2026 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR) | 979-8-3315-4967-1/26/$31.00 ©2026 IEEE | DOI: 10.1109/AIXVR67263.2026.00014

EMSIM: AI-Driven VR Simulation for De-Escalation Training in Law Enforcement

Duo Wang
College of Media
University of Illinois Urbana-Champaign
Champaign, IL, United States
dw36@illinois.edu

Kyrian Liang
Electrical and Computer Engineering
University of Illinois Urbana-Champaign
Champaign, IL, United States
jingxil2@illinois.edu

Wenxuan Song
School of Information Sciences
University of Illinois Urbana-Champaign
Champaign, IL, United States
js129@illinois.edu

Qingxiao Zheng
Computer Science and Engineering
University at Buffalo SUNY
Buffalo, NY, United States
qingxiao@buffalo.edu

Zyra Sheikh
Industrial Enterprise Systems Engineering
University of Illinois Urbana-Champaign
Champaign, IL, United States
zyras2@illinois.edu

Jen Whiting
College of Education
University of Illinois Urbana-Champaign
Champaign, IL, United States
whiting6@illinois.edu

Victor Lu
Electrical and Computer Engineering
University of Illinois Urbana-Champaign
Champaign, IL, United States
vclu2@illinois.edu

Mike Yao
Institute of Communications Research
University of Illinois Urbana-Champaign
Champaign, IL, United States
Mzyao@illinois.edu

Caroline Cao
University of Ottawa
Ottawa, Ontario, Canada
caroline.cao@uottawa.ca

***Abstract*—De-escalation is a critical communication skill for law enforcement, yet traditional training methods are resource-intensive and offer limited feedback. We investigate the use of AI-driven virtual reality (VR) to provide scalable, adaptive training for high-stakes encounters. We developed EMSIM , a scenario-based VR system where trainees engage in open-ended dialogue with an emotionally responsive virtual subject. A rubric-guided evaluator powered by a large language model assesses each utterance and modulates scenario escalation. In a mixed-methods study with 14 trainees and 5 trainers, we analyzed usability, workload, behavioral trajectories, and post-session reflections. Results revealed two user strategies, careful de-escalation and boundary testing, and showed how trainees adapted across trials. Findings highlight how AI-assisted XR can support reflective, feedback-rich communication training.**

***Index Terms*—Empathy Training, Law Enforcement, Scenario-Based Training, Virtual Reality, Large Language Model.**

I. INTRODUCTION

In high-stakes professions such as law enforcement, the ability to de-escalate emotionally volatile situations through effective communication is critical to public safety and trust [1], [2]. Yet current training approaches, such as role-play or scenario-based exercises, often rely on limited instructor availability, lack real-time feedback, and are difficult to scale with consistency [3]–[6]. These limitations have spurred interest in immersive technologies for simulating stressful encounters in repeatable and controllable ways [7], [8].

At the same time, large language models (LLMs) have opened new possibilities for generating dynamic, context-aware virtual characters capable of engaging in open-ended conversations. When embedded into immersive environments, these AI agents can enable rich experiential learning by reacting to trainee input with realistic verbal and emotional responses [9], [10]. However, the integration of LLMs into XR training systems, especially for high-risk, emotionally sensitive communication tasks, remains underexplored. Questions persist about how trainees interpret AI feedback, how realistic and pedagogically useful these experiences feel, and how such systems can be reliably evaluated in practice.

To address these gaps, we introduce **EMSIM** , a scenario-based VR training system designed to support de-escalation skill development. **EMSIM** allows trainees to converse freely with an emotionally distressed virtual subject whose behavioral trajectory is driven by the trainee's conversational tone, content, and timing. A rubric-guided LLM-powered communication evaluator scores each conversation and informs the system's escalation logic. We use empathic communication signals as part of a broader escalation control model that allows trainees to experience the impact of their words in a safe yet consequential training loop.

This work contributes:

(1) A novel integration of LLM-driven conversation scoring into a real-time VR simulation for high-stakes training,

(2) A qualitative and behavioral analysis of how trainees interact with escalation-feedback systems, and

(3) Design implications for feedback-rich, AI-assisted XR training platforms targeting emotionally complex professional skills.

2771-7453/26/$31.00 ©2026 IEEE
DOI 10.1109/AIxVR67263.2026.00014

Authorized licensed use limited to: University of Illinois. Downloaded on April 10,2026 at 21:08:15 UTC from IEEE Xplore. Restrictions apply.

II. Related Works

A. De-escalation in Police Training

Law enforcement agencies have increasingly adopted de-escalation training in response to public concerns about police use of force, with nearly all of the 150 surveyed departments reporting they provide such training to officers [3]. A range of training models has been implemented to prepare officers for these high-stakes scenarios. Notable among them is the Integrating Communications, Assessment, and Tactics (ICAT) model, which emphasizes verbal engagement, tactical repositioning, and threat assessment for situations involving unarmed or less-lethally armed individuals [11], and the Crisis Intervention Team program, which delivers curriculum on mental health awareness, verbal de-escalation, and disposition decision-making [4]. Increasingly, programs have integrated scenario-based training [5], [6], body-worn camera video analysis [12], [13], and trauma-informed relational frameworks to build empathy and cultural sensitivity [6].

Empirical evaluation has demonstrated substantial benefits across multiple domains of police performance and community outcomes, such as reductions in use of force incidents, citizen injuries, and officer injuries [11]. Beyond force reduction, community members who interacted with trained officers reported higher levels of satisfaction [12]. Trained officers were significantly more likely to attempt rapport-building, less likely to use condescending or patronizing tones, and demonstrated greater use of procedural justice principles [12]. Despite demonstrated benefits, de-escalation training remains difficult to scale. Programs typically require 10–40 hours per officer [4], [12] and department-wide implementation can extend over 10 months [11]. Developing curricula alone may take up to 18 months [13]. Delivery is further constrained by the need for intensive human resources: professional actors for scenario enactments [4], expert officers as instructors [13], and coordination among mental health professionals and community stakeholders [6]. These demands create significant bottlenecks, particularly for smaller agencies with limited budgets. Moreover, wide variation in curricula across departments hinders the establishment of standardized models [11], while evidence of training decay necessitates periodic refreshers, which compounds time and staffing demands [11].

B. VR and AI in De-escalation Training

Virtual Reality (VR) and Artificial Intelligence (AI) are the two primary avenues being explored to enhance, extend, or automate de-escalation training by simulating high-stakes scenarios in a controlled manner. The primary benefit of VR is its ability to generate a strong sense of presence, allowing trainees to practice responding to high-risk situations in a safe and reproducible environment [7], [8], [14]. A recent meta-analysis found that VR training significantly outperforms traditional methods in both knowledge acquisition and retention [8]. Crucially, immersive VR can elicit measurable psychophysiological stress responses such as heightened heart rate and reduced HRV, which are key to practicing emotional regulation under pressure [15].



(a) (b)

Fig. 1. In-scenario user interface and outcome feedback in **EmSim** . (a) The initial decisions of the simulation include options to begin the scenario. (b) A sample outcome screen is shown at the end of a training session.

In parallel, AI has been leveraged to create more dynamic interactions. This work has progressed from early systems that relied on rigid, predictable AI, such as multiple-choice menus [16], keyword analysis of speech-to-text input [9], or pre-scripted behavioral trees [17], to more sophisticated approaches incorporating cognitive models that allow virtual agents to maintain internal states and exhibit more varied, human-like behaviors [10]. However, both technologies introduce their own limitations. VR faces persistent challenges with tracking, locomotion, motion sickness, user privacy, and a potential "novelty effect" that may inflate initial positive findings [7], [9]. Similarly, AI agents struggle with the immense complexities of human language, leading to inaccurate speech interpretation [9], stereotypical behavior from rigid conversation trees [10], and overly scripted simulations [18].

This reveals a research gap. First, the application of modern Large Language Models (LLMs) for dynamic, naturalistic conversation remains largely unexplored in the specific domain of police de-escalation training. Although one recent study demonstrated that an LLM-powered system effectively taught general conflict resolution skills [18], many existing systems in law enforcement contexts still rely on older AI techniques or a human-in-the-loop to control the avatar [19], [20]. Second, few systems have integrated both an immersive VR environment and an advanced conversational AI agent (most pair VR with humans, or AI with non-immersive environments) for de-escalation training.

III. System Design and Implementation

A. Key Feature and User Experience Design

The **EmSim** system is designed to immerse trainees in realistic de-escalation scenarios while providing trainers with a dashboard to monitor trainees' performance in real-time. The system offers a complete training cycle that starts from scenario development and ends at post-simulation review. This section describes the key designs and user experience through the system for both trainees and trainers.

1) Pre-Training Scenario Development: Before a training session begins, the trainer can customize the simulation to address particular training needs or skill gaps. For instance, train-

Authorized licensed use limited to: University of Illinois. Downloaded on April 10,2026 at 21:08:15 UTC from IEEE Xplore. Restrictions apply.

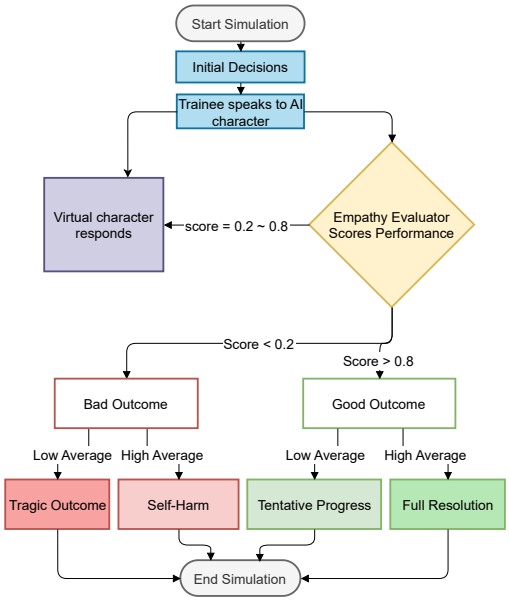


Fig. 2. User experience flow in the **EMSIM** system.

ers can adjust the virtual character's profile, which includes background information, personality, voice, and appearance. Furthermore, the virtual environment itself can be adjusted to align with the scenario context. This flexible customization empowers trainers to create targeted learning experiences and present trainees with diverse empathy challenges.

For the initial implementation and evaluation of **EMSIM** , we selected a scenario involving a suicidal individual reported outside their home with a razor blade. This choice was guided by several factors. Firstly, it presents a lower immediate physical threat to the officer compared to scenarios involving aggression towards others. This allows the trainee to focus on developing de-escalation skills, rather than complex tactical maneuvers. Secondly, to ensure authenticity, the scenario was developed in collaboration with a local police training institute. It was adapted directly from a real-world case they utilize within their role-play training. Finally, this mental health crisis fundamentally demands a high degree of empathetic engagement from the responder, making it an ideal context for **EMSIM** .

2) Scenario-Based De-escalation Training: Once the scenario is finalized, the trainee enters the simulation. We implemented several phases to mimic the progression of a real-world encounter, as shown in Figure 2 and below:

Initial Decisions. Upon arriving at the scene, the trainee makes several crucial initial decisions. These choices involve tactical considerations such as where to park their vehicle and whether to use emergency lights and sirens. These decisions directly influence the virtual character's initial emotional state and receptiveness, allowing trainees to immediately observe how their preliminary approach can set the tone for the entire interaction.

Conversation Loop. Following the initial choices, the core of the simulation involves an interactive conversation loop where the trainee engages verbally with the virtual character to gather information and de-escalate the situation. In each round of conversation, the trainees' input is analyzed in real-time to generate an escalation management score that dynamically influences the virtual character's state and responses. If the trainee is consistently evaluated as highly empathic, the character may become more cooperative. Conversely, interactions lacking empathy can cause the character to become more agitated and hostile.

Scenario Endings. The simulation concludes dynamically based on the trainee's performance throughout the simulation. If the trainee demonstrates strong and effective de-escalation skills, a positive outcome is triggered. On the other hand, if the trainee struggle to make significant progress, triggering neither a positive nor a negative outcome over an extended period, a neutral ending is triggered. The system determines the specific ending by considering both the performance score of the final exchange and the trainee's cumulative average across the simulation.

Read-time Feedback. While the trainee is engaged in the scenario, the trainer can monitor the session in real-time using the Performance Dashboard, allowing them to see the trainee's empathy score as it fluctuates with each exchange. Additionally, the trainer can see a running transcript of the conversation between the trainee and the virtual character. This capability allows trainers to assess performance as it happens, understand the trainee's decision-making process, and identify key moments for later discussion.

3) Post-Training Reflection: Once the simulation concludes and the ending is revealed, both the trainee and the trainer can engage in a detailed post-training review on the Performance Dashboard, which includes the trainee's overall performance score and conversation highlights. Additionally, trainees can visualize communication patterns through tools like a word cloud. These metrics and snippets offer significant opportunities for reflection, learning consolidation, and future improvement.

4) Rubric-Guided LLM-powered Empathy Evaluator: To support adaptive scenario progression, the **EMSIM** system continuously evaluates the trainee's verbal input for signs of empathic communication. This evaluation is not intended to diagnose or quantify empathy as a psychological trait; instead, it serves as a mechanism for guiding the virtual subject's behavior. This approach aligns with real-world crisis communication protocols, where perceived empathy can meaningfully shape the direction and intensity of an encounter.

Thus, we developed an empathy rubric that is composed of five categories. The scores from each category are summed to produce a final Empathy Score ranging from 0 to 1.0, where higher scores indicate a more empathetic input. The first four categories are derived from the Empathy Assessment Index (EAI), a validated tool for measuring empathy through five dimensions: affective response, self–other awareness, perspective taking, emotion regulation, and empathic attitudes [21].

Authorized licensed use limited to: University of Illinois. Downloaded on April 10,2026 at 21:08:15 UTC from IEEE Xplore. Restrictions apply.



Fig. 3. Virtual environment and character design in **EMSIM** (a) The virtual character is standing in front of a suburban house. (b) A top-down view of the simulation environment. (c) A police vehicle was parked in front of the house.

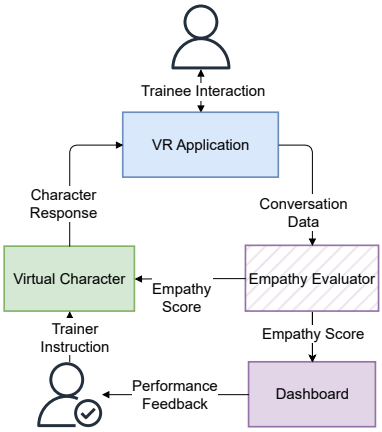


Fig. 4. System architecture of **EMSIM** .

Conversational Empathetic Response	
Level 0: Denial of patient's perspective	*"You're not actually thinking about doing this, right?"*
Level 1: Automatic recognition	*"That sounds tough."*
Level 2: Implicit recognition	*"I hear you. Let's focus on getting through this safely."*
Level 3: Acknowledgment without pursuit	*"That must be really painful..."*
Level 4: Acknowledgment with pursuit	*"Can you tell me more about what makes you feel this way?"*
Level 5: Confirmation	*"I'm here to listen..."*
Level 6: Shared feeling	*"You're not alone..."*
Emotional Reaction	
Level 0: No Emotional Reaction	
Level 1: Weak Emotional Reaction	*"Everything will be fine."*
Level 2: Strong Emotional Reaction	*"I feel really sad for you."*
Interpretation	
Level 0: No Interpretation	
Level 1: Weak Interpretation	*"I understand how you feel."*
Level 2: Strong Interpretation	*"This must be terrifying," or "I also have anxiety attacks..."*
Exploration	
Level 0: No Exploration	
Level 1: Weak Exploration	*"What happened?"*
Level 2: Strong Exploration	*"Are you feeling alone right now?"*
Cumulative Score	
Positive Increase: De-escalation	
Negative Decrease: Escalation	

Fig. 5. Rubric designed to evaluate empathetic communication in the **EMSIM** system. The rubric is structured around four key dimensions of empathetic input: Conversational Empathetic Response, Emotional Reaction, Interpretation, and Exploration.

To ensure empathy is fully covered, each rubric category is aligned with one or more EAI components. The fifth category (Cumulative Score) factors in the previous responses of the ongoing conversation. Although each category is designed to be mutually exclusive and captures a unique dimension of empathy in test-based conversations, they overlap in the conceptual components of empathy. The Conversational Empathetic Response Level category is adapted from the Empathetic Communication Coding System (ECCS) [22]. The other categories Emotional Reactions, Interpretations, and Explorations are based on a conceptual framework developed to analyze empathy in text-based interactions [23]. This framework is particularly well-suited to the **EMSIM** simulation environment, as it was specifically designed to assess empathy in non-face-to-face communication contexts.

The Cumulative Score category aims to factor in trends of empathetic or non-empathetic inputs from the trainee to escalate or de-escalate the situation. The Cumulative Score allows for progressive conversations and personalized feedback for the trainee. The final Empathy Score is calculated by summing the individual scores from each of the five rubric categories.

The weighting of the rubric categories can be adjusted based on the specific training scenario and objectives. For this study, we consulted with law enforcement trainers to determine appropriate weightings to align with the instructional goals of simulating a suicidal scenario. As a result, the five categories were weighted as follows: Conversational Empathetic Response (40%), Emotional Reaction (20%), Interpretation (20%), Exploration (10%), and Cumulative Score (10%).

B. System Implementation

1) Architecture: The **EMSIM** system employs a distributed architecture with several components, including: a VR Application where trainees interact in an immersive environment with the virtual character; an virtual character System that could be modified by the trainer to manage response generation, animation, and profile configuration; a web-based monitoring dashboard where trainer can review trainee's performance; and an Empathy Scoring System that evaluates trainee's performance. These components communicate through the directional data flows illustrated in Figure 4, collectively creating the training experience.

2) AI-Driven Virtual Character: The Virtual Character system, implemented using Convai, a character creation platform, provides real-time voice-to-voice interaction with digital hu-

Authorized licensed use limited to: University of Illinois. Downloaded on April 10,2026 at 21:08:15 UTC from IEEE Xplore. Restrictions apply.

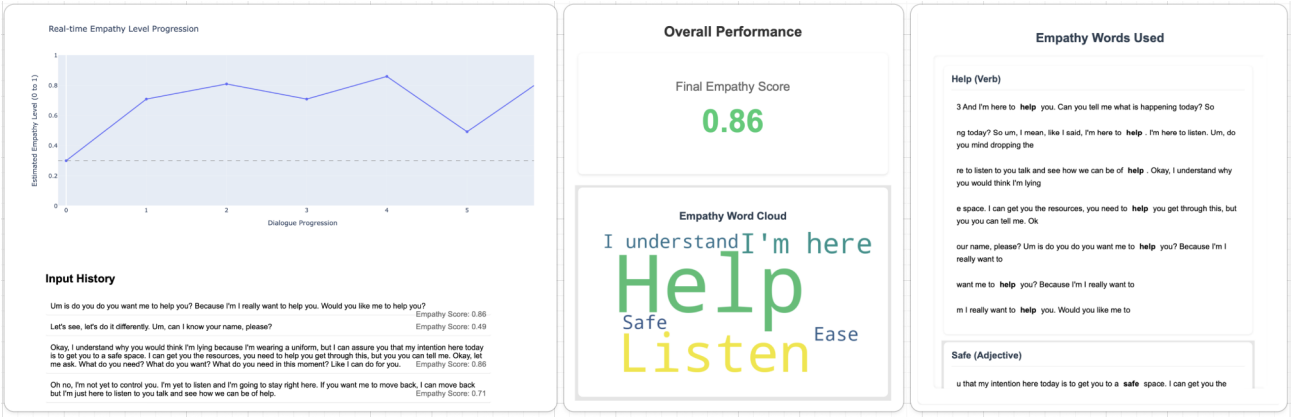


Fig. 6. Performance dashboard interface in **EMSIM** .

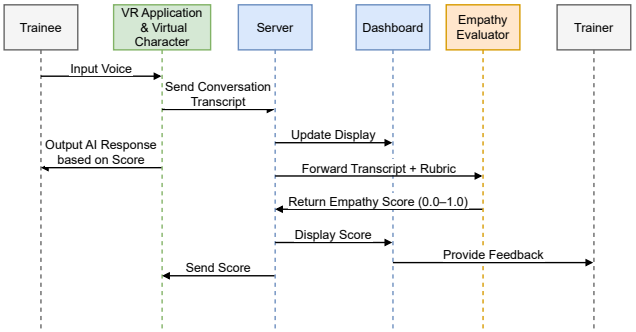


Fig. 7. System interaction sequence in **EMSIM** .

mans by chaining speech-to-text and text-to-speech together. Character profiles are managed through a configuration interface, offering control over characters' personality, background information, and response style. During the interaction, the system also generates real-time lip sync and body animation. Additionally, the system keeps a collection of pre-defined character states (e.g., escalating, de-escalating) that can be triggered dynamically by specific empathy scores, allowing branching behaviors in an otherwise open-ended interaction.

3) VR Application: The VR application delivers the core training experience by integrating several key technologies. At its foundation, the system uses Unreal Engine's digital human and environment tooling, which enables high-fidelity animation and visual effects for scenario endings. The application is designed to be compatible with the PC VR platform, allowing us to leverage powerful hardware such as NVIDIA RTX 4090 used in the study and stream content to headsets like Meta Quest 3 via QuestLink. Other than these visual components, we also developed a robust state manager that communicates with the dashboard, scoring, and character system to coordinate scenario progression. This provides fine-grained control over the simulation flow while maintaining the modularity necessary for extensibility and scenario customization.

4) Performance Dashboard: The Performance Dashboard (Figure 6), which facilitates both real-time monitoring and post-training review, is implemented as a web application using Flask and hosted on cloud infrastructure (AWS EC2). It communicates with other system components via RESTful APIs to exchange data. The dashboard's real-time capabilities are implemented to provide immediate insights during simulations. The backend accesses an LLM (GPT-4-turbo) via API calls to perform evaluations using the empathy rubric for the trainee input. The front-end employs standard web technologies, such as asynchronous JavaScript (AJAX) polling and libraries like Chart.js, to dynamically render the live transcript and visualize the evolving empathy score history in a line graph. The dashboard presents key performance indicators, including the final empathy score, conversation highlights, and an empathy word cloud. The word cloud visualization is generated using a Python library(wordcloud). All information is consolidated to display on the front-end technologies mentioned above.

5) System Interaction Sequence: While the trainee engages in conversations with the Virtual Character, the system follows a precise sequence to create a dynamic simulation (Figure 7). The interaction begins with collecting voice input from the VR headset microphone. The voice input is immediately transcribed into text, which is then routed to both the Character System and the server. The server updates the monitoring dashboard, allowing trainers to view the conversation in real time. Meanwhile, the transcript is bundled with an empathy rubric and sent to the Scoring System. This system evaluates the performance and returns a normalized empathy score between 0.0 and 1.0. The score serves a dual purpose: it appears on the dashboard, and it is returned to the VR application, which updates the character's internal state and prompts the Character System for a response (e.g., escalate if there is a low score). The Character System then generates an appropriate verbal response and animation, which are rendered through a digital human in the VR environment. This entire feedback loop finishes within seconds, maintaining a natural

Authorized licensed use limited to: University of Illinois. Downloaded on April 10,2026 at 21:08:15 UTC from IEEE Xplore. Restrictions apply.

conversation while delivering real-time feedback and adaptive behavior that grounds the training to learning objectives.

IV. User Study

A. User Evaluation

To evaluate the usability and perceived cognitive load of the **EmSim** system, a quantitative evaluation was conducted with 14 trainees at a police training institute, all of whom were in week 14 of their 16-week training program. Most participants reported prior experience with de-escalation training (80%) and empathy training (86.7%). In contrast, familiarity with XR or VR technologies was more limited, with 46.7% of participants indicating they had never used such devices.

Participants began by completing a consent form and a pre-test interview that captured demographic information as well as relevant training and technical experience related to the project. Each trainee then completed two rounds of the **EmSim** simulation, selecting their own strategies for interaction. All sessions were screen-recorded, including both the VR first-person view and the dashboard interface displaying dynamic escalation scores. Following the session, participants completed a post-test survey consisting of the System Usability Scale (SUS), the NASA Task Load Index (NASA-TLX), and open-ended feedback on the system's potential as an empathy training tool from the Learner's perspective.

B. Expert Evaluation

To assess the usefulness and effectiveness of **EmSim** as an de-escalation training tool, a qualitative study was conducted with five experts (E1-E5) from a law enforcement training institute. These participants have extensive experience in empathy training in de-escalation scenarios, and most of them reported limited or no familiarity with XR or VR technologies. Before engaging with the **EmSim** system, the experts completed a consent form and a pre-test questionnaire. Qualitative data were gathered through semi-structured interviews, which were audio-recorded and transcribed for analysis.

V. Findings

A. System Usability and Cognitive Load

The System Usability Scale (SUS) results indicated that participants perceived **EmSim** to be highly usable, with an average score of 81.61. This score corresponds to a grade of "A" on the usability scale, reflecting a top-tier user experience. These findings suggest that minimal effort was required to operate the system effectively, supporting its potential for broader implementation with limited need for additional user training or adaptation.

Results from the NASA Task Load Index (NASA-TLX) further indicated that participants experienced a low overall workload while interacting with the system. The average raw TLX score was 2.79, suggesting that the task was perceived as easy, low-stress, and minimally demanding. While a few participants reported slightly elevated scores (3.33, 3.50, and 3.83) in areas such as mental demand and effort, the overall trend remained consistent across categories such as frustration, temporal demand, and physical effort. These consistently low ratings highlight a smooth and manageable user experience. The SUS score indicates a favorable usability outcome, possibly due to the intuitive and familiar format of using virtual reality and first-person perspectives.

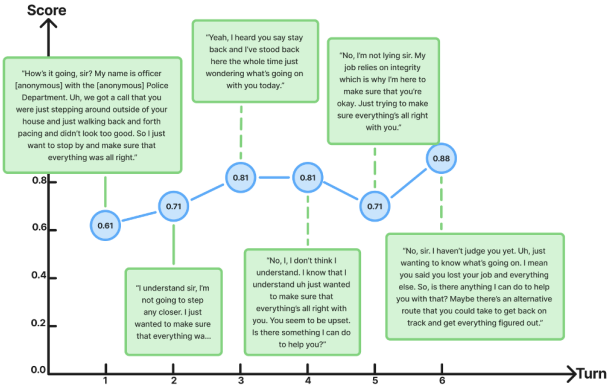


Fig. 8. User Log of Careful De-escalator (P2, Session 1)

B. Behavioral Strategies and Interaction Patterns

To investigate how participants engaged with the simulation and adapted their conversational approaches, we analyzed screen-recorded interaction trajectories from the dashboard system across two trials. Each participant completed the simulation twice, with the freedom to vary their responses. We extracted representative interaction sequences and visualized them as user journey maps, plotting turn-level scores alongside corresponding dialogue excerpts (Figure 8 and Figure 9). These visualizations reveal two contrasting behavioral strategies: careful de-escalation and exploratory volatility.

1) Careful De-escalatory Trajectories: In the first pattern, observed across multiple participants in their initial session, users adopted a measured and deliberate conversational tone. These interactions typically began with moderate empathy scores and showed a gradual upward trend over successive turns, indicating increasing alignment with the de-escalation goal. Participants used affirming language, offered respectful reassurances, and demonstrated reflective listening. Utterances such as "I just want to make sure everything's all right with you" or "Is there something I can do to help you?" contributed to sustained increases in the system's feedback score. This pattern suggests that participants were treating the simulation seriously, applying de-escalatory training principles, and adapting based on visible feedback.

2) Exploratory Volatility and Intentional Downward Trajectories: In contrast, several participants in later sessions demonstrated exploratory behaviors, pushing conversational boundaries either to probe the system's limits or experiment with less appropriate strategies. In these trajectories, empathy scores often dropped sharply after the initial turn and remained low for several interactions. For example, as shown in Figure 9, one participant opened with a respectful introduction but

Authorized licensed use limited to: University of Illinois. Downloaded on April 10,2026 at 21:08:15 UTC from IEEE Xplore. Restrictions apply.

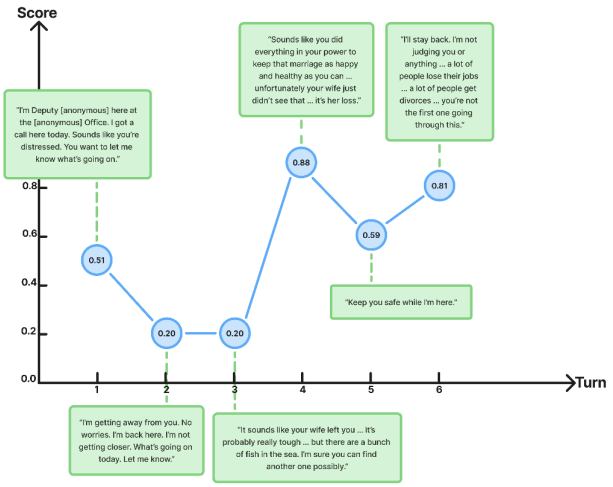


Fig. 9. User Log of Exploratory Pathway (P4, Session 2)

quickly shifted to dismissive or confrontational language, producing scores as low as 0.20. Some participants appeared to deliberately "fail" the simulation, asking antagonistic questions or issuing vague threats, resulting in a prolonged downward trend.

Notably, some participants switched strategies between trials, attempting escalation first and de-escalation second, or vice versa. These strategy shifts reflect the simulation's utility not just as an evaluative tool, but as an environment for safe experimentation. Participants could test conversational approaches and immediately observe consequences, whether beneficial or detrimental. This dynamic suggests that the real-time feedback helped reinforce an experiential understanding of high-stakes communication.

C. Reflections from Trainers

1) Emotionally Difficult Characters Could Enhance De-escalation Training: Trainers highlighted the value of exposing trainees to emotionally difficult virtual characters that reflect the complexity of real encounters in the fieldAs E1 described, "The standoffish person, the person that just has to be pissed off at the world... the one that just doesn't like the police," represents exactly the kind of persona that forces officers to navigate emotionally charged situations with composure. These scenarios require trainees to tailor their approach and practice de-escalation in a way that goes beyond scripted communication. To support this kind of adaptive communication, trainers often relied on the Verbal Judo framework, a widely used model for managing conflict through calm and persuasive dialogue. "Most of it's based off of the verbal judo template because it's multifaceted. You can use it in a lot of different ways," E4 explained. This flexibility made it especially well-suited for engaging with diverse character types in the simulation.

2) Lifelike Character Movement Deepens Immersion: Trainers noted that the virtual character's limited movement



Fig. 10. A law enforcement trainee using **EmSim** during an experiment session.

reduced the intensity and realism of the interaction. As E4 explained, "The character could have been like a little more animated... if they're agitated, especially they're going to move around more... even if it's just pacing or waving their arms." E5 shared a similar view, suggesting, "The character could probably be a little bit more life-like... maybe a little bit more animated rather than static." These reflections highlight the importance of incorporating emotional behaviors into the character's body language. Movements that signal distress or escalation (e.g., fidgeting, pacing, or a raised voice) are especially critical for reflecting the realities of field encounters. When characters convey emotional states through dynamic gestures, it becomes easier for trainees to recognize cues and respond with empathy. E3 added that embedding realistic behavioral prompts (e.g., subtle hand movements) could reinforce officer training: "Make the character put their hands in their pocket a lot... if the AI can... then they address... please take your hands out of the pockets... it just reminds them of what they need to do." These features help simulate high-pressure situations in which both verbal and nonverbal communication are essential.

3) Natural Dialogue and Conversational Authenticity Play a Critical Role: Dialogue plays a central role in shaping empathy-driven interactions, and trainers consistently emphasized its importance for effective de-escalation. How an officer initiates contact can significantly influence the interaction's trajectory; establishing rapport through clear, respectful communication was viewed as essential. As E2 explained: "A lot of it is the approach of the officer... hand me your license and immediately it sets up that brick wall... instead of... introducing yourself, telling them why you stopped them." This underscores how early tone-setting builds trust and reduces resistance. Furthermore, trainers highlighted that practicing these skills is particularly valuable given potential generational differences in communication habits. E2 noted that newer recruits may lack face-to-face experience, saying, "They spend all the time on the phone. Yeah. No face-to-face communication." Given the fundamental importance of nuanced verbal exchange in de-escalation, the ability of the AI system to facilitate natural, uninterrupted conversation is crucial for effective training.

Authorized licensed use limited to: University of Illinois. Downloaded on April 10,2026 at 21:08:15 UTC from IEEE Xplore. Restrictions apply.

4) Reflection Facilitates Coaching: The system's analytics features were seen as essential for making that learning process visible and actionable. Visual tools such as graphing and conversation summaries allowed trainers to isolate specific dialogue moments and provide targeted feedback. "The graphing, I think, is really important… being able to access what was said… to give them their point value… is a critical training tool," E1 noted. These features helped both trainees and instructors identify when emotional communication succeeded and where it fell short. As E3 reflected, "It gives you good insight on what to say, the empathetic things to say, the key things not to say." By linking performance data to communication quality, the system helps trainees better understand how their words and tone influence outcomes.

VI. Discussion

A. Contrasting Perspectives on Controllability in AI-Driven Training Systems

1) Delay and Turnaround Time: One unexpected finding in our evaluation was the perception of system delay between trainee input and AI response. Due to computational processing time and LLM response generation, the system incurs a brief delay, typically several seconds, between each conversational turn. Although this delay was not deliberately designed, expert participants expressed that it served a valuable pedagogical function. Specifically, they noted that the pause gave trainees an opportunity to reflect before responding, which encouraged more thoughtful and emotionally attuned communication. This aligns with prior work showing that increased wait time can enhance learning outcomes and improve response quality in instructional settings [24]. In simulation-based contexts, controlled pacing has also been noted to support learner focus and emotional engagement [25]. In real-world policing scenarios, such reflection time is rare, but for novice trainees, the additional pause may enhance empathy-building by encouraging cognitive regulation during emotionally charged dialogue. Although partially constrained by system performance, this delay could be formalized as an adjustable feature in future iterations, enabling instructors to calibrate response pacing according to trainees' skill levels.

2) AI-Generated Response Unpredictability: The virtual character's responses were generated in real-time by a large language model (LLM), based on system prompts and trainee input. While the conversational output is guided by scenario context and system constraints, the model ultimately determines the specific wording and tone of each response. This introduces a degree of unpredictability: responses sometimes included repetition, indirect answers, or emotionally ambiguous language. Interestingly, participants did not perceive these responses as flawed or distracting. Several noted that such imperfections mirrored real human behavior, particularly in emotionally intense situations, where individuals may repeat themselves, deflect, or struggle to articulate their thoughts. This observation supports broader HCI research suggesting that mild AI imperfections can enhance authenticity and trust in human-AI interaction, especially when the goal is to simulate real human variability [26], [27].

From a design perspective, while controllability is often prioritized in educational systems to ensure consistency and learning reliability [28], in this context, a controlled degree of unpredictability may actually enrich the learning experience, reflecting the nuanced, emotionally complex nature of real-world communication.

B. Expectation and System Improvements

While the current empathy evaluator focuses on textual and conversational content, future iterations of **EmSim** would benefit from integrating multimodal inputs, particularly vocal tone analysis and biometric indicators, to deliver a more nuanced and context-sensitive assessment of empathic behavior. Biometric data such as heart rate, blood pressure, and stress levels could help visualize the physiological impact of high-stress encounters. As E2 reflected, "Once that heart rate gets above 160 beats per minute, things start happening—auditory exclusion kicks in, fine motor skills disappear, their ability to communicate diminishes tremendously." Integrating these metrics could offer insight into when and how physiological stress impairs communication, enabling trainers to coach officers not just in what to say, but when to regulate themselves before speaking. Participants also pointed out limitations in scenario interactivity, particularly around access to standard officer tools.

C. Limitations and Future Research

While the initial evaluation of **EmSim** yielded promising results, several limitations constrain the scope of our findings and present opportunities for future research. The current version of **EmSim** was evaluated using a single scenario: a mental health crisis involving a suicidal individual. However, real-world policing demands adaptive empathy across a wide range of high-stress encounters. Future research could include a broader library of scenarios.

VII. Conclusion

We introduced **EmSim** , an AI-driven VR training system designed to support de-escalation communication through open-ended interaction with emotionally responsive virtual characters. By combining a rubric-guided evaluator and a real-time performance dashboard, the system enables trainees to explore conversational strategies while providing actionable feedback. Our exploratory study with law enforcement trainees and trainers suggests that **EmSim** is usable, engaging, and effective for practicing high-stakes communication. This work highlights the potential of large language models to support adaptive, feedback-rich training in immersive environments. Future efforts will refine the rubric and scale evaluation across broader contexts.

References

[1] C. Cooper, "Patrol police officer conflict resolution processes," *Journal of Criminal Justice*, vol. 25, no. 2, pp. 87–101, 1997.

Authorized licensed use limited to: University of Illinois. Downloaded on April 10,2026 at 21:08:15 UTC from IEEE Xplore. Restrictions apply.

[2] A. R. Grubb, "Effective police negotiation: Synthesising the strategies and techniques that promote success within hostage or crisis situations," in *Police Conflict Management, Volume I: Challenges and Opportunities in the 21st Century*. Springer, 2023, pp. 285–314.

[3] CBS News. We asked 155 police departments about their racial bias training. here's what they told us. CBS News.

[4] M. T. Compton, E. Jackson, E. Fu, H. F. Andrews, R. Bruno, E. Comartin, H. Galfalvy, D. Kamin, L. G. Pope, E. Vega, and A. C. Watson, "Crisis Intervention Team Mental Health Training for Law Enforcement Officers: Protocol for a Multi-Site, Randomized, Controlled Trial," vol. 7, no. 1, pp. 63–70.

[5] R. S. Engel, H. D. McManus, and T. D. Herold, "Does de-escalation training work?: A systematic review and call for evidence in police use-of-force reform," vol. 19, no. 3, pp. 721–759.

[6] J. A. A. Lavoie, N. Alvarez, and Y. Kandil, "Developing Community Co-designed Scenario-Based Training for Police Mental Health Crisis Response: A Relational Policing Approach to De-escalation," vol. 37, no. 3, pp. 587–601.

[7] O. Zechner, L. Kleygrewe, E. Jaspaert, H. Schrom-Feiertag, R. I. V. Hutter, and M. Tscheligi, "Enhancing Operational Police Training in High Stress Situations with Virtual Reality: Experiences, Tools and Guidelines," vol. 7, no. 2, p. 14.

[8] D. Scorgie, Z. Feng, D. Paes, F. Parisi, T. Yiu, and R. Lovreglio, "Virtual reality for safety training: A systematic literature review and meta-analysis," vol. 171, p. 106372.

[9] N. Moore, N. Ahmadpour, M. Brown, P. Poronnik, and J. Davids, "Designing Virtual Reality–Based Conversational Agents to Train Clinicians in Verbal De-escalation Skills: Exploratory Usability Study," vol. 10, no. 3, p. e38669.

[10] T. Bosse and S. Provoost, "Integrating Conversation Trees and Cognitive Models Within an ECA for Aggression De-escalation Training," in *PRIMA 2015: Principles and Practice of Multi-Agent Systems*, Q. Chen, P. Torroni, S. Villata, J. Hsu, and A. Omicini, Eds. Springer International Publishing, vol. 9387, pp. 650–659.

[11] R. S. Engel, N. Corsaro, G. T. Isaza, and H. D. McManus, "Assessing the impact of de-escalation training on police behavior: Reducing police use of force in the Louisville, KY Metro Police Department," vol. 21, no. 2, pp. 199–233.

[12] M. D. White, C. Orosco, and S. Watts, "Beyond force and injuries: Examining alternative (and important) outcomes for police De-escalation training," vol. 89, p. 102129.

[13] ——, "Can police de-escalation training reduce use of force and citizen injury without compromising officer safety?" *Journal of Experimental Criminology*, vol. 21, no. 1, pp. 45–70, 2025.

[14] D. Wang, X. Wang, Q. Zheng, B. Tao, and G. Zheng, "How interaction paradigms affect user experience and perceived interactivity in virtual reality environment," in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 223–234.

[15] J. E. Muñoz, J. A. Lavoie, and A. T. Pope, "Psychophysiological insights and user perspectives: Enhancing police de-escalation skills through full-body VR training," vol. 15, p. 1390677.

[16] Bosse Tibor, Gerritsen Charlotte, and De Man Jeroen, "An Intelligent System for Aggression De-Escalation Training," in *Frontiers in Artificial Intelligence and Applications*. IOS Press.

[17] C. Maathuis, K. Van Den Steen, and R. Van Dijke, "Virtual Reality Training Prototype for Police-Bystander Conflict," in *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, pp. 313–315.

[18] O. Shaikh, V. E. Chai, M. Gelfand, D. Yang, and M. S. Bernstein, "Rehearsal: Simulating Conflict to Teach Conflict Resolution," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, pp. 1–20.

[19] C. E. Hughes and K. M. Ingraham, "De-escalation training in an augmented virtuality space," in *2016 IEEE Virtual Reality (VR)*. IEEE, pp. 181–182.

[20] J. Holmes, L. Sokoloff, N. McNaughton, S. Gardner, L. Truong, K. Nemethy, and K. Joseph, "VR-based simulation training for de-escalation of responsive behaviours in persons with dementia: Efficacy and feasibility," p. bfqc7623.

[21] C. A. Lietz, K. E. Gerdes, F. Sun, J. M. Geiger, M. A. Wagaman, and E. A. Segal, "The empathy assessment index (eai): A confirmatory factor analysis of a multidimensional model of empathy," *Journal of the Society for Social Work and Research*, vol. 2, no. 2, pp. 104–124, 2011. [Online]. Available: https://doi.org/10.5243/jsswr.2011.6

[22] A. Gilbert, A. Miles, S. Carnell, and B. Lok, "Using virtual patients to support empathy training in health care education: An exploratory study." *Simulation in Healthcare*, vol. 19, no. 3, pp. 151–157 – 157, 2024.

[23] A. Sharma, T. Althoff, A. Miner, and D. Atkins, "A computational approach to understanding empathy expressed in text-based mental health support warning: This paper contains content related to suicide and self-harm." in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, no. EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, (1)Paul G. Allen School of Computer Science and Engineering, University of Washington, 2020, pp. 5263–5276 – 5276.

[24] M. B. Rowe, "Wait time: Slowing down may be a way of speeding up!" *Journal of teacher education*, vol. 37, no. 1, pp. 43–50, 1986.

[25] P. Dieckmann, D. Gaba, and M. Rall, "Deepening the theoretical foundations of patient simulation as social practice," *Simulation in Healthcare*, vol. 2, no. 3, pp. 183–193, 2007.

[26] E. Luger and A. Sellen, "" like having a really bad pa" the gulf between user expectation and experience of conversational agents," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 5286–5297.

[27] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 12, no. 2, pp. 293–327, 2005.

[28] K. VanLehn, "The behavior of tutoring systems," *International journal of artificial intelligence in education*, vol. 16, no. 3, pp. 227–265, 2006.

Authorized licensed use limited to: University of Illinois. Downloaded on April 10,2026 at 21:08:15 UTC from IEEE Xplore. Restrictions apply.