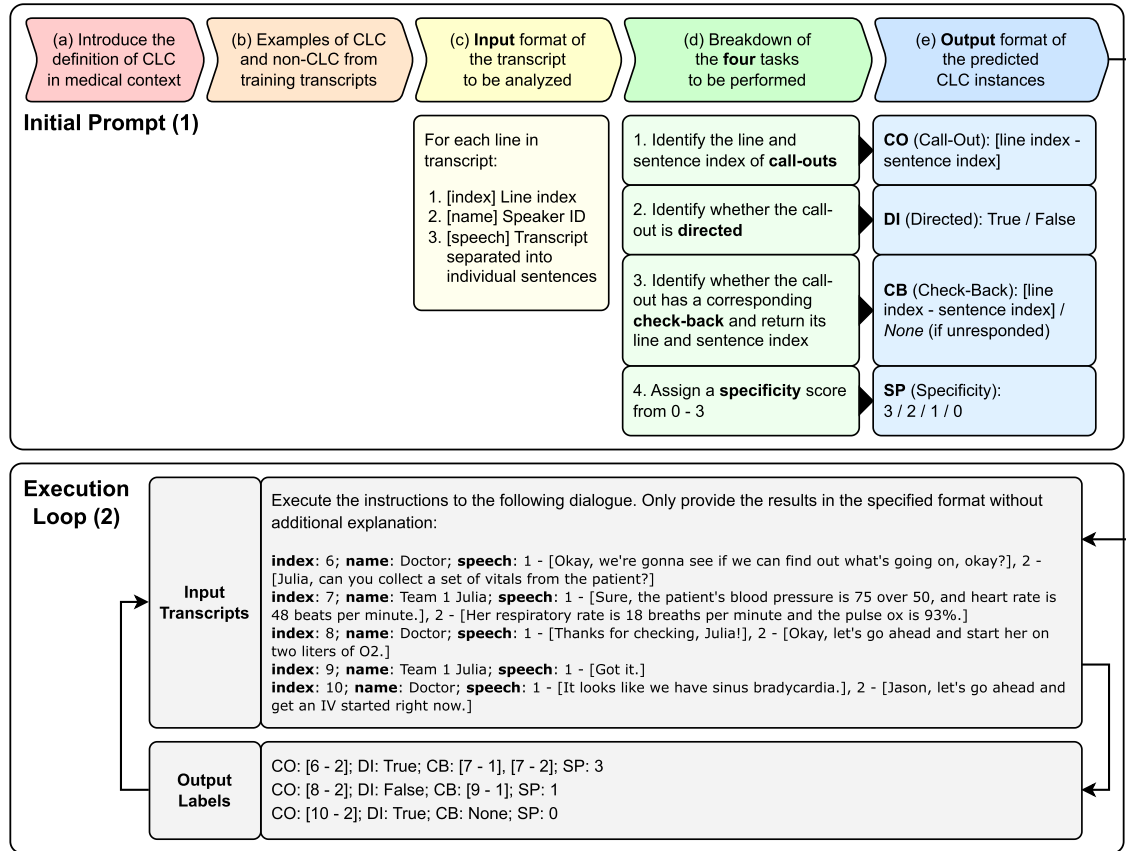


Fig. 3. The prompt engineering pipeline for evaluating Closed-Loop Communication (CLC), including the initial prompt and the execution loop through which sections of the transcripts are analyzed.

- For each response identified in Task 3, evaluate how specific the response is. Provide a score of 1–3, with 3 being the most specific. Return 3 if the response repeats all critical information in the request; 2 if it contains part of the information or any action verb; 1 if it repeats no information and merely provides an acknowledgment such as “okay,” “got it,” or “on it.” For unresponded instances identified in part 3, return 0 for this task.

The last component explains the output format (Figure 3(1e)). The example shown in the *Output Labels* (Figure 3(2)) contains three instances of Call-Outs, which are [line 6 sentence 2], [line 8 sentence 2], and [line 10 sentence 2]. Each instance contains four components indicated by “CO” (Call-Out), “DI” (Directed), “CB” (Check-Back), and “SP” (Specificity), which correspond to the four tasks mentioned previously. The first and third instances are directed, while the second one is not. The first and second instances have corresponding responses, which are line 7 sentences 1 and 2, and line 9 sentence 1, respectively. The last instance ([10 - 2]) is unresponded, as indicated by the “None.” Within the two instances with a response, the first response, which contains two sentences ([7 - 1], [7 - 2]), contains all key information and scores a 3 for specificity, while the second one ([9 - 1]) only provides a vague acknowledgment and scores a 1. For unresponded instances, the fourth task is not applicable, as indicated by the “0” in the last line. Note that all Check-Backs are within five lines of the Call-Outs.

After the *Initial Prompt*, sections of the transcripts containing 15 lines of dialogue each are entered into the ChatGPT interface in the *Execution Loop*. The 15-line window is a parameter determined through experimentation with various window sizes during a pilot study on three videos, through which we identified 15 lines to be an optimal balance between capturing context and accurately identifying all CLC occurrences, as this window typically encompasses a complete topic segment during medical simulation. Each group of adjacent sections overlaps by five lines to compensate for the Call-Out and Check-Back pairs spanning multiple sections. All the *Output Labels* are compiled into a single text file for post-analysis and generating the report. The entire customized ChatGPT prompt for transcript analysis is provided in the Appendix A for reference.

4.2 Description of the CLC Report

Using all predicted CLC instances from a training transcript, i.e., *Output Labels* in Figure 3, a CLC report is generated based on the design goals summarized in Section 3.3. Figure 4 illustrates an example of the AI-generated CLC report. The top section presents descriptive team communication statistics, revealing that the session lasted 10.4 minutes and contained 129 lines of dialogue. A total of 48 CLC Call-Out statements were identified, indicating instances where a team member initiated a specific medical instruction or request. Of these 48 Call-Outs, 30 (62.5%) received a Check-Back response. A Check-Back involves the receiver acknowledging and repeating back the original message to confirm understanding. However, 18 Call-Outs (37.5%) went unresponded. The report highlights that, on average, there were 2.89 Check-Backs and 1.73 unresponded Call-Outs per minute, indicating room for improvement in maintaining consistent CLC practices [6].

To further exemplify the varying levels of CLC quality, the three top-ranked conversational interactions from each of the three categories are automatically generated into a report. The AI-predicted CLC instances with a specificity score of 3 (“SP: 3”) are categorized as “good CLCs”; the ones with a score of 1 (“SP: 1”) are categorized as “bad CLCs”; the ones with a score of 0 (“CB: None; SP: 0”) are unresponded. The “good CLCs” section showcases three instances where Call-Outs were followed by clear, specific Check-Backs, such as “Epinephrine is in” or “Synchronized cardioversion at 200.” In contrast, the “poor CLCs” section highlights instances where Check-Backs were non-specific or vague, such as “Drawing up,” “Yep,” or “Sure.” Finally, the “Unresponded Call-Outs” section provides examples where team members did not provide any Check-Back response, representing a breakdown in CLC. Later, in Section 6, the usability and effectiveness of the CLC reports were evaluated through a user study with 11 medical educators, all with extensive experience in cardiac arrest training.

5 Technical Evaluation

This section presents the technical evaluation of the AI-enhanced CLC evaluation tool. The evaluation is divided into two parts: the accuracy of ChatGPT predicted CLC instances (Section 5.2) and the quality of generated CLC reports (Section 5.3). To collect a set of ground-truth labels and compare the AI-predicted CLC instances against, a user annotation interface was designed to allow trained medical professionals to identify and classify Call-Outs and Check-Backs in the transcripts. The specific features and functionalities of the user interface are described in Section 5.1.

5.1 User Annotation Interface Description

The user interface of the AI-enhanced CLC evaluation tool was developed to streamline the annotation and evaluation of transcripts from medical training sessions, enabling easy identification and quality assessment of CLC instances. This interface addresses the gap in tools capable of tracking communication usage and evaluating medical professionals’ communication skills in closing the loop. This section provides a detailed overview of the user interface components and their functionalities, which were designed to optimize user experience and facilitate efficient navigation.

Before using the annotation tool, participants were provided with video instructions, and a study coordinator was available to answer any questions they may have. Upon entering the web interface, users are prompted to

AI-generated report for VR training session: [session_name]

Total lines of dialogues: 129

Duration of the training session: 10.4 minutes

Total number of CLC call-outs: 48

Within them, 30 (62.5%) call-outs have a check-back; 18 (37.5%) call-outs are unresponded.

On average, there are 2.89 check-backs per minute and 1.73 unresponded call-outs per minute.

Examples of good CLCs:

[call-out] Team Lead (8:20 : 8:29): Epinephrine and one milligram IV please.

[check-back] Speaker 3 (8:29 : 8:40): Epinephrine is in. Epinephrine is in at 7:05.

[call-out] Team Lead (3:29 : 3:42): And give a dose of epinephrine this time please.

[check-back] Speaker 3 (3:58 : 3:59): One milligram of epinephrine is in.

[call-out] Team Lead (2:24 : 2:45): So can we go for synchronized cardioversion?

[check-back] Speaker 3 (2:48 : 2:51): Synchronized cardioversion at 200.

Examples of poor CLCs (non-specific check-back):

[call-out] Team Lead (6:57 : 7:11): Can we draw up another 150 of amiodarone and don't give it yet.

[check-back] Speaker 3 (7:11 : 7:12): Drawing up.

[call-out] Team Lead (7:55 : 7:58): Steve, can you start compressions real fast?

[check-back] Speaker 4 (7:58 : 7:59): Yep.

[call-out] Team Lead (2:24 : 2:45): Can we make sure we are synced please?

[check-back] Speaker 3 (2:45 : 2:46): Sure.

Examples of unresponded call-outs:

[call-out] Team Lead (5:22 : 5:45): And can we get an end-tidal monitor on this patient whenever we get a chance?

[check-back] (No response)

[call-out] Team Lead (1:20 : 1:27): Can we start a liter of LR on him please?

[check-back] (No response)

[call-out] Team Lead (5:01 : 5:12): And can we have a hand on the pulse, please?

[check-back] (No response)

Fig. 4. AI-generated CLC report with team performance statistics and examples of good, poor, and unresponded CLC instances.

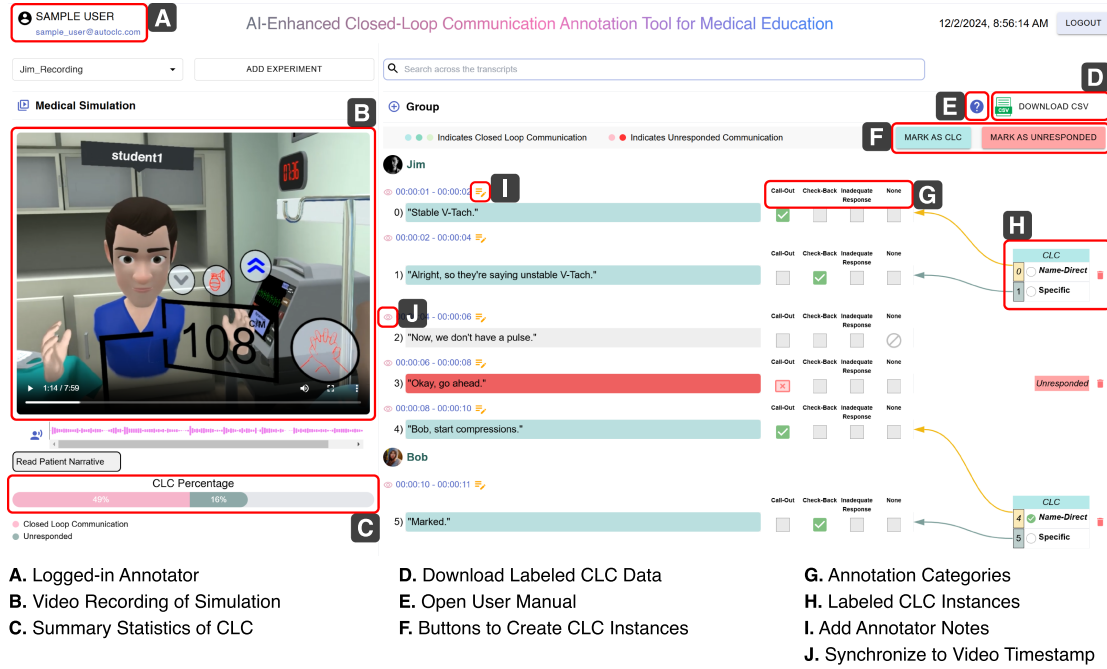


Fig. 5. Interface of the CLC annotation tool for team-based clinical simulation training.

enter their name and e-mail and select an experiment, ensuring personalized access to annotated data (Figure 5(a)). The tool also features a selection menu for switching between various video simulations to be annotated. A user manual is readily available, guiding users through the annotation process step-by-step (Figure 5(e)). The interface displays the video simulation on the left side (Figure 5(b)), with corresponding timestamped transcripts on the right. Users can easily navigate between the video and transcript using icons at the start of each transcript line (Figure 5(j)).

The transcript labeling interface allows users to categorize communication instances into four categories: Closed-Loop (Call-Out + Check-Back), Unresponded (Call-Out without Check-Back), None (inquiries not involved in CLC evaluation), and Inadequate Response (responses without confirmation or comprehension) (Figure 5(g)). Users can annotate transcripts as closed-loop or unresponded (Figure 5(f)) and add notes or explanations (Figure 5(i)).

When users group transcript lines, a visual representation connects each chosen transcript line with an arrow and a descriptive table (Figure 5(h)). The table shows if the grouped transcript is a CLC or unresponded, and whether call-outs are name-directed and check-backs are specific. Check-backs can be specific (e.g., “1 mg of epinephrine is in at 7:05”) or not (e.g., “Got it, drawing up”), where a specific check-back mirrors the call-out (e.g., “Tom, give 1 mg of epinephrine, now”). Users can undo annotations using the trash icon. The interface automatically generates a summary of the overall CLC percentage in the simulation video based on user labels (Figure 5(c)). Additional features include search word functionality, patient narrative viewing, and the ability to download annotated transcripts in CSV format (Figure 5(e)). This annotation tool facilitates the evaluation of medical professionals’ communication skills by providing visual context, easy navigation, and efficient annotation capabilities.

5.2 ChatGPT Performance Statistics

To evaluate the accuracy of the CLC instances predicted by ChatGPT using the Prompt Engineering Pipeline introduced in Section 4.1, 12 VR simulation session recordings were transcribed using otter.ai [63] with speaker identifications and converted into the input format specified in Figure 3. The average length of each recording is

Table 1. ChatGPT Performance in Evaluating Call-Outs and Check-Backs

Metrics	Call-Out	Check-Back
ChatGPT (Precision)	78.90%	74.27%
ChatGPT (Recall)	72.92%	71.54%
Inter-rater reliability	83.46%	79.67%

10.0 minutes (min: 6.1, max: 15.3). The 12 transcripts were analyzed using the pipeline with 12 lists of predicted CLC instances generated. On average, ChatGPT identifies 38 instances of Call-Out (min: 23, max: 55).

To compare the performance of ChatGPT against ground-truth labels, three medical students with backgrounds in emergency medical services and familiarity with the ACLS training were recruited to annotate the 12 transcripts using the interface introduced in Section 5.1. The same definition of CLC in the ChatGPT prompt was provided to the annotators. All three annotators performed the labeling manually, with no reference to any AI-assisted features. Excluding the time spent watching the tutorial video and familiarizing themselves with the interface, the annotators spent around 15–30 minutes labeling each transcript, depending on its length. All labeling sessions were conducted remotely through Zoom with a member of the research team being present to answer any questions. At the end of each labeling session, the human-annotated CLC labels were downloaded from the web interface and converted into the same format output by ChatGPT. All three sets of labels were concatenated to form a single list of ground-truth labels. In total, the three annotators watched 361 minutes of simulation training recordings with 4,293 lines of dialogue combined. This process yielded 1,297 labeled CLC events, generated exclusively based on the annotators' medical expertise and their understanding of CLC.

For all 12 transcripts, the AI-generated labels were compared against the ground-truth labels, and the ChatGPT performances are summarized in Table 1. A predicted Call-Out or Check-Back is counted as correct if at least one sentence overlaps with the Call-Out or Check-Back in a CLC instance within the ground-truth labels, i.e., the specific sentences within a line of dialogue will also need to be correctly identified. On average, for identifying Call-Outs, ChatGPT achieves 78.90% precision, i.e., the percentage of AI-predicted Call-Outs that were also identified by at least one human annotator, and 72.92% recall, i.e., the percentage of the Call-Outs in the ground-truth labels that were identified by ChatGPT. Within all correctly identified Call-Outs, ChatGPT achieves 74.27% precision and 71.54% recall in predicting the corresponding Check-Backs. Note that while the three annotators were provided with the same definition of CLC and labeling tutorials, the inconsistency between each annotator imposes a fundamental upper limit on the accuracy achievable by ChatGPT regardless of the prompt engineering efforts, i.e., the AI cannot perfectly agree with all labelers at the same time. To quantify the effect of such inconsistency, inter-rater reliability metrics were calculated by evaluating the accuracy of each annotator using the other two annotators' labels as ground truth. As shown in the last row of Table 1, the inter-rater reliability is 83.46% for Call-Out and 79.67% for Check-Back, which is only around 5% higher than ChatGPT's precision (4.56% for Call-Out and 5.40% for Check-Back).

5.3 CLC Report Quality Rating

The CLC reports generated in Section 4.2 are useful to the medical educators if the selected examples conform to the operational definition of good, poor, and unresponded CLC instances. To verify the accuracy of the CLC instances in the reports, all examples from the 12 reports were rated by three human annotators in terms of the quality of the Check-Back responses and how well they closed the loop to prevent miscommunication and missteps.

A Google Form survey was created for each CLC report. For the good and poor examples of CLC, the annotators were asked to rate the quality, i.e., specificity, of the Check-Back on a Likert scale of 1–7. The descriptions of the seven scores are listed in Table 2. Note that the question prompt does not indicate whether the Check-Back was

Table 2. Distribution of the Annotators' Ratings for the Check-Back Quality within Good and Poor CLC Instances in the 12 CLC Reports Generated by the AutoCLCtool

CLC Quality Rating	Description of Rating	Instances of Each Score within AI-predicted Good CLC	Instances of Each Score within AI-predicted Poor CLC
1	Very Poor	0	4
2	Poor	1	44
3	Relatively Poor	0	29
4	Neutral	12	11
5	Relatively Good	23	13
6	Good	26	4
7	Very Good	46	0

For human ratings, scores 1–4 are considered poor-quality Check-Backs, and scores 5–7 are considered high-quality Check-Backs.

an instance of good or poor CLC predicted by AI. For the examples of unresponded Call-Outs, the annotators were asked to indicate whether they agreed with the AI prediction (unresponded) or not (responded). The line containing the highlighted Call-Out statement and the next five lines of dialogue are presented in the survey to provide the context. In addition, for unresponded Call-Outs, the survey has a “not sure” option if the annotator believes that the Check-Back is ambiguous. In total, 12 survey forms were created, and the order of questions within each survey was randomized.

Three medical students with background in emergency medical services were recruited to complete the survey. Overall, 108 human labels were collected for good CLCs predicted by AI, 105 for poor CLCs, and 102 for unresponded (some CLC reports contain fewer than three poor or unresponded CLC instances). Table 2 shows the distribution of the annotators' ratings for the Check-Back quality within good and poor CLC instances, e.g., the third column shows the number of instances of each human-rated score within AI-predicted Good CLCs. The average human-rated score is 5.95 (SD: 1.11) for instances of good CLC predicted by AI and 2.97 (SD: 1.23) for poor CLC. A Mann–Whitney U test confirms that the distributions of the two groups of scores are distinctly separated ($p < 0.005$) [3]. For instances of good CLCs generated by AI, 88.0% of human labels agree with the prediction (score 5–7); for instances of poor CLCs generated by AI, 83.8% of human labels agree with the prediction (score 1–4). For the instances of unresponded Call-Outs, 91.2% of human labels agree with the prediction (unresponded), 4.9% disagree (responded), and 3.9% indicate unsure.

6 User Study

This section details a user study conducted to evaluate AutoCLC, focusing on its effectiveness, usability, and practicality in assessing and enhancing CLC during team-based healthcare training. By engaging experienced ACLS instructors, the study examines how the tool's automated analysis aligns with current evaluation practices, its potential to improve feedback quality and its feasibility for integration into both simulation and clinical settings.

6.1 Participants

We recruited ACLS instructors through an e-mail sent to two medical schools and a local community college. Table 3 lists the participant information. Participants included 11 ACLS instructors (6 males, 5 females) who were either paramedics or nurses, with ACLS teaching experience ranging from 1 to 40 years. Informed consent was obtained prior to individuals agreeing to participate in the study. We chose ACLS instructors because they are specifically trained in evaluating team communication and regularly assess CLC as part of their teaching responsibilities. Additionally, most ACLS instructors are practicing clinicians who use CLC during real patient

Table 3. Participant Details

ID	Gender	Profession	Years Teaching ACLS
P1	Female	Paramedic	6
P2	Male	Paramedic	5
P3	Male	Paramedic	4
P4	Male	Paramedic	40
P5	Female	Paramedic/Registered Nurse	1
P6	Male	Paramedic	15
P7	Female	Registered Nurse	34
P8	Female	Registered Nurse	11
P9	Female	Registered Nurse	20
P10	Male	Registered Nurse	28
P11	Male	Paramedic	38

care, so they are well suited to evaluate the accuracy and utility of our automated assessment approach and its potential transfer from simulation to clinical settings.

6.2 Procedure

A single interviewer conducted semistructured interviews with participants through Zoom. Interviews lasted approximately 45–60 minutes. Interviews started with collecting data on participants' demographics and professional backgrounds. Participants were provided an overview of CLC (based on AHRQ's TeamSTEPPS Fundamentals Course [18]), which included essential CLC terminology such as Call-Outs and Check-Backs.

Next, participants were asked how they evaluate and provide feedback on CLC. They then observed a 10 minutes VR-based video demonstration of the simulated cardiac arrest scenario performed by a team of resident physicians. Participants were asked to evaluate the CLC performance using techniques typically used to teach ACLS. Specifically, they were asked to characterize the overall quality of CLC, provide an estimate of the percentage score for the team's CLC based on the number of Call-Outs that were closed, recall specific instances of effective and ineffective CLC, and interpret how CLC quality impacted the team's overall performance and patient care.

Participants were then asked to review the automated CLC report generated using the AutoCLC. They were then asked how the tool impacted their previous evaluation of CLC quality observed in the VR-based video and how this would impact the feedback they would provide to learners. They were asked whether the tool helped them notice aspects of the team's communication dynamics they may have missed or overlooked and how they might incorporate insights from this CLC report into instruction and feedback. Participants were then asked questions regarding usefulness, usability, and suggestions for improvement of the automated CLC report and AutoCLC tool.

Interviews were transcribed and manually coded using a priori codes based on six predetermined themes. A deductive thematic analysis was conducted based on the Braun and Clarke procedure [16]. These main themes can be found in Table 5, along with subthemes that emerged from the data.

After the interviews, participants were asked to complete an online survey consisting of the following: Acceptability of Intervention Measure, Intervention Appropriateness Measure, and Feasibility of Intervention Measure. This survey was developed and validated using acceptability, appropriateness, and feasibility as criteria that are key indicators of the success of implementation processes [65, 79]. The online survey also included the **Systems Usability Scale (SUS)**, a widely used questionnaire in industry and academic research to assess the subjective usability of products or systems, including software, web sites, and hardware [55].

Table 4. Acceptability, Appropriateness, and Feasibility of AutoCLC

Outcome Measure	Mean	SD
Acceptability		
AI-powered CLC tool meets my approval	4.73	0.47
AI-powered CLC tool is appealing to me	4.55	0.69
I like AI-powered CLC tool	4.55	0.82
I welcome AI-powered CLC tool	4.64	0.67
Total Acceptability	4.61	0.66
Appropriateness		
AI-powered CLC tool seems fitting	4.64	0.50
AI-powered CLC tool seems suitable	4.55	0.69
AI-powered CLC tool seems applicable	4.73	0.47
AI-powered CLC tool seems like a good match	4.73	0.47
Total Appropriateness	4.66	0.45
Feasibility		
AI-powered CLC tool seems implementable	4.64	0.50
AI-powered CLC tool seems possible	4.82	0.40
AI-powered CLC tool seems doable	4.73	0.47
AI-powered CLC tool seems easy to use	4.55	0.93
Total Feasibility	4.68	0.46

6.3 User Perception Evaluation of AutoCLC

Results from the online survey demonstrated that participants rated acceptability, appropriateness, and feasibility highly, suggesting that the tool is well received by users, aligns effectively with the intended purpose, and is feasible to integrate into practice. Table 4 provides the scores for each criterion. The high acceptability score of 4.61 indicates that users find the tool agreeable and satisfactory, which is crucial for its adoption. The appropriateness score of 4.66 reflects that the tool is perceived to be a good fit and relevant for addressing the targeted needs, enhancing its potential effectiveness. Moreover, the feasibility score of 4.68 signifies that the tool can be effectively utilized given the existing resources and constraints, indicating a strong likelihood of successful adoption and effectively meeting the users' needs.

The total calculated SUS score was 76.36 (SD=16.25). An SUS score of 76 indicates a notably high level of usability. This score falls within the "Good" range on the adjective rating scale, signifying that the tool is effective in assisting users in achieving their goals, efficient in the effort and resources required, and generally satisfactory in the user experience. Specifically, a score of 76 suggests that the CLC tool can effectively produce automated CLC reports that accurately calculate the total number of CLC instances and offer pertinent examples of both "good" and "poor" CLC. This functionality likely supports ACLS instructors in thoroughly assessing and providing feedback on their students' communication skills during simulated cardiac arrest scenarios. Additionally, by automating the analysis and reporting of CLC instances, the tool can improve efficiency by reducing the time and effort required by instructors to evaluate communication patterns. Overall, the SUS score implies the tool's interface and usability are well designed, enabling instructors to utilize its features without significant frustration or complexity.

6.4 Comprehensive Evaluation of the AutoCLC and Insights

Saturation was met within each of the 6 predetermined themes, leading to the emergence of 22 individual sub-themes. Each subtheme, along with a brief explanation, can be found in Table 5. The following are details of the results.

Table 5. Detailed Insights on Current Practices, Tool Impressions, and Barriers to Adoption

Main and Sub-Theme	Descriptions
(1) Current methods	Current methods of recording and evaluating CLC.
Unstructured	There is currently little structure to the process of observing and evaluating CLC.
Difficulty evaluating	It is difficult to evaluate CLC accurately due to the dynamic nature of the scenarios and the focus on technical skills (e.g., decision-making, procedural skills).
Notetaking/checklist	Unstructured notetaking and some checklist use are the current practices for recording observed CLC.
CLC evaluation is secondary	Evaluators tend to focus on observing and evaluating technical skills, not CLC.
(2) Impression of Generated Report	Participants' impression of the automated CLC report generated from the demonstration they observed.
Surprising	Report generated many findings that the participants did not notice.
Quantifiable	Report was able to quantify the frequency of CLC. This was difficult using current methods.
Specific examples	Valuable examples were provided in the report.
More accurate	Report appeared more accurate than current methods.
(3) Usefulness	Uses for the generated CLC report during debriefing.
Targeted feedback	Specific examples generated by the report can be used to give specific feedback to learners.
Reminds debriefer of CLC	Having the report would remind the instructor to provide feedback on CLC.
Recognize oversights	Instructor may review the report prior to feedback and identify communication they did not observe.
(4) Ease of use	Reported ease of reading and interpreting the report.
Easy to interpret	The report was easy to read and quickly interpret.
Categorized examples	The categorized examples (i.e., great, good, bad) made it easy to interpret.
Concise	The report was free of clutter and focused only on key findings.
(5) Suggestions	Suggestions for improvement to the generated report.
Real-time feedback	Seeing the report being generated in real time during the simulation could allow instructors to pause simulations and provide immediate feedback.
Provide all examples	Provide all CLC examples that could be searched to find specific incidents the instructor may want to discuss.
Benchmarks	Set ideal standards for total CLC frequency.
Color and visual charts (pie, bar, etc.)	Use color and visual charts to enhance the ability for instructors to read and interpret the results.
(6) Barriers	Barriers to adoption.
Cost	Cost of the tool could pose a barrier to adoption as organizational budgets are often limited.
Timeframe to receive results	If results are not provided in a timely manner after a simulation is complete, this could result in the report not being used.
Trusting the accuracy of the results	Some instructors may not trust the results provided in the report.

6.4.1 Current Methods. The key findings about instructors' prior CLC evaluation practices align closely with our formative study findings (see Section 3.2). In brief, participants reported unstructured assessment methods, difficulty in the simultaneous evaluation of technical and communication skills, and inconsistent observation methods.

6.4.2 Impression of Generated Report.

- *Surprising*: After being provided with the CLC report generated by AutoCLC, most participants expressed some degree of astonishment with the results and mentioned they gained more accurate insight regarding using CLC during the simulation. Most participants indicated that they did not observe as much CLC as described in the CLC report. P5 stated: *"So it looks like there was a lot more close loop communication than what I saw."*
- *Quantifiable*: One of the features participants mentioned the most was the ability of the tool to quantify the number of CLC instances occurring during the simulation. P10 stated: *"I think it does quantify all those incidents of closed-loop communication. I like how it shows when there was closed-loop communication and where there was not."* Some participants preferred having a benchmark for the percentage of Call-Outs that received Check-Backs. P5 stated: *"It shows that there was more closed-loop communication than what I saw, but it was still not to the benchmark that I would like to see."*
- *Specific Examples*: Participants highlighted the examples provided in the report as a particularly valuable feature. Reports listed up to three examples each of good, poor, and unresponded CLC. Participants specifically commented on how this could be used to enrich the quality of debriefing by showing learners examples of their own behaviors and instructing them on how they can improve their communication skills. P7 stated: *"You could say to [learners], you know, this wasn't as good as I'd like, but you're getting better. See this example of good closed-loop communication versus poor. Do you see? When someone called out, that was not closed. That's not considered good communication."*
- *More Accurate*: Several participants indicated that they felt the report captured the frequency of CLC more accurately than their live observation, particularly given the challenges of manual note-taking during fast-paced simulations. P1 explained: *"It's nice because as fast as I was trying to write things, more things were happening."* This limitation is especially pronounced in debriefing sessions, which typically last only 10–15 minutes, preventing instructors from thoroughly reviewing video recordings to identify teachable moments. One participant mentioned that they teach ACLS in a high-fidelity simulation environment where the instructor is situated in a separate room from the trainees (behind a one-way mirror), and they must rely on audio in the control room to examine the quality of communication. P11 described: *"The instructor is not even in the room when we do our classes; we're in a control room. We can't always easily see and hear all the participants. This would be a great tool."* In these scenarios, the tool provides a more comprehensive record of communication patterns that might otherwise be missed during live observation.

6.4.3 Usefulness.

- *Targeted Feedback*: Most participants discussed the usefulness of the examples provided in the report to offer targeted feedback to learners. Participants discussed that they could highlight the specific examples to learners and better explain how to improve communication. Specifically, the fact that the examples are categorized (i.e., Good, Poor, Unresponded) facilitates the ability of instructors to easily compare and contrast different CLC qualities and improve feedback for learners. P1 stated: *"Well, it helps to point out specific incidents where you know there might have been a failure in closed-loop communication. You can also highlight good things that happened. I like to provide positive and negative examples instead of just hammering them with the negative. That always helps clarify what we are looking for in a code [team]."*
- *Reminds Debriefers of CLC*: Some participants discussed how this report could help remind the instructor to discuss CLC during debriefing. As previously mentioned, participants indicated that technical skills were

often the focus of evaluation and debriefing, leaving non-technical skills as a secondary consideration. Having this report available to instructors could help remind them to discuss CLC. P6 stated: *"I think a tool like this would prompt the debriefer to discuss communication because it's often not discussed in debriefing. Having this tool reminds the debriefer that this is something they need to talk about."*

- *Recognize Oversights*: Participants discussed how they are prone to missing incidents of communication when observing simulations due to the dynamic nature and communication clutter that is common during those sessions. Some participants discussed how this report could help identify some communication they missed when observing. Without the tool, the instructor is limited to discussing only what they were able to observe or reviewing a video of the session, which can be time-consuming. P3 stated: *"If I'm evaluating a code that's 15 minutes long, remembering who answered each time can be difficult. That's why I took some notes, but I still didn't catch everything. I may have missed the acknowledgement of an order that was actually given. So, this [CLC tool] is very helpful."*

6.4.4 Ease of Use.

- *Easy to Interpret*: Although suggestions were made to improve the information provided in the report, all participants provided positive feedback on the sample they were provided. Participants felt the current layout of the CLC report was easy to review and interpret in a timely manner for the purposes of debriefing. This included the layout of examples and description of CLC frequency. P2 stated: *"I like the way this is set up. This is very easy to interpret."*
- *Categorized Examples*: The most frequent feedback regarding the ease of use of the report referred to the categorized examples. Participants appreciated that examples were separated into different categories (i.e., good, poor, unresponded), which made it easy to read and interpret. Additionally, participants felt that this would make it easier to use in debriefing, as they can point out examples of communication varying in quality. P3 stated: *"I like that there are good examples and poor examples; it's pretty nice. It gives you the different examples of what's good and what's not so great, as well as the examples of Call-Outs that went unresponded."*
- *Concise*: Some participants discussed that the limited information in the report kept it concise, providing only the necessary information to debrief CLC. One feature mentioned by participants was that using only a few examples in each category reduced clutter in the report. Although participants later suggested having the option to look through all communication, having only three examples in each category was ideal for debriefing. P6 reported that *"I think if it had too much information on the display, I would not have time to sit there and read through it before students sat down to be debriefed. So, having only 3 examples in each category is really good."*

6.4.5 Recommendations.

- *Real-Time Feedback*: Some participants discussed the potential value of having real-time feedback. These participants indicated that they often stop simulations, provide immediate feedback on the current performance, and then recommence the simulation. The interruption of simulations is often due to poor non-technical skills. Having the report generated in real time to be used during these interruptions is valuable in providing immediate feedback. P11 stated: *"One of the things I would like to see is this available in real-time, during the simulation. That would help me provide immediate feedback during the simulation if needed."*
- *Provide All Examples*: Although participants preferred only three examples of CLC in each category on the report, most preferred an option to see more examples, including the entire transcript. This could be a searchable or a scrollable feature in the application that generates the report. Participants indicated that this could help them point out and discuss specific examples they may have observed during the simulation or examine communication surrounding a specific event in the simulation. P5 stated: *"I'd probably like to see every example. I would like to see it all. A search bar would be a plus."*

- *Benchmarks*: Several participants indicated that they would like to see a benchmark that is ideal for the percentage of Call-Out receiving Check-Backs. Participants mentioned that they do not currently know what percentage should be reasonably expected during simulations. One participant (P10) specified that the benchmark should not simply be based on the number of Call-Out receiving Check-Back but only record those that are of high quality. He stated: *“I don’t think the number of Call-Outs with Check-Backs is the prime number that you should be looking at. I think it’s more the percentages of good closed-loop communication we should focus on.”*
- *Color and Visual Charts*: Several participants indicated that they preferred the use of color and visual charts in the display. Some participants indicated that words like “Good” or “bad” should be highlighted or colored green and red to assist instructors in identifying those classifications quicker. One participant discussed the value of using different types of charts in the report. P8 stated: *“Something like a pie chart would be helpful. So that I could quickly pull it up and see it. That would be easy. Just like the feedback for the CPR [on manikin software]. You know. It’s very visual, like green and orange. The feedback is quick.”*

6.4.6 Barriers.

- *Cost*: Some participants indicated that cost could be an issue for them being able to use the technology at their organization. They specifically described this as a potential barrier that could result in their organization not purchasing the technology. However, there was little elaboration on what price could be a deciding factor. P7 stated: *“I think if it were something that an institution put into place, I would just use it, but I know that for them it would be a cost factor whether they had it in place or not.”*
- *Timeliness to Receive Results*: There was a concern among some participants regarding the time it would take for the report to be generated. Participants indicated that they have limited time for the simulation and debriefing during these sessions, with one participant (P11) indicating that they are often only provided 5 minutes to debrief. Waiting for the report to be generated could reduce overall debriefing time. P8 stated: *“I think it has to be available as quickly as possible. Timewise, it needs to be easy to pull up so that you can provide that feedback to them.”*
- *Trusting the Accuracy of the Results*: Although participants felt that the report was more accurate than their observations of CLC, there was concern that some instructors may not trust the report’s accuracy. Additionally, overtrust in the report could lead to improper feedback or instructional decisions that could harm learning and student success. Participants stressed the importance of this tool being accurate in recording CLC. P6 stated: *“I think making sure it’s accurate is very important; It’s accuracy and the actual quality of the feedback it’s providing.”*

7 Discussion

AutoCLC represents a step forward in applying LLMs and virtual environments to real-world challenges related to team collaboration, providing a novel method for communication analysis and feedback in team-based simulations. The following subsections discuss the insights derived from our formative study, technical evaluation, and user study of the AutoCLC tool, including design improvements, broader implications and potential applications, limitations, and future directions of our research.

7.1 Enhancing Feedback Practices and Assessment of Team Communication in Simulation Training

Without guidance, most instructors tend to focus their debriefing on technical skills, potentially limiting the learning gains on non-technical skills, including ones related to communication. In prior studies, teams that received non-technical skill debriefing performed better than those that received only technical skill debriefing [32, 75]. Participants in our study saw AutoCLC as a tool that could help motivate them to prioritize CLC more in their training practices.

By comparing AutoCLC approach to current debriefing methods involving CLC, instructors perceived significant improvements in their ability to evaluate and provide targeted feedback on CLC performance. Participants explained that their current methods of evaluating the frequency and quality of CLC were often unstructured and difficult due to the dynamic nature of team communication and overemphasis on technical skills [1]. According to the study participants, AutoCLC could increase their ability to assess CLC during simulation training by providing them with a tool for structured debriefings on the quality and frequency of communications during team-based cardiac arrest simulations.

Participants generally considered the report generated by AutoCLC as usable, with a System Usability Scale score of 76.36, indicating above-average usability. Across the interview sessions, participants regarded the report's ability to highlight both effective and ineffective communication instances as a valuable contribution to the debriefing process. For instance, in prior studies, Hughes et al. [40] used a video review process after simulation training to manually assess the quality and accuracy of CLC. However, this time-consuming process could potentially impede timely feedback and increase the instructor's cognitive load. By contrast, the AutoCLC report was perceived as a more efficient tool for providing structured debriefing on the quality and frequency of communication.

AutoCLC as a Supplement Rather than a Replacement for Instructor Judgment. While our current focus is not on real-time human-AI collaboration, we believe that our work provides important insights for thoughtfully integrating AI tools into educational and professional learning settings. During our user study with the medical instructors, they interacted with AI as they reviewed the AI-generated reports. The participants provided valuable insights on the role of AI in an instructional context, including the preference for using AI-generated reports as reference points rather than definitive assessments and maintaining control over how to integrate these insights into their teaching. This approach supports instructor decision-making and meaning-making without diminishing their essential role in complex social contexts. Rudolph et al. [69] emphasize the importance of "debriefing with good judgment," where instructors provide explicit feedback enabling comprehensive exploration of key learning insights. The shift toward "good judgment" debriefing combines data-informed feedback with inquiry [51]. Providing data-driven CLC details alongside comprehensive discussion may foster this type of deeper learning. The insights gained from this study provide important considerations for the design of future systems that may incorporate more direct human-AI collaboration.

7.2 Design Implications: Suggestions and Potential Barriers to Adoption

Incorporating the option to view all examples of CLC reported by AutoCLC emerged as a key feature desired by participants. Currently, the automated CLC report provides a limited number of examples that best represent each of the three defined CLC categories (i.e., good, poor, unresponded). However, participants expressed that the ability to access and search all auto-generated CLC examples would enable them to locate specific events they observed during the simulation and link them to possible causes and teachable moments (e.g., knowledge gap, high cognitive load). This capability would allow instructors to provide more targeted feedback, as noted in prior work [44, 45, 54], to optimize trainee learning and facilitate more comprehensive debriefing discussions.

Participants also emphasized the importance of visual aids, such as color-coding and visual charts, to enhance the report's interface readability and facilitate quick identification of communication patterns. Aligning with the principles of efficient information processing and decision-making, visualizations such as graphs and charts have been recognized as beneficial tools for enhancing understanding and communication in educational and scientific contexts [7]. The use of these visual aids could help facilitate better data interpretation and deciding on teachable moments [11, 28, 61]. Although research on the use of charts and graphs for feedback on team performance in healthcare is limited, studies have shown that visual aids can facilitate students' learning outcomes and skills acquisition [29, 56, 87].

Despite the promising suggestions for optimizing the AI-generated report, potential barriers to adoption were identified. One barrier involves the timeframe required for report generation. Some participants were concerned

that it may take too long for the report to be generated, which could impede the workflow of simulation training sessions, especially in cases where there is limited time for debriefing. While the data processing outlined in this work involves several manual steps, these can be automated with sufficient software engineering efforts. It is envisioned that future iterations of the AutoCLC tool will complete the analysis and report generation in under 3 minutes, thereby providing instructors with just-in-time metrics and examples to provide immediate feedback to learners in debriefing sessions.

Although participants themselves believed that the reports generated by AutoCLC were accurate, they were concerned that some colleagues may not. This could inhibit the adoption of the technology into simulation debriefing sessions, e.g., across schools or institutions. Trust in the AutoCLC accuracy is essential as it can influence users' perceptions of its utility and benefits in improving the debriefing experience. This requires careful consideration of risks related to over-reliance on AI, including hallucinations, overconfidence in inaccurate predictions, or propagation of subtle biases embedded in training data. Over-reliance on AI occurs when users accept AI-generated recommendations without judgment, leading to errors in task performance and/or decision-making. This is particularly problematic in simulation training where instructors may develop "automation bias" [36], emphasizing the need for balanced trust and doubt. In the CLC assessment, this could manifest as trainees and/or instructors accepting AutoCLC's categorization of the team's communication as "poor" based on surface-level linguistic patterns, while missing nuanced contextual factors. For example, when a call-out states "Mark, stop CPR and check pulse," and the check-back reports "Pulse is 120," an LLM may incorrectly identify an incomplete communication loop because "stop CPR" was not explicitly acknowledged. However, the pulse reading itself confirms that CPR was stopped, as this is a clinical prerequisite for pulse assessment.

Another potential issue is the propagation of subtle biases embedded in training data. When training data overrepresent communication patterns from specific institutions, professional hierarchies, or demographic groups, the system may systematically penalize legitimate cultural and/or institutional variations in communication styles and protocols. These subtle algorithmic biases become particularly insidious because they appear objective and evidence-based while perpetuating existing inequities in healthcare and medical education assessment [5, 84].

AutoCLC may also generate plausible but incorrect assessments—confidently identifying nonexistent communication breakdowns—that become difficult for busy instructors to detect when presented with detailed, convincing CLC data. To mitigate these risks, we believe that AutoCLC or similar tools must be implemented as decision-support tools with regular bias auditing across diverse trainee populations, and clear guidelines prioritizing human judgment in ambiguous cases.

One effective strategy to enhance trust in AI systems in education is to prioritize explainability [8, 48]. Through the implementation of **Explainable AI (XAI)** techniques, educators and trainees can gain insights into how AI systems reach specific recommendations or decisions. This transparency enables users to comprehend the reasoning behind AI-generated outputs and help address the issues of potential bias and hallucinations. Furthermore, using the human-in-the-loop design frameworks [57] and offering explanations for AI-driven actions can empower users to make well-informed decisions based on AI recommendations [48]. For example, reasoning models (e.g., o1 from OpenAI [2]), which employ techniques like Chain-of-Thought and Tree-of-Thoughts [78] to generate intermediate steps in their decision-making process, are promising for enhancing XAI in educational contexts by making their multi-step reasoning visible to users. However, the tradeoffs may include computational cost, being overly verbose, and generating plausible yet incorrect reasoning paths. These reasoning models can be tested in future iterations of AutoCLC to help explain to instructors and trainees the rationale behind the good and poor categorizations of CLC instances.

7.3 Broader Implications and Potential Applications of AutoCLC

In terms of broader implications, AutoCLC's fundamental contribution lies in its automated approach to analyzing structured communication patterns using LLMs, with particular focus on the two-part structure: initial

message/request → acknowledgment/confirmation. This structure represents a pattern that appears across various high-reliability domains. For example, in aviation safety, this pattern manifests when a pilot provides navigation coordinates, the co-pilot independently cross-checks them, and both confirm navigational accuracy before proceeding. The methodology we developed for detecting and analyzing these communication patterns has broad applicability beyond healthcare. Many organizations demanding high-reliability have adopted some form of CLC as part of effective communication practices within crew resource management [68, 82]. Beyond that, in remote or hybrid team meetings, the system could assess participation patterns and turn-taking behaviors. In educational settings that use team-based learning, it could evaluate argument quality by distinguishing between evidence-based and unsupported claims. For recorded online meetings, it could track acknowledgment of contributions and generate detailed interaction reports [19]. AutoCLC's foundational approach can be adapted to create automated toolchains for communication analysis across different contexts.

Importantly, evaluation of AutoCLC across diverse medical scenarios is crucial for demonstrating broader utility. Our focus on cardiac arrest scenarios, where CLC is both critical and frequent, is a strategic first step in developing and evaluating our approach. ACLS protocols offer standardized team structures and communication requirements across institutions, making them an ideal test ground for refining our methods. Once reliability and initial validity are established in this controlled context, we will systematically expand to other medical scenarios such as trauma resuscitation, operating room procedures, and emergency department handoffs, which all practice CLC protocol. Our ongoing larger-scale empirical studies ultimately aim to investigate how AutoCLC affects group dynamics, learning gains, and transfer to real-world clinical settings.

However, re-applying AutoCLC to new contexts presents both opportunities and challenges. Although the fundamental CLC structure remains consistent, various clinical scenarios differ significantly in their specific vocabulary (e.g., names, medications, procedures), workflows, and nuanced communication patterns. Effective generalization thus requires targeted fine-tuning of the underlying language model to align with these domain-specific characteristics. Additionally, broader deployment into real-world clinical environments introduces further complexities, such as background noise, overlapping speech, and inconsistent recording quality, which can impact transcription accuracy and NLP performance. Moreover, the applicability of AutoCLC in less structured or emergent scenarios, where standard communication patterns may deviate significantly, should be systematically evaluated to ensure robust and adaptive performance. Addressing these challenges will necessitate rigorous testing, iterative adaptation, and enhancements to acoustic robustness. Furthermore, extending AutoCLC to multilingual environments or hospitals outside the United States will involve retraining with relevant datasets and culturally appropriate adjustments to prompts and analytical frameworks. To address these challenges and enhance both scientific and social impact, we have developed a strategic plan for future directions. This includes domain-specific fine-tuning, multilingual model development, and improvements in acoustic robustness. Additionally, we have made our code publicly available, enabling other practitioners and institutions to adapt our analytical pipeline to their specific contexts, potentially scaling the assessment of CLC beyond simulated clinical settings.

7.4 Limitations and Future Work

While this study serves as foundational work, we acknowledge several limitations of AutoCLC. Currently, the report generated by AutoCLC is primarily designed for instructors, who oversee and review the entire team's performance during debriefing sessions. Future work will focus on developing more detailed, individualized reports for each trainee, incorporating CLC-related performance metrics, tailored suggestions for improvement, and team-level CLC feedback. In addition, the transcription process from recordings to texts is automated after uploading the audio recording, with limited manual corrections applied to speaker identifications. However, we acknowledge the need for further software engineering to fully automate transcript processing, which should progress in tandem with the ongoing evolution of machine learning and LLM technologies. LLMs also face notable limitations in their current forms, particularly in contextual understanding, requiring further refinement.

Nonetheless, continuous advancements in machine learning, hardware, and LLM technology are expected to significantly enhance the performance of AutoCLC. Finally, we aim to deploy this system in real clinical scenarios. Currently, the AutoCLC pipeline includes steps requiring human intervention, such as file transfer and format conversion. Integrating alternative tools or APIs could streamline these processes and facilitate future deployment in real-world environments.

8 Conclusion

This study developed and evaluated AutoCLC, an AI-driven system leveraging NLP to assess and provide learners feedback on CLC in healthcare team-based simulation training. Motivated by formative study insights from ACLS instructors who report a paucity of objective and accurate CLC assessment tools, AutoCLC's core component is an LLM-powered pipeline that efficiently and accurately detects, computes, and classifies CLC instances. AutoCLC generates reports that both quantify CLC frequency and classify and provide examples of good, poor, and unresponded CLC—both features that instructors found valuable. A multi-part evaluation involving technical assessments and CLC instance ratings demonstrated AutoCLC's efficacy, compared to human annotators. In the user study, instructors positively perceived AutoCLC's potential to enhance communication analysis, feedback, and debriefing quality. These insights into future development are valuable not only for improving AutoCLC but also for designing AI-driven tools in other areas. Overall, this foundational work indicates that AI-driven CLC assessment tools have the potential as valuable assets in healthcare simulation training, but careful consideration must be given to their usability and consistency to increase the likelihood of future deployment in more challenging real-world clinical scenarios.

References

- [1] American Heart Association. 2023. *The ACLS Instructor Manual*. American Heart Association. This product has been updated with new science from the 2020 Guidelines for CPR and ECC and replaces the previous version (15-1003). A digital version of this product is also available: 20–3101.
- [2] OpenAI. 2023. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. Retrieved May 24, 2024 from <https://openai.com/index/gpt-4/>
- [3] SciPy. 2024. scipy.stats.mannwhitneyu. Retrieved May 24, 2024 from <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html>
- [4] Maria Celine Aas. 2024. The effectiveness of closed-loop communication training for healthcare providers on use, medical errors, and task performance: A systematic review. Retrieved from <https://munin.uit.no/handle/10037/33922>
- [5] Alaa Abd-Alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, Padraig Mark Healy, Syed Latifi, Sarah Aziz, Rafat Damseh, Sadam Alabed Alrazak, and Javaid Sheikh. 2023. Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Medical Education* 9, 1 (2023), e48291.
- [6] Ibrahim Abd El-Shafy, Jennifer Delgado, Meredith Akerman, Francesca Bullaro, Nathan A. M. Christopherson, and Jose M. Prince. 2018. Closed-loop communication improves task completion in pediatric trauma resuscitation. *Journal of Surgical Education* 75, 1 (2018), 58–64.
- [7] Sakinah S. J. Alhadad. 2018. Visualizing data to support judgement, inference, and decision making in learning analytics: Insights from cognitive psychology and visualization science. *Journal of Learning Analytics* 5, 2 (2018), 60–85.
- [8] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bannetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.
- [9] Caitlin H. Bailey and Julie D. Gesch. 2023. Team strategies and dynamics during resuscitation. *Emergency Medicine Clinics of North America* 41, 3 (Aug. 2023), 587–600. DOI : <https://doi.org/10.1016/j.emc.2023.03.011>
- [10] David P. Baker, Rachel Day, and Eduardo Salas. 2006. Teamwork as an essential component of high-reliability organizations. *Health Services Research* 41, 4p2 (2006), 1576–1598.
- [11] Seyyed Kazem Banihashem, Omid Noroozi, Stan van Ginkel, Leah P. Macfadyen, and Harm J. A. Biemans. 2022. A systematic review of the role of learning analytics in enhancing feedback practices in higher education. *Educational Research Review* 37 (2022), 100489.
- [12] Avneesh Bhangu, Lowyl Notario, Ruxandra Pinto, Dylan Pannell, Will Thomas-Boaz, Corey Freedman, Homer Tien, Avery Nathens, and Luis Da Luz. 2022. Closed loop communication in the trauma bay: Identifying opportunities for team performance improvement through a video review analysis. *Canadian Journal of Emergency Medicine* 24 (Apr. 2022), 419–425. DOI : <https://doi.org/10.1007/s43678-022-00295-z>

- [13] Tiffany M. Bisbey, Molly P. Kilcullen, and Eduardo Salas. 2021. Maintaining resilience in times of crisis: Insights from high-reliability organizations. *Oxford Research Encyclopedia of Politics* (2021).
- [14] Clint A. Bowers, Florian Jentsch, Eduardo Salas, and Curt C. Braun. 1998. Analyzing communication sequences for team training needs assessment. *Human Factors* 40, 4 (1998), 672–679.
- [15] Marie-Stéphanie Bracq, Estelle Michinov, and Pierre Jannin. 2018. Virtual reality simulation in nontechnical skills training for healthcare professionals: A systematic review. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare* 14 (Dec. 2018), 1. DOI: <https://doi.org/10.1097/SIH.0000000000000347>
- [16] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.
- [17] C. S. Burke, Eduardo Salas, K. Wilson-Donnelly, and H. Priest. 2004. How to turn a team of experts into an expert medical team: Guidance from the aviation and military communities. *Quality and Safety in Health Care* 13, Suppl 1 (2004), i96–i104.
- [18] Alissa S. Chen, Bernice Yau, Lee Revere, and Jennifer Swails. 2019. Implementation, evaluation, and outcome of TeamSTEPPS in interprofessional education: A scoping review. *Journal of Interprofessional Care* (2019).
- [19] Xinyue Chen, Shuo Li, Shipeng Liu, Robin Fowler, and Xu Wang. 2023. Meetscript: Designing transcript-based interactions to support active participation in group video meetings. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–32.
- [20] Adam Cheng, David J. Magid, Marc Auerbach, Farhan Bhanji, Blair L. Bigham, Audrey L. Blewer, Katie N. Dainty, Emily Diederich, Yiqun Lin, Marion Leary, et al. 2020. Part 6: Resuscitation education science: 2020 American heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation* 142, 16 Suppl 2 (2020), S551–S579.
- [21] Simon Cooper and Alan Wakelam. 1999. Leadership of resuscitation teams: ‘Lighthouse Leadership’. *Resuscitation* 42, 1 (1999), 27–45.
- [22] Gianpaolo Coro, Serena Bardelli, Armando Cuttano, and Nicoletta Fossati. 2022. Automatic detection of potentially ineffective verbal communication for training through simulation in neonatology. *Education and Information Technologies* 28 (Mar. 2022), 9181–9203. DOI: <https://doi.org/10.1007/s10639-022-11000-z>
- [23] Kris Demuynck and Tom Laureys. 2002. A comparison of different approaches to automatic speech segmentation. *Speech Communication* 36, 4 (2002), 313–331. DOI: [https://doi.org/10.1016/S0167-6393\(01\)00088-3](https://doi.org/10.1016/S0167-6393(01)00088-3)
- [24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), 4171–4186.
- [25] Maria Carmen G. Diaz and Kimberly Dawson. 2020. Impact of simulation-based closed-loop communication training on medical errors in a pediatric emergency department. *American Journal of Medical Quality* 35, 6 (2020), 474–478.
- [26] Tracy M. Dodson, Janet M. Reed, and Kimberly Cleveland. 2023. Exploring undergraduate nursing students’ ineffective communication behaviors in simulation: A thematic analysis. *Teaching and Learning in Nursing* 18, 4 (2023), 480–485. DOI: <https://doi.org/10.1016/j.teln.2023.06.019>
- [27] Andrew J. Doorey, Zoltan G. Turi, Elizabeth H. Lazzara, Erika G. Mendoza, Kirk N. Garratt, and William S. Weintraub. 2020. Safety gaps in medical team communication: Results of quality improvement efforts in a cardiac catheterization laboratory. *Catheterization and Cardiovascular Interventions* 95, 1 (Jan. 2020), 136–144. DOI: <https://doi.org/10.1002/ccd.28298>
- [28] Karin Eberhard. 2023. The effects of visualization on judgment and decision-making: A systematic literature review. *Management Review Quarterly* 73, 1 (2023), 167–214.
- [29] Vanessa Echeverria, Roberto Martinez-Maldonado, Lixiang Yan, Linxuan Zhao, Gloria Fernandez-Nieto, Dragan Gašević, and Simon Buckingham Shum. 2022. HuCETA: A framework for human-centered embodied teamwork analytics. *IEEE Pervasive Computing* 22, 1 (2022), 39–49.
- [30] Debora Field, Simon Worgan, Nick Webb, M. Hepple, and Yorick Wilks. 2008. Automatic Induction of Dialogue Structure from the Companions Dialogue Corpus. Retrieved from https://www.researchgate.net/publication/262485912_Automatic_Induction_of_Dialogue_Structure_from_the_Companions_Dialogue_Corpus
- [31] John M. Flach, John S. Carroll, Marvin J. Dainoff, and W. Ian Hamilton. 2015. Striving for safety: Communicating and deciding in sociotechnical systems. *Ergonomics* 58, 4 (2015), 615–634.
- [32] A. L. Garden, D. M. Le Fevre, H. L. Waddington, and J. M. Weller. 2015. Debriefing after simulation-based non-technical skill training in healthcare: A systematic review of effective practice. *Anaesthesia and Intensive Care* 43, 3 (2015), 300–308.
- [33] Ehsan Garosi, Reza Kalantari, Ahmad Farahani, Mojgan Zuaktafi, Esmail Roknabadi, and Ehsan Bakhshi. 2019. Concerns about verbal communication in the operating room: A field study. *Human Factors* 62 (Jul. 2019), 001872081985827. DOI: <https://doi.org/10.1177/0018720819858274>
- [34] Liane Ginsburg and Lorna Bain. 2017. The evaluation of a multifaceted intervention to promote “speaking up” and strengthen interprofessional teamwork climate perceptions. *Journal of Interprofessional Care* 31, 2 (Mar. 2017), 207–217. DOI: <https://doi.org/10.1080/13561820.2016.1249280>
- [35] K. Gjæraa, T. P. Møller, and D. Østergaard. 2014. Efficacy of simulation-based trauma team training of non-technical skills. A systematic review. *Acta Anaesthesiologica Scandinavica* 58, 7 (Aug. 2014), 775–787. DOI: <https://doi.org/10.1111/aas.12336>
- [36] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. 2012. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127.

- [37] Oren Guttman, Elizabeth Lazzara, Joseph Keebler, Kristen Webster, Logan Gisick, and Anthony Baker. 2021. Dissecting communication barriers in healthcare: A path to enhancing communication resiliency, reliability, and patient safety. *Journal of Patient Safety* 17, 8 (Dec. 2021), e1465–e1471. DOI: <https://doi.org/10.1097/PTS.0000000000000541>
- [38] Maria Härgestam, Marie Lindkvist, Christine Brulin, Maritha Jacobsson, and Magnus Hultin. 2013. Communication in interdisciplinary teams: Exploring closed-loop communication during in situ trauma team training. *BMJ Open* 3, 10 (2013), e003525.
- [39] Pei-Yun Hsueh. 2008. Audio-based unsupervised segmentation of multiparty dialogue. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 5049–5052.
- [40] Kate E. Hughes, Patrick G. Hughes, Thomas Cahir, Jennifer Plitt, Vivienne Ng, Edward Bedrick, and Rami A. Ahmed. 2020. Advanced closed-loop communication training: The blindfolded resuscitation. *BMJ Simulation & Technology Enhanced Learning* 6, 4 (Jul. 2020), 235–238. DOI: <https://doi.org/10.1136/bmjstel-2019-000498>
- [41] Sabina Hunziker, Anna C. Johansson, Franziska Tschan, Norbert K. Semmer, Laura Rock, Michael D. Howell, and Stephan Marsch. 2011. Teamwork and leadership in cardiopulmonary resuscitation. *Journal of the American College of Cardiology* 57, 24 (Jun. 2011), 2381–2388. DOI: <https://doi.org/10.1016/j.jacc.2011.03.017>
- [42] Maria Härgestam, Marie Lindkvist, Christine Brulin, Maritha Jacobsson, and Magnus Hultin. 2013. Communication in interdisciplinary teams: Exploring closed-loop communication during in situ trauma team training. *BMJ Open* 3, 10 (2013), e003525. DOI: <https://doi.org/10.1136/bmjopen-2013-003525>
- [43] Toshiki Kawai, Hiroki Chino, Kohei Matsumura, Haruo Noma, Osamu Sugiyama, Manabu Shimoto, Shigeru Ohtsuru, and Tomohiro Kuroda. 2018. What’s going on? A disaster in disaster communication between medical teams at hospitals. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW ’18 Companion)*. ACM, New York, NY, 265–268. DOI: <https://doi.org/10.1145/3272973.3274071>
- [44] Nathanael L. Keiser and Winfred Arthur Jr. 2021. A meta-analysis of the effectiveness of the after-action review (or debrief) and factors that influence its effectiveness. *Journal of Applied Psychology* 106, 7 (2021), 1007.
- [45] Nathanael L. Keiser and Winfred Arthur Jr. 2022. A meta-analysis of task and training characteristics that contribute to or attenuate the effectiveness of the after-action review (or debrief). *Journal of Business and Psychology* 37, 5 (2022), 953–976.
- [46] Mike Kentros, Manos Kamarianakis, Michael Cole, Vitaliy Popov, Antonis Protopsaltis, and George Papagiannakis. 2025. Non-linear, team-based VR training for cardiac arrest care with enhanced CRM toolkit. *Eurographics Technical Report Series* (2025).
- [47] Mike Kentros, Manos Kamarianakis, Michael Cole, Vitaliy Popov, Antonis Protopsaltis, and George Papagiannakis. 2025. Non-linear, team-based VR Training for cardiac arrest care with enhanced CRM Toolkit. In *Eurographics 2025—Short Papers*. Duygu Ceylan and Tzu-Mao Li (Eds.), The Eurographics Association. DOI: <https://doi.org/10.2312/egs.20251054>
- [48] Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. 2022. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence* 3 (2022), 100074.
- [49] Heidi B. King, James Battles, David P. Baker, Alexander Alonso, Eduardo Salas, John Webster, Lauren Toomey, and Mary Salisbury. 2008. *TeamSTEPPS™: Team Strategies and Tools to Enhance Performance and Patient Safety*. Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK43686/>
- [50] Linda Kohn, Janet Corrigan, and Molla Donaldson. 2000. *To Err Is Human: Building a Safer Health System*. National Academies Press. DOI: <https://doi.org/10.17226/9728>
- [51] Michaela Kolbe, Bastian Grande, Nale Lehmann-Willenbrock, and Julia Carolin Seelandt. 2023. Helping healthcare teams to debrief effectively: Associations of debriefers’ actions and participants’ reflections during team debriefings. *BMJ Quality and Safety* 32, 3 (2023), 160–172.
- [52] Jim Lara. 2023. Achieve Perfection with Simplified, Closed-Loop Communication Principles. Retrieved from <https://www.graystoneadvisors.com/blog/closed-loop-communication/>
- [53] Virginia LeBaron, Tabor Flickinger, David Ling, Hansung Lee, James Edwards, Anant Tewari, Zhiyuan Wang, and Laura E. Barnes. 2023. Feasibility and acceptability testing of CommSense: A novel communication technology to enhance health equity in clinician–patient interactions. *Digital Health* 9 (Jan. 2023), 20552076231184991. DOI: <https://doi.org/10.1177/20552076231184991>
- [54] Myung-Nam Lee, Shin-Jeong Kim, Kyung-Ah Kang, and Sunghee Kim. 2020. Comparing the learning effects of debriefing modalities for the care of premature infants. *Nursing and Health Sciences* 22, 2 (2020), 243–253.
- [55] James R. Lewis. 2018. The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction* 34, 7 (2018), 577–590.
- [56] Roberto Martinez-Maldonado, Vanessa Echeverria, Gloria Fernandez-Nieto, Lixiang Yan, Linxuan Zhao, Riordan Alfredo, Xinyu Li, Samantha Dix, Hollie Jaggard, Rosie Wotherspoon, et al. 2023. Lessons learnt from a multimodal learning analytics deployment in-the-wild. *ACM Transactions on Computer-Human Interaction* 31, 1 (2023), 1–41.
- [57] Bahar Memarian and Tenzin Doleck. 2024. Human-in-the-loop in artificial intelligence in education: A review and entity-relationship (ER) analysis. *Computers in Human Behavior: Artificial Humans* 2, 1 (2024), 100053.

- [58] T. Daniel Midgley. 2003. Discourse chunking: A tool in dialogue act tagging. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 2 (ACL '03)*. Association for Computational Linguistics, USA, 58–63. DOI: <https://doi.org/10.3115/1075178.1075186>
- [59] Ida Molin, Liselott Fornander, and Peter Berggren. 2021. An analysis of trauma team communication using the verbal response mode taxonomy. *Surgery* 170, 6 (2021), 1849–1854. DOI: <https://doi.org/10.1016/j.surg.2021.05.043>
- [60] Dmitry Mozhey, José Agustín González Almeida, María del Cristo, Adrián de Ganzo, Alejandro U. Gómez Correa, and Javier Almenara de Luz. 2024. Using the closed loop communication system to prevent accidents caused by human errors. *Journal of Maritime Research* 21, 1 (2024), 361–369.
- [61] Julius Murumba. 2022. Learning Analytics and Educational Data Visualization in the Digital Era. *Journal of Innovations in Data Science and Big Data Management* 1, 3 (2022). Retrieved from https://www.researchgate.net/publication/359093879_Learning_Analytics_and_Educational_Data_Visualization_in_the_Digital_Era
- [62] Brahmajee K. Nallamothu, Timothy C. Guetterman, Molly Harrod, Joan E. Kellenberg, Jessica L. Lehrich, Steven L. Kronick, Sarah L. Krein, Theodore J. Iwashyna, Sanjay Saint, and Paul S. Chan. 2018. How do resuscitation teams at top-performing hospitals for in-hospital cardiac arrest succeed? A qualitative study. *Circulation* 138, 2 (2018), 154–163.
- [63] Inc. Otter.ai. 2024. Otter.ai. Retrieved May 24, 2024 from <https://otter.ai/>
- [64] Vitaliy Popov, Natalie Mateju, Caris Jeske, and Kadriye O. Lewis. 2024. Metaverse-based simulation: A scoping review of charting medical education over the last two decades in the lens of the ‘marvelous medical education machine. *Annals of Medicine* 56, 1 (2024), 2424450.
- [65] Enola Proctor, Hiie Silmere, Ramesh Raghavan, Peter Hovmand, Greg Aarons, Alicia Bunger, Richard Griffey, and Melissa Hensley. 2011. Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health* 38 (2011), 65–76.
- [66] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.
- [67] Daniel B. Raemer, Michaela Kolbe, Rebecca D. Minehart, Jenny W. Rudolph, and May C. M. Pian-Smith. 2016. Improving anesthesiologists’ ability to speak up in the operating room: A randomized controlled experiment of a simulation-based intervention and a qualitative analysis of hurdles and enablers. *Academic Medicine* 91, 4 (2016), 530–539.
- [68] F. Eric Robinson, David Grimm, Dain Horning, Jamie C. Gorman, Jennifer Winner, and Christopher W. Wiese. 2024. An automated content-based measure of closed loop communication among critical care air transport teams. *Military Psychology* 37 (2024), 1–7.
- [69] Jenny W. Rudolph, Robert Simon, Peter Rivard, Ronald L. Dufresne, and Daniel B. Raemer. 2007. Debriefing with good judgment: combining rigorous feedback with genuine inquiry. *Anesthesiology Clinics* 25, 2 (2007), 361–376.
- [70] Grace V. Ryan, Shauna Callaghan, Anthony Rafferty, Mary F. Higgins, Eleni Mangina, and Fionnuala McAuliffe. 2022. Learning outcomes of immersive technologies in health care student education: Systematic review of the literature. *Journal of Medical Internet Research* 24, 2 (Feb. 2022), e30082. DOI: <https://doi.org/10.2196/30082>
- [71] Irin Salik and John Ashurst. 2020. Closed Loop Communication Training in Medical Simulation. Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK549899/>
- [72] Antonio Casal Sanchez, Laura Calviño Pereira, Jose Flores Arias, Luis Sanchez Santos, and Antonio Iglesias Vazquez. 2014. Advanced medical simulation as a tool to assess the leadership and teamwork in emergencies. *Resuscitation* 85 (May. 2014), S72. DOI: <https://doi.org/10.1016/j.resuscitation.2014.03.180>
- [73] Stayko Spiridonov. 2017. Causes for ineffective communication between medical specialists. *Journal of IMAB - Annual Proceeding (Scientific Papers)* 23 (Jul. 2017), 1623–1626. DOI: <https://doi.org/10.5272/jimab.2017233.1623>
- [74] Michael St. Pierre, Gesine Hofinger, and Cornelius Buerschaper. 2008. *Crisis Management in Acute Care Settings: Human Factors and Team Psychology in a High Stakes Environment*. Springer.
- [75] Scott I. Tannenbaum and Christopher P. Cerasoli. 2013. Do team and individual debriefs enhance performance? A meta-analysis. *Human Factors* 55, 1 (2013), 231–245.
- [76] Roberto Trevi, Chiappinotto, Stefania Alvisa Palese, and Alessandro Galazzi. 2024. Virtual reality for cardiopulmonary resuscitation healthcare professionals training: A systematic review. *Journal of Medical Systems* 48, (May. 2024), 50. DOI: <https://doi.org/10.1007/s10916-024-02063-1>
- [77] C. Vinotheni and S. Lakshmana Pandian. 2021. Deep learning-based text segmentation in NLP using fast recurrent neural network with Bi-LSTM. *Advances in Parallel Computing* 38 (2021), 87–93. DOI: <https://doi.org/10.3233/APC210017>
- [78] Jason Wei, Xuezhong Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed. H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Curran Associates Inc., Red Hook, NY, Article 1800, 1–14.
- [79] Bryan J. Weiner, Cara C. Lewis, Cameo Stanick, Byron J. Powell, Caitlin N. Dorsey, Alecia S. Clary, Marcella H. Boynton, and Heather Halko. 2017. Psychometric assessment of three newly developed implementation outcome measures. *Implementation Science* 12 (2017), 1–12.

- [80] Myron L. Weisfeldt and Lance B. Becker. 2002. Resuscitation after cardiac arrest. *JAMA* 288, 23 (Dec. 2002), 3035. DOI : <https://doi.org/10.1001/jama.288.23.3035>
- [81] Jennifer Weller, Matt Boyd, and David Cumin. 2014. Teams, tribes and patient safety: Overcoming barriers to effective teamwork in healthcare. *Postgraduate Medical Journal* 90, 1061 (2014), 149–154.
- [82] Jennifer Winner, Jayde King, Jamie Gorman, and David Grimm. 2022. Team coordination dynamics measurement in enroute care training: Defining requirements and usability study. In *Proceedings of the 11th International Symposium on Human Factors and Ergonomics in Health Care*. SAGE Publications, Los Angeles, CA, 21–25.
- [83] Nicole Yamada, Janene Fuerch, and Louis Halamek. 2015. Impact of standardized communication techniques on errors during simulated neonatal resuscitation. *American Journal of Perinatology* 33, 04 (Oct. 2015), 385–392. DOI : <https://doi.org/10.1055/s-0035-1565997>
- [84] Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A. Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W. Bates, Raja-Elie E. Abdulnour, et al. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. *The Lancet Digital Health* 6, 1 (2024), e12–e22.
- [85] Zhan Zhang, Aleksandra Sarcevic, and Claus Bossen. 2017. Constructing common information spaces across distributed emergency medical teams. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, 934–947. DOI : <https://doi.org/10.1145/2998181.2998328>
- [86] Linxuan Zhao, Dragan Gašević, Zachari Swiecki, Yuheng Li, Jionghao Lin, Lele Sha, Lixiang Yan, Riordan Alfredo, Xinyu Li, and Roberto Martinez-Maldonado. 2024. Towards automated transcribing and coding of embodied teamwork communication through multimodal learning analytics. *British Journal of Educational Technology* 55, 4 (2024), 1673–1702.
- [87] Linxuan Zhao, Zachari Swiecki, Dragan Gasevic, Lixiang Yan, Samantha Dix, Hollie Jaggard, Rosie Wotherspoon, Abra Osborne, Xinyu Li, Riordan Alfredo, et al. 2023. METS: Multimodal learning analytics of embodied teamwork learning. In *13th International Learning Analytics and Knowledge Conference (LAK '23)*, 186–196.
- [88] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. arXiv:2303.18223. Retrieved from <https://arxiv.org/abs/2303.18223>

Appendix A

Closed-Loop Communication (CLC) is the process of acknowledging the receipt of information and clarifying with the sender of the communicated message that the information received is the same as the original, intended information. In CLC, the person receiving instruction or information repeats it to ensure that the message is understood correctly and accurately. In a medical scenario, the request part or call-out must imply an action that needs to be executed by another individual; the response part or check-back should indicate the name of the medication and its dosage or other quantitative information if they are included in the callout.

The following are examples of CLC and non-CLC in a medical context:

— *Example 1:*

Instructor: “So energy select 200. Going charge when you are ready.”

Student: “Charge, charging.”

In this example, the dialogue is NOT considered a CLC since the student did not mention 200 as the energy setting.

— *Example 2:*

Instructor: “Go ahead and restart compressions.”

Student: “Restarting compressions.”

In this example, the dialogue is considered a CLC since the student mentioned the key action “restarting compressions.”

— *Example 3:*

Instructor: “Okay, Joe. Do we have a set of vitals from Miss Fernandez?”

Student: “Yeah, blood pressure is 70 over 40. Heart rate is 45 beats per minute. Respiratory rate is 16 breaths per minute, and Pulse ox is 92%.”

In this example, the dialogue is NOT considered a CLC since the instructor only inquires about the patient’s condition and does not ask the student to perform additional measurements.

– *Example 4:*

Instructor: “I’m going to intubate the patient.”

Student: “Okay, I will observe you.”

In this example, the dialogue is NOT considered a CLC since the instructor states his next action instead of directing the student to perform any operation.

You will be presented with several lines of dialogue from a medical training session. The components of each line are delimited by semicolons. Each line contains three components. The number after the phrase “index:” is the line index of the dialogue. The phrases after “name:” indicate the speaker. The words after “speech:” are the actual dialogue, which is separated into individual sentences. Each sentence is indexed by a number followed by a dash, with the speech content delimited by a pair of square brackets. Note that adjacent lines in the dialogue could have the same speaker.

You will perform FOUR tasks based on the dialogue:

- (1) Identify the occurrences of any questions or requests (call-outs) in the dialogue. Any questions or requests must correspond to a medical instruction that needs to be explicitly performed by another person. A statement about one’s own actions or an inquiry about the patient’s status should not be included. Return the line and sentence index of all occurrences. Note that a single call-out could contain multiple sentences from the same line. In addition, a single line of dialogue could contain multiple callouts directed at different individuals in separate sentences. Especially focus on sentences that contain the name of an individual.
- (2) For each instance in Part 1, identify whether the question or request is directed, i.e., whether the requester mentions the name or identity of the receiver in the call-out.
- (3) For each instance in Part 1, identify whether the question or request has been responded to by another speaker within the next five lines of dialogue. If so, provide the line and sentence index of the response (check-back). If not, return None as the result. A single check-back could contain multiple sentences from the same line. Note that the response cannot have the same speaker as the requester identified in Part 1.
- (4) For each response identified in Part 3, evaluate how specific the response is. Check whether the response repeats all key information in the call-out, including actions to be performed, medical terminologies, names of medications and their dosages, or other quantitative information. Provide a score of 1–3, with 3 being the most specific. Return 3 if the response repeats all critical information in the request; 2 if it contains part of the information or any action verb; 1 if it repeats no information and merely provides an acknowledgment such as “okay,” “got it,” or “on it.” For example, if the request is “Please prepare one milligram of epinephrine,” the response “Preparing one milligram of epinephrine” would score a 3; the response “Preparing” or “Preparing epinephrine” would score a 2; the response “Okay” or “Got it” would score a 1. Note that the specificity score should be at least 2 if the action in the request is repeated, e.g., “charging” or “preparing.” For unresponded instances identified in part 3, return 0 for this task.

Note that one call-out or check-back could contain multiple sentences from the same line if the sentences imply related instructions. All sentences should appear at most once among all call-outs and check-backs, i.e., a single sentence will never be a response to multiple requests. In addition, any dialogue between the medical professionals and the patients or their relatives is excluded and should not be considered in the tasks above.

Here is an example format you should follow. Do not provide any additional explanation:

– Q: [1 - 3]; D: True; R: [2 - 1], [2 - 2]; S: 3

– Q: [5 - 2], [5 - 4]; D: False; R: [6 - 1]; S: 1

– Q: [8 - 3]; D: True; R: None; S: 0

The example above contains three instances of call-outs, which are [line 1 sentence 3], [line 5 sentence 2, line 5 sentence 4], and [line 8 sentence 3]. The second call-out contains two sentences. Each instance contains four

components indicated by “Q” (question), “D” (directed), “R” (response), and “S” (specificity), which correspond to the four tasks mentioned previously. The first and third instances are directed, while the second one is not. The first and second instances have corresponding responses, which are line 2 sentence 1 and sentence 2, and line 6 sentence 1, respectively. The last instance ([8 - 3]) is unresponded, as indicated by the “None.” Within the two instances with a response, the first response, which contains two sentences ([2 - 1], [2 - 2]), repeats all key information and scores a 3 for specificity, while the second one ([6 - 1]) only provides a vague acknowledgment and scores a 1. For unresponded instances, the fourth task is not applicable, as indicated by the “0” in the last line. Note that all check-backs are within 5 lines of the call-outs.

Clarify and explain your understanding of the instructions.

Execute the instructions to the following dialogue. Only provide the results in the specified format without additional explanation:

[Transcript content]

Received 13 March 2025; revised 17 July 2025; accepted 4 August 2025