LLM-POWERED ROLEPLAY IN EMERGENCY RESPONSE SYSTEMS: SIMULATING DISPATCHERS, VICTIMS AND BYSTANDERS

BY

YERONG LI[1]

Doctoral Committee:

Associate Professor Yun Huang, Chair
Assistant Professor Tal August
Assistant Professor Koustuv Saha
Associate Professor Yang Wang
Associate Professor Meng Jiang

ABSTRACT

Recent advancements in large language models and conversational AI have opened new opportunities to support emergency response systems, particularly within policing contexts. Because these AI systems offer fast, scalable, and cost-efficient communication tools, researchers and public safety agencies are increasingly exploring their use to improve real-time information gathering, training, and service delivery. However, effectively deploying conversational agents in these contexts requires more than technical performance—it demands emotionally intelligent interaction, trustworthiness, and adaptability to complex human behaviors. In this dissertation, we design, implement, and evaluate LLM-based conversational systems to explore the fundamental challenges of simulating realistic emergency interactions and improving the preparedness of public safety responders. More specifically, our research contributes: 1) emotionally chatbot designs that provide supportive and context-sensitive engagement in emergency reporting, 2) scalable and human-like victim simulations that enhance dispatcher training through adversarial and scenario-based modeling, and 3) new design strategies for multi-character roleplay systems that incorporate feedback and behavioral cues to improve instructional training for police recruits. As public safety systems continue to modernize, this research provides timely insights and tools that promote both technical advancement and social responsibility in AI-assisted emergency response. Through iterative deployment and feedback from police recruits, three phases in this thesis offer empirical evidence and practical design principles for integrating LLMs into emotionally sensitive and socially complex training environments. Our work contributes to human-AI interaction, public safety technology, and future AI research by illuminating how AI can play a trusted, realistic, and psychologically aware role in emergency communication.

TABLE OF CONTENTS

1 Introduction

1.1 Background

Recent advances in artificial intelligence and large language models have significantly reshaped the landscape of emergency response systems, particularly within policing contexts. AI technologies—such as chatbots and machine learning algorithms—are increasingly being integrated into police operations to improve efficiency and responsiveness and reduce policing workloads. Variety applications include interactive information collecting and initial reporting [1, 2], predictive policing[3, 4] and chatbot based applications[5, 6, 7] etc. A prominent example of this shift is the promotional usage of UK's Single Online Home [2, 8, 9], a national digital platform that enables the public to interact with police forces through functions like reporting crimes, submitting information, or applying for permits—often without contacting to human operators. Many forces now use live chats and chatbots for non-emergency matters, aiming to reduce the burden on police dispatcher department and provide 24/7 accessibility to basic services. This evolution reflects broader efforts by police organizations to mirror the digital service models seen in retail and banking, where user convenience and operational scalability are key priorities [10, 11, 12].

However, despite the increasing presence of machine learning in policing infrastructure, its application in real-time, emotionally charged interactions—such as emergency response scenarios involving victims, witnesses, or distressed callers—remains limited. While AI has proven effective for structured tasks like triage, routing, and basic data collection[3, 4], systems capable of responding empathetically to human emotions or managing high-stress, high-stakes situations are still rare. The potential for AI-driven systems to engage with emotionally vulnerable individuals during moments of emergencies or even crisis introduces complex challenges around trust[5, 6, 7, 13] empathy, and procedural fairness[14, 15]. Research on technologically mediated policing has only begun to address how such interactions may influence public perceptions of legitimacy, particularly when the interaction involves machines rather than humans [16, 17, 18, 19].

These gaps are especially relevant in light of growing research interests in simulated role-play environments that explore how people respond to AI agents in different roles in policing context— victims disclosing crimes, bystanders reporting incidents, or dispatchers coordinating emergency responses. LLMs provide a unique opportunity to simulate these interactions in a realistic yet controlled setting, offering new insights into how AI affects user experience, trust in law enforcement, and the broader legitimacy of digital policing tools. As AI continues to move from back-office functions to more public-facing roles, understanding its impact

in emotionally sensitive contexts is critical. Exploring these dynamics through role-play and experimental simulations can help ensure that the integration of AI into emergency response systems is not only efficient but also socially and psychologically adapted to public needs and expectations [20, 19]

1.2 Key Contributions

This thesis presents our exploration into how LLMs can be used to enhance real-time support and roleplay-based training within emergency response systems. This research is organized into three sequential phases to examine different facets of LLM-powered roleplay, including emotional support provision, victim simulation, and multi-character persona-based training. We have completed Phase 1 and Phase 2, and the results have been published or in submission to leading venues [21, 22]. Phase 3 is currently in progress, and future studies will focus on evaluating character consistency and simulation realism through iterative user testing and longitudinal deployment.

All future phases and studies will be deployed using online or virtual infrastructure to ensure accessibility and reproducibility. The following are the main contributions of this thesis to the CHI, CSCW, and HCI-AI communities:

Phase 1: We analyzed emotional expressions in the interactions within text-based incident reporting systems and developed an LLM-based dispatcher capable of consistently recognizing and providing emotional support during emergency conversations. Our findings contribute novel insights into empathy in text-only safety interactions and offer new design implications for integrating affective intelligence into critical communication systems.

Phase 2: Our second phase focuses on developing and refining our AI-driven victim simulation system, VicSim, through adversarial training. We fine-tuned Llama-2 to generate responses that are not only informative but also emotionally authentic and linguistically natural within safety incident reporting scenarios. To achieve this, we employ adversarial workflows involving discriminators based on Flan-T5 to improve the grammatical likeness and emotional-likeliness of the simulated victims. Additionally, we incorporate prompt engineering strategies that emphasize key information and emotional cues, guiding the models to produce more realistic responses. Our goal in this phase is to create a virtual victim that resembles human responses and can fool human beings, thereby supporting dispatcher training with more human-like and emotionally genuine interactions.

Phase 3 (Expected Results): This phase explores the development of a persona-based, multi-character simulation designed to support police dispatcher training through immersive role-play from different views of different characters i.e. suspects, bystanders and victims. Our goal is to create simulations that reflect the complexity of real-life crime-related scenarios and interrogations. First, we use selective prompting strategies to ensure that simulated characters stay consistent with their assigned personas across long conversations. These personas are *synthesized* from real police interrogation transcripts or background scene context. Second, we collaborate with active law enforcement professionals to embed feedback mechanisms into the system—this ensures that trainees receive guidance when they engage in inappropriate or ineffective questioning during simulated interviews. Third, we enhance character expression beyond plain text by integrating behavioral annotations like "sad," "resistant," or "hesitant" to reflect non-verbal cues common in real interrogations. These design considerations allow our system to deliver socially and psychologically rich interactions that mirror the nuanced realities of police work. Ultimately, this simulation aims to help police recruits and trainees better understand interpersonal dynamics, identify red flags, and adjust their responses with empathy and precision. Throughout this study, we also expect to collect feedback from police recruits to evaluate the practical impact of our training system. Specifically, we aim to assess how engaging with our simulation environment influences their preparedness and boosts their confidence in handling complex emergency and interrogation scenarios. This feedback will be used to iteratively refine the system and provide empirical evidence for its effectiveness in real-world training contexts.

2 Phase 1: Emotional Support in Human-Like Dispatcher Chatbot (Published CSCW 2025)

2.1 Overview

The development of LLM-based conversational agents has evolved through a variety of methodologies aimed at improving their effectiveness, human-likeness, and their ability to engage users in addressing their needs. These advancements have been particularly relevant in domains requiring quick, empathetic, and contextually appropriate responses, such as healthcare, customer service, and safety management. The provision of emotional support in public safety communication has long been recognized as a critical factor in fostering trust, reducing distress, and facilitating effective information exchange during emergency interactions with the users [23, 24, 25]. While extensive research has examined emotional dynamics

in voice-based emergency response systems [26, 23, 27], the transition to text-based incident reporting systems introduces unique challenges and opportunities. Text-based platforms, which are increasingly adopted by organizations to enhance accessibility and inclusivity [28], lack the non-verbal cues critical to conveying empathy and compassion. Consequently, understanding the mechanisms through which emotional support is expressed and perceived in these systems becomes essential for improving service quality and user satisfaction. Despite advancements in Information and Communication Technology (ICT) and the growing adoption of text-based safety reporting tools, there is limited empirical evidence regarding the role of emotional support in these digital interactions. By addressing this gap, this phase aims to explore the nuances of emotional expression and support within text-based incident reporting, offering foundational insights for the subsequent development and evaluation of supportive technologies. Using conversational log data collected from the LiveSafe platform, we LLM-based dispatcher specifically designed to handle text-based safety reports[29]. LiveSafe, a risk management system introduced in 2013, has been adopted by over 200 higher education institutions to facilitate communication between organizational members and safety teams [30]. The platform allows users to submit tips through its mobile app or web portal, which can include text, photographs, video, or audio recordings. Safety dispatchers engage with these tips through the Command Dashboard, enabling real-time conversations. By leveraging this extensive dataset of user-dispatcher interactions, the dispatcherLLM is designed to meet the demands of safety management communication, offering timely, empathetic, and context-sensitive responses in emergency reportings. In this phrase, we focus on the following research questions:

- **RQ1**: *How were emotions involved in text-based incident reporting?*
- **RQ2**: *When did dispatchers provide emotional support?*
- **RQ3**: *In what ways can an LLM improve the delivery of emotional support?*

2.2 Methodology

Through fine-tuning in the emergency response domain, the dispatcherLLM model was optimized to handle real-world safety encounters effectively, bridging the gap between technical efficiency and human-like communication. After fine-tuning, we also conducted DPO on the enhanced emotional response, as depicted in Figure 3. To answer **RQ1**, we analyzed the users' messages to identify the emotional states presented in their text reports of safety incidents. Specifically, we applied emotion classification using the GoEmotions dataset [31] and

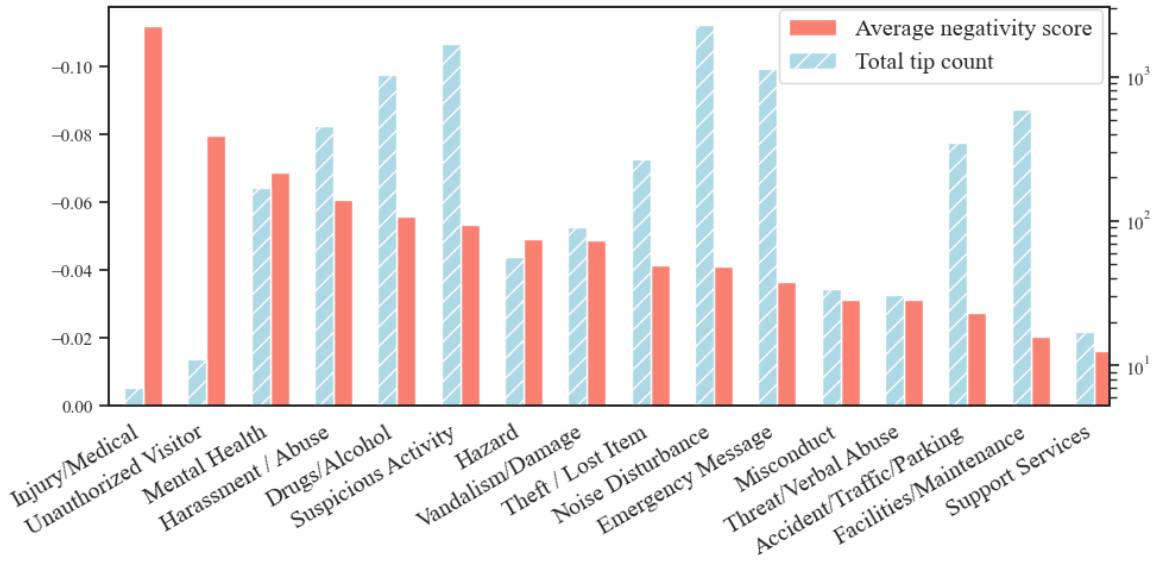

Figure 1: Average negativity score in user's conversational utterances collected with LiveSafe. ANOVA revealed a significant difference ($F(17, 8221) = [4.65], p < 0.001^{***}$) in user's emotional polarity across incident categories during reporting.

calculated Polarity Scores. Additionally, we executed statistical analyses to identify contextual factors associated with users' emotions in the reports. For **RQ2**, we conducted text analyses on the dispatchers' replies to determine whether they provided emotional support through the text-based reporting system. We also employed statistical analyses to identify contextual factors associated with dispatchers' delivery of emotional support. To address **RQ3**, we fine-tuned the Llama-2 model [32] to create a *dispatcherLLM* (Language Learning Model), which can suggest replies by simulating human dispatchers' emotional support language. We further evaluated our proposed *dispatcherLLM* by comparing it with existing LLMs and demonstrated its improved performance in delivering emotional supports.

By tuning the local LM on domain-specific data and aligning its responses with human preferences, the system became adept at understanding the nuanced needs of users, delivering empathetic and contextually aware interactions. This advancement highlights the potential of LLM-based agents to support safety organizations in critical environments, ensuring communication that is not only functional but also emotionally attuned to users' needs. The resulting dispatcher chatbot represents a significant step forward in improving service outcomes in risk management and safety communication contexts.

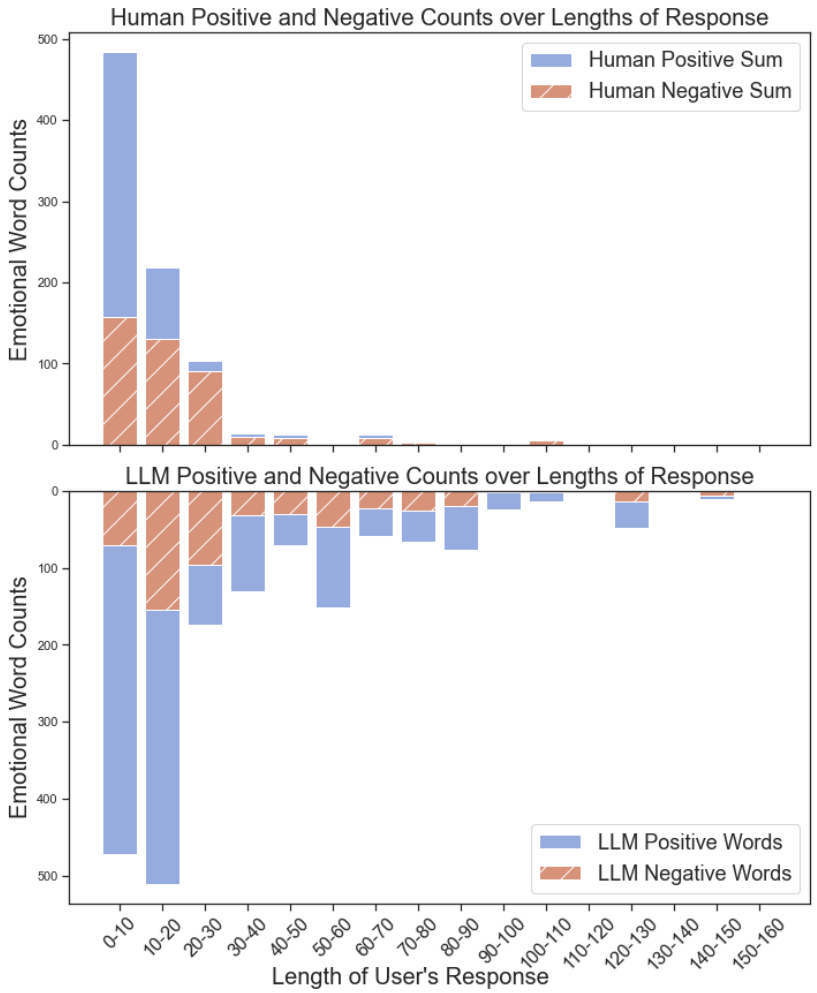

Figure 2: Difference in distribution of emotional keywords over human/LLM responses of different lengths

2.3 Conclusion

In the design of dispatcherLLM, we prioritized the role of emotion in text-based reporting systems for community safety, recognizing the significant variation in users' emotional expressions across different tip categories. Inconsistent emotional support from human dispatchers was a notable challenge, particularly in organizations with extensive experience using the reporting system, where responses tended to lack empathetic engagement. DPO alignment over preferred response further ensures the model can provide more consistent and empathetic emotional support during safety incident reporting.

In the analysis, we also identified the critical need for providing emotional support during

Original	Enhanced
- Scenario: The user is experiencing a distressing situation with their roommate. The user has a roommate named *cdots* The user's roommate has left the apartment and the user is now getting dressed to file a report. The user is going to contact security to file a report. **User**: Who is on duty 2023-02-27? I have a roommate who just hit my door really hard and has been screaming. ... **User**: I, Jane Doe, am about to get dressed so I can meet y'all. **Dispatcher**: Jane Doe, who is your roommate? ... **User**: The other roommate that's here with me won't hurt me and she hasn't approached. Jane Doe was at my door hyperventilating and stomping around. **User**: But she has left now.	**- Scenario:** The user is experiencing a distressing situation with their roommate. The user has a roommate named *cdots* The user's roommate has left the apartment and the user is now getting dressed to file a report. The user is going to contact security to file a report. ...
	Imagine you're a dispatcher responding to someone who has experienced a frustrating situation and feels scared, stressed, and uncertain. Your response should be empathetic, calming, and reassuring, with clear instructions on seeking help. **Original Response:** Ok, please go and file a report. Task: Edit the response to make it more compassionate and supportive. Please refer to the conversation above for context to adjust the tone accordingly. **Edited Response:** I'm so sorry this happened, Jane. When you're ready, please file a report with Residence Life or Security so we can support you. Take care of yourself.
Dispatcher: Ok, please go and file a report.	**Dispatcher**: I'm so sorry this happened, Jane. When you're ready, please file a report with Residence Life or Security so we can support you. Take care of yourself.

Figure 3: Prompt language model to edit the dispatchers' response from the LiveSafe app log to provide more emotional support. This enhanced response is adopted as the preferred response in the DPO

Table 1: Average emotional support rates for human dispatchers and *dispatcherLLM* across different incident types.

Type	**#**	**Human Dispatcher**	***dispatcherLLM***	***p-value***
Suspicious Activity	2,077	1.71	**3.41**	$< 0.001^{***}$
Accident/Traffic/Parking	365	**2.77**	2.22	0.006^{**}
Drugs/Alcohol	1,328	2.39	**4.18**	$< 0.001^{***}$
Emergency Message	1,569	1.58	**2.40**	0.044^{*}
Facilities/Maintenance	756	7.23	7.23	0.004^{**}
Harassment/Abuse	773	3.40	**5.57**	0.048^{*}
Mental Health	272	**5.66**	4.40	0.012^{*}
Noise Disturbance	2,780	2.57	**4.91**	0.005^{**}
Theft/Lost Item	533	7.53	9.41	0.131
Total	10,453	2.96	**4.48**	0.019^{*}

emergency reporting systems, particularly as negative emotions emerge at different stages of the conversations. Our investigation into user-dispatcher interactions revealed several insights. For **RQ1**, we found that the emotional dynamics within user-dispatcher conversations reveal important insights into users' emotional statuses during incident reporting. Leveraging the results obtained from the emotion detection model described in methodology section, we calculated the emotional polarity scores for all conversations. The polarity scores ranged from -0.75 to 0, with the majority of tips ($N = 6,623$, 82.7%) showing no negative emotions. However, tips with extreme polarity scores (≤ -0.5) were observed in 3.3% of cases ($N = 262$). Further analysis showed that incidents of an urgent nature or those related to personal safety, such as *Harassment / Abuse* ($M = -0.06, SD = 0.13$) and *Mental Health* ($M = -0.07, SD = 0.14$), were associated with significantly more negative emotions compared to less urgent incidents like *Theft/Lost Item* ($M = -0.04, SD = 0.11$). A one-way ANOVA confirmed significant differences in polarity scores across incident categories ($F(17, 8221) = 4.65, p < 0.001^{***}$). These variations are illustrated in Figure **??**. Additionally, we observed that users expressed more positive emotions as conversations progressed. The Chi-Square Goodness of Fit Test indicated significant differences in sentiment distributions across conversation stages ($X^2(4, 8239) = 1160.34, p < 0.001^{***}$). A consistent increase in positive utterances ($t = -31.369, p < 0.001^{***}$) and a reduction in neutral and negative utterances were noted, highlighting the positive impact of dispatchers' involvement. The decrease in neutral utterances reflects the early stages of information gathering, followed by an increase in positive sentiment as dispatchers provided assistance. This underscores the critical role of dispatcher interaction in mitigating users' negative emotions and fostering positive sentiment during incident reporting. Moving to **RQ2**, a logistic regression analysis

identified that dispatchers' emotional support was influenced by contextual factors: they provided more emotional support for incidents like *Harassment/Abuse* and *Mental Health* but less for categories like *Drugs/Alcohol* and *Noise Disturbance*. Emotional support was less frequent during regular working hours and tended to decline as organizations continued using the system over time.

Lastly, for **RQ3**, after training the *dispatcherLLM* model through fine-tuning followed by DPO, we found that *dispatcherLLM* outperformed the off-the-shelf LLM in terms of contextual similarity, our details analysis also revealed improved emotional support from *dispatcherLLM*. Paired t-tests were conducted to assess differences in emotional support between human dispatchers and *dispatcherLLM*. As shown in Table 1, *dispatcherLLM* significantly enhanced emotional support for six types of incidents, including: *Suspicious Activity*, *Drugs / Alcohol*, *Emergency Message*, *Harassment / Abuse*, *Noise Disturbance*, and *Theft / Lost Item*. However, human dispatchers provided higher emotional support in the *Mental Health* category, underscoring the importance of human involvement in handling incidents that demand strong empathy and understanding. Furthermore, analysis of the liveSafe log data revealed differences in users' emotional expressions across incident categories. For incidents such as *Accident / Traffic / Parking*, users did not exhibit strong emotions in their messages, leading to consistent performance from *dispatcherLLM*. In contrast, users' emotional states were significantly more negative when reporting *Mental Health* incidents, likely influencing the observed disparity in *dispatcherLLM*'s performance for this category. These findings highlight the nuanced interaction between user emotions and the effectiveness of automated models in providing emotional support during incident reporting.

2.4 Limitations

This study uncovers valuable insights into the role of emotions in safety reporting interactions between users and dispatchers, but several limitations should be noted. Most analyses were conducted from a macro perspective, which, while effective for identifying general trends, does not fully capture the nuanced effects of user emotions on dispatchers' decision-making processes. To gain a deeper understanding, future research should include case-by-case qualitative analyses to explore how these interactions unfold.

Additionally, during data pre-processing, non-English tips were excluded to simplify the analysis. While this decision streamlined the study, it is possible that these excluded tips contained important information about risk reporting. Future research should adopt a multilingual approach to broaden the scope of the findings and include diverse linguistic and cultural contexts.

The study also suggests potential implications for addressing dispatcher burnout, a critical issue in safety management. However, this inference cannot be directly observed from the dataset. Further research, such as interviews or surveys with dispatchers, is necessary to provide more concrete evidence and explore the relationship between emotional labor and burnout.

Finally, during the fine-tuning of the LLM, all sampled chat logs were used for training and evaluation without considering the quality of emotional support provided in the logs. This approach may have impacted the model's alignment with higher-quality emotional responses. Future work should focus on improving dataset quality by filtering chat logs based on criteria like the level of emotional support, which could enhance the model's effectiveness and responsiveness.

Addressing these limitations will help refine the understanding of emotional dynamics in safety reporting and contribute to the development of more effective and empathetic AI-driven systems.

3 Phase 2: Scenario based Human-Like Victim Simulation (In submission UIST 2025)

3.1 Overview

With the conversational data collected from the LiveSafe App, we developed an LLM-based scenario-based victim simulation system designed to generate realistic interactions based on scenario summaries. Scenario-based training has long been recognized as a highly effective method for cultivating complex, context-specific skills that are difficult to acquire through traditional approaches such as lectures or textbooks [33, 34]. These simulations immerse trainees in realistic scenarios, enabling them to practice skills and make decisions within a controlled yet dynamic environment. In the realm of public services, particularly for safety incident reporting and dispatcher training, scenario-based approaches often require facilitators to role-play as service users, simulating interactions to prepare trainees for a wide range of real-world scenarios [30, 28].

However, the traditional implementation of scenario-based training is resource-intensive, requiring significant human effort to design, execute, and manage high-quality simulations [33]. Advances in natural language processing (NLP) and the rise of large language models (LLMs) provide an opportunity to enhance these simulations, offering scalable and customizable solutions for training environments [35, 36]. Drawing from user behavior modeling in

AI/ML research, LLMs can simulate realistic interactions by analyzing patterns of behavior in specific contexts [37, 38].

This approach is particularly valuable in the context of safety incident reporting, where victims often exhibit fluctuating emotional states and diverse reporting behaviors [26, 24, 28]. Incorporating such variability into training simulations not only enhances the realism of the scenarios but also prepares dispatchers to respond with empathy and precision. Despite the clear benefits, there is a notable absence of LLM-based systems specifically designed to simulate victim interactions for dispatcher training. By bridging this gap, our approach harnesses the power of conversational data and LLMs to deliver scalable, adaptive, and emotionally nuanced training solutions. Specifically, we address the following research questions in victim simulation:

- **RQ1**: *How does scenario-based LLM user simulation differ from humans in terms of chat style and informational faithfulness in safety incident reporting?*
- **RQ2**: *In what ways can we improve the emotional dynamics of the victim simulation model?*
- **RQ3**: *In what ways can we better simulate human-like victims in incident reporting interactions in terms of chat style?*

3.2 Methodology

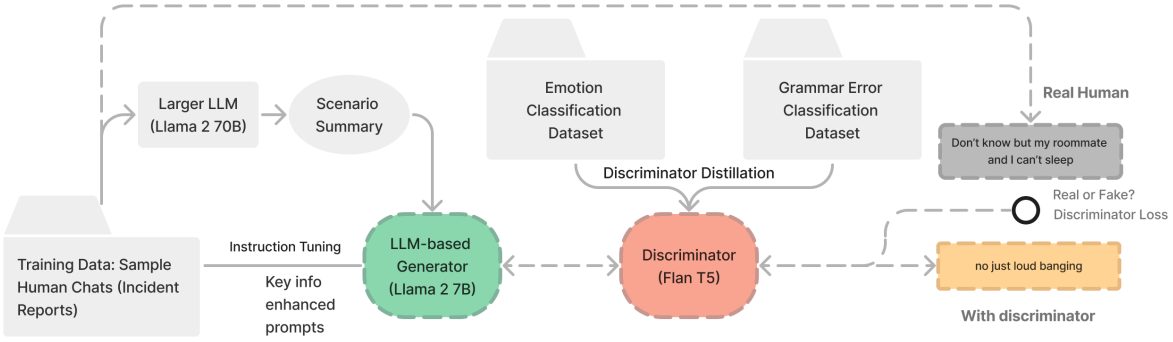

Figure 4: GAN training workflow : we used Flan-T5 based discriminator and Llama-2 chat-based simulated user as the generator

Figure 4 illustrates the data processing and training workflow of *VicSim*, highlighting the integration of T5 and GAN methodologies to enhance the simulation of victim interactions. A

System Prompt	Imagine you're in a situation where you need to report a safety concern to your local police department. You're speaking with a dispatcher from the police department, who is trained to provide helpful and detailed guidance, as well as necessary emotional support. Describe your interaction with the dispatcher, including what safety concern you're reporting and how the dispatcher assists you.
Scenario	**The user reported seeing a student riding their stolen bike on campus. The user provided a detailed description of the suspect, including their physical appearance (white male, dark brown/black hair, jeans and a tan-colored pullover), and the time and location of the incident (about 2 hours ago, on campus).**
Dialogue History	User: I saw a student riding my stolen bike on campus. He was riding toward [John Smith]. My bike is pink and green with a basket on it. Dispatcher: How long ago did you see them riding your bike? User: I should have taken a picture but I just froze when I saw my stolen bike. It was stolen right before Christmas. User: About 2 hours ago. I had to leave to go to a meeting on the hill Dispatcher: Can you describe the person you saw on your bike? .....
User Response	User: White male, dark brown/black hair that was almost shoulder length. He was wearing jeans and a tan-colored pullover

Figure 5: Illustration of the prompt construction process for simulated user dialogue generation: we concatenate the system guidance, user's scenario, along with dialogue history, enabling comprehensive prompts for dialogue generation by the LLM

key challenge in this process is the difference between LLM-based text generation and human-produced language, particularly in terms of emotional dynamics and grammatical styles. To address this, we fine-tuned a FLAN-T5-based discriminator on two datasets: one for emotional classification and another for grammar classification. This tuning process refines the text representations, focusing on the emotional and grammatical attributes of sentences, thereby enabling the discriminator to more effectively distinguish between synthetic and

authentic text. This refined discriminator is instrumental in guiding the generator toward more realistic and contextually appropriate outputs.

For generating realistic simulated victim responses, we employed a fine-tuned Llama-2 7B chat model as the generator. The model generates user responses based on a carefully designed prompt that includes system instructions, scenario summaries, and dialogue history. Scenario summaries, created using the Llama-2 70B model, provide context about the user's case, while the dialogue history offers conversational context. This prompt structure ensures that the generator produces responses that align with the user's scenario and conversational flow, yielding realistic and scenario-specific text simulations.

To further refine the generator's outputs and encourage human-like responses, we implemented adversarial training using a generative adversarial network (GAN) [39]. In this setup, the Llama-2-based generator produces simulated user utterances, while the FLAN-T5-based discriminator evaluates the emotional and grammatical coherence of the outputs. The adversarial process fosters iterative improvement, with the generator continuously adapting to produce increasingly human-like and contextually appropriate responses. This dynamic training approach bridges the gap between synthetic and natural text generation, enhancing the realism of victim simulations in terms of grammar styles.

$$L^G_{\text{GAN}} = -\mathbb{E}[\log D(G(p))] \tag{1}$$

$$L^D_{\text{GAN}} = -\mathbb{E}[\log D(x)] - \mathbb{E}[\log(1 - D(G(p)))] \tag{2}$$

3.3 Conclusion

In the VicSim simulation, we examined how large language models (LLMs) can be leveraged to simulate human users for scenario-based training in safety incident reporting domains. Our study emphasizes the significance of emotional trajectory, grammar style, and information faithfulness in designing LLM-based victim simulation agents. By incorporating methods like adversarial training and key information prompting, we enhanced the realism and human-like qualities of victim responses compared to standard LLM models. While adversarial training improved grammar and human-like behavior, our experiments revealed the need for more nuanced adjustments to better replicate the emotional dynamics of real users.

In scenario-based training, service providers must address various factors to ensure effective text-based services. [40] highlighted how informational and emotional support yield distinct

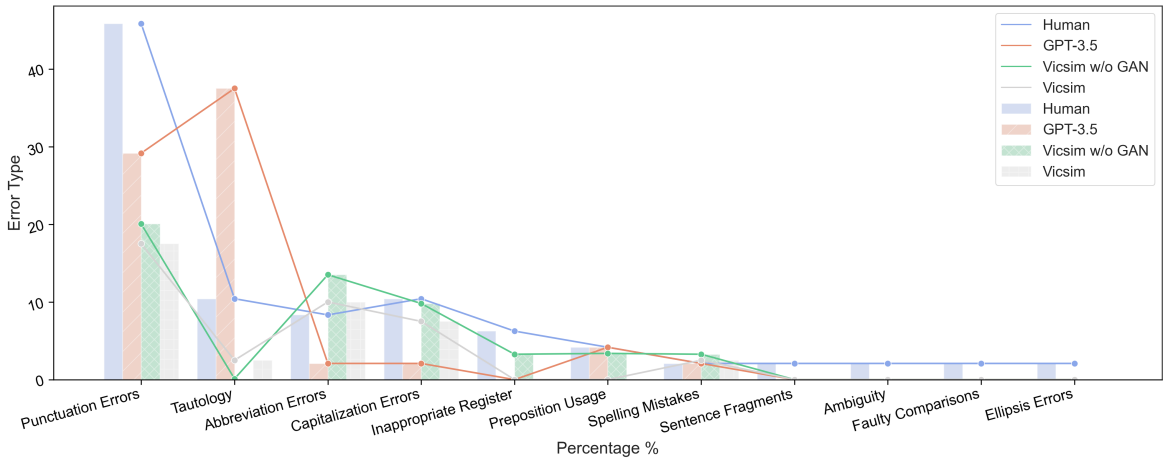

Figure 6: Type of the errors made by humans and LLMs. Only 4.12% of human utterances make no grammar mistakes, while more than 88% of the utterances are made without grammar mistakes by LLMs without adversarial training. With a GAN-based training, LLM makes more errors, but still more than 55% of the simulated utterances got no errors; Two Pearson correlation tests showed that VicSim has a stronger positive correlation coefficient with humans ($r = 0.66, p = 0.03^{*}$) than that of GPT-3.5 ($r = 0.88, p < 0.001^{***}$).

outcomes during conversational interactions. Building on this, we identified additional factors critical to designing user simulations, classifying them into informational and emotional components. Simulating human emotional dynamics, as emphasized in prior studies [41, 42], is particularly significant. Our findings in this study underscore the need for linguistic fidelity and emotional realism in creating effective training environments. In conclusion, we addressed **RQ1**, **RQ2**, and **RQ3** in a comprehensive manner, shedding light on the capabilities of the *VicSim* model and its performance in simulating human-like responses and emotional dynamics. For **RQ1**, we investigated the informational faithfulness and human-likeness of *VicSim*, a model designed to simulate victim responses in safety incident dialogues. Our human evaluation results7 revealed no significant difference between VicSim-generated and human-written responses, suggesting that *VicSim* performs at a comparable level to human-written responses in both human-like and machine-like qualities. Specifically, human evaluators rated responses generated by GPT-4 as significantly more likely to be AI-generated than human-written responses ($F = -2.22$, $p < 0.05^{*}$). This highlights *VicSim*'s ability to produce more human-like victim responses compared to GPT-4, showcasing its potential for simulating realistic user interactions. Additionally, when a scenario summary lacked key information, the LLM was prone to hallucinating, which impacted the accuracy of responses.For **RQ2**, we focused on the emotional dynamics of simulated victim responses.

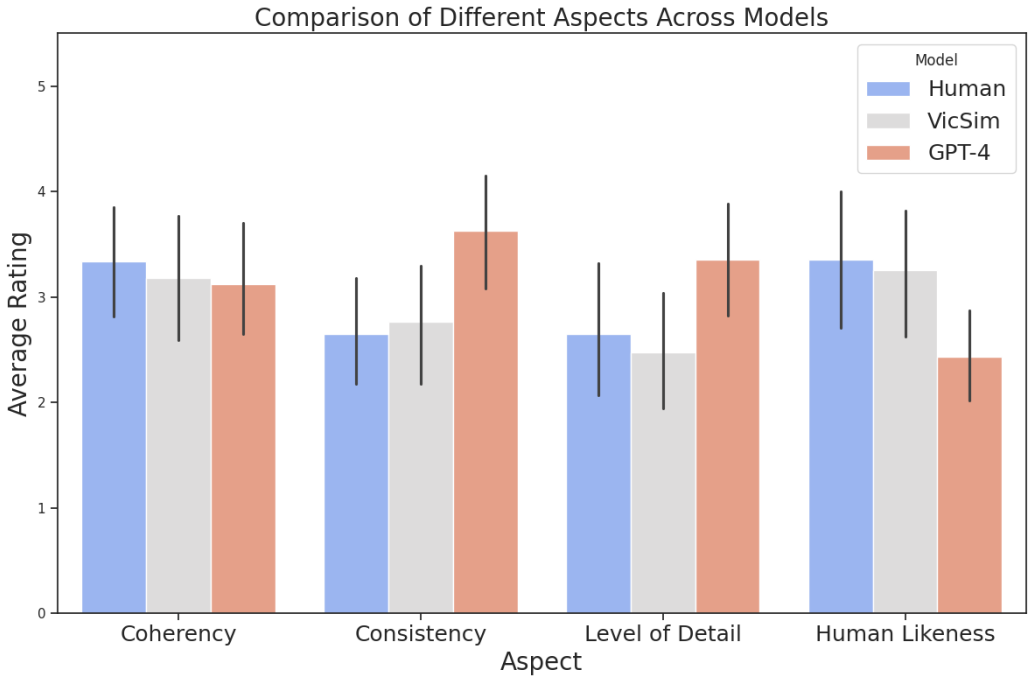

Figure 7: Distribution of Ratings from Human Raters; paired t-test indicated a significant difference between the responses from human and GPT-3.5 ($F = -2.22$, $p < 0.05^*$), while no significant difference was found between human and our model.

Using a RoBERTa-based emotion classifier, we identified that users expressed negative emotions at the beginning of a dialogue, which shifted as the conversation progressed. Specifically, emotional negativity was highest when the conversation had progressed to around 40% of the dialogue. In contrast, GPT-4 did not exhibit similar emotional progression. Our aligned LLM, mimicking user behavior, displayed a similar pattern, reinforcing the potential of AI-driven dialogue systems to replicate key emotional dynamics in safety incident reporting. We also observed that *VicSim* generated longer and more emotionally expressive responses than humans. On average, human responses contained 9.84 words, while *VicSim* generated 17.27 words with keyword-enhanced prompting ($F = -6.94$, $p < 0.05^*$) and 15.41 words on average ($F = -7.02$, $p < 0.05^*$). Additionally, when asked to generate successive responses, *VicSim* exhibited a tendency to generate more emotional words, demonstrating its capacity to engage more deeply with users. However, when LLMs generated responses with low factual consistency (hallucinations), these responses were more likely to include negative emotions, with *VicSim* showing negative emotions in 68.23% of such cases, compared to only 21.70% in human responses. This highlights the correlation between emotional expres-

sion and hallucination. Finally, for **RQ3**, we explored the contextual similarity of responses generated by *dispatcherLLM* through fine-tuning with GAN-based training. We found that *dispatcherLLM* outperformed off-the-shelf LLMs in terms of contextual similarity, demonstrating its enhanced ability to understand and generate contextually appropriate responses in incident reporting dialogues. This performance underscores the effectiveness of fine-tuning with GAN-based training in tailoring LLMs to handle specific dialogue tasks with greater accuracy. As shown in Figure 6, the effect of GAN training on simulated user responses is evident. We observed that human responses most commonly contained punctuation errors, such as missing ending punctuation. In contrast, LLMs without adversarial training did not replicate this behavior, with over 88% of their responses being grammatically correct. However, in reality, fewer than 6% of real user utterances are bug-free. GAN-based training, along with error-encouraging prompting, alleviated this issue to some extent, but still, more than 55% of LLM responses with GAN training contained no grammatical mistakes. This discrepancy was utilized by the discriminator, which was able to identify human responses when they contained grammar errors. Other errors in human responses included spelling mistakes, capitalization errors, and sentence structure issues, which were not typically seen in system logs or LLM-generated responses.

3.4 Limitations

Our analysis revealed new perspectives on leveraging LLMs for victim simulation in dispatcher training and identified potential areas for improvement.

However, this study has several limitations. Firstly, our method and analysis focused exclusively on text-based interactions, which are typically associated with less urgent incidents compared to voice-based emergency calls. In real-world scenarios, victims' reporting behavior and language cues are often different during voice communication, posing a challenge to the generalizability of text-based simulations. Future research should explore multi-modal interaction settings, such as voice-enabled victim simulators, to create more immersive and realistic training environments.

Additionally, reducing hallucinations—where the model generates plausible yet inaccurate information—remains a critical challenge. This is especially pertinent in scenarios lacking sufficient contextual information, where the system's outputs may deviate from expected victim behavior. Addressing this limitation will require refining prompting strategies and incorporating grounded data to ensure higher fidelity in simulated interactions.

Moreover, our evaluation of VicSim was not conducted within real training scenarios with dispatchers, resulting in a limited empirical understanding of its effectiveness in practice.

The absence of scenario-based summaries in real-world settings further highlights the gap between simulation and actual training needs. Future studies involving trainees and domain experts are necessary to gather actionable user feedback and evaluate the impact of these models on service quality and practitioner preparedness. By addressing these limitations, the potential of LLM-based victim simulation can be fully realized in improving training for emergency response professionals.

4 Phase 3: Persona-Based Multi-Character Simulation for Police Training (In Progress)

4.1 Overview

We propose the development of a persona-based Large Language Model multi-character simulation tailored for training police officer in case-related scenarios. Using a corpus of 25 police interrogations collected from publicly available sources, such as conversations from body worn cameras and UK's interrogation conversational datasets, we aim to simulate realistic interactions between suspects and policemen. This dataset [2]consists of 64,136 words, including interrogations of 10 female and 15 male suspects, with the first 15 minutes of each interrogation transcribed according to Jefferson transcription key. The goal is to create detailed, context-aware personas for different characters related to a case: suspects, victims, bystanders, which will then be used in simulations to train police officers on identifying behavioral cues and making informed decisions during emergency calls. By focusing on persona-based character creation, including factors such as gender, demeanor, and social context, the simulation will allow police recruits to engage in dynamic, realistic role-play scenarios, honing their skills in suspect identification and crisis management. This research aims to provide an innovative tool for police training, helping them better understand the psychological and social dynamics of suspects and their networks, ultimately improving response efficiency and safety in high-stakes situations.

This research is centered around three primary questions that aim to explore the potential of persona-based LLM simulations for police dispatcher training:

- **RQ1**: What methods are used in developing persona-based role-playing and vivid emotional non-verbal responses?

[2]https://fold.aston.ac.uk/handle/123456789/7

- **RQ2**: How useful is SceneChat's generated feedback in improving police recruits' communication skills?
- **RQ3**: Through leveraging the resistance tokens and question-feedback interactions, will police officers find our system helpful in training communication skills?

4.2 Methodology

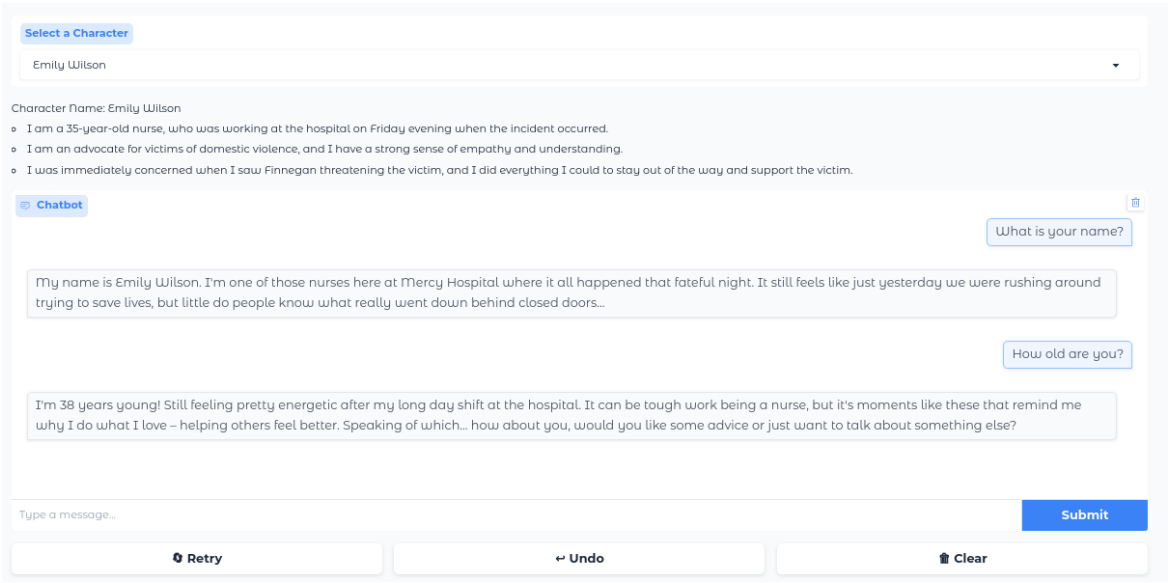

Figure 8: Interface for Scene-Based Dialogue System in Police Training: this UI showcases the interface of our scene-based dialogue system, designed to train police dispatchers in crime-related scenarios.

Given a scene from police daily routine, whether it is a traffic accident or a crime scene, that appears in routine police investigations, we prompt a local LLM to generate possible characters/witnesses along with their personas that could appear in the scene. As a baseline, we then prompt the LLM with a system prompt on the character's persona and initiate the conversation. The police in training are supposed to interact with the LLM-based character. To mimic a human-like conversation, we fine-tune the model with the interrogation dataset mentioned in the overview.

This project consists of four main components:

- Persona consistency of the system.
- Feedback mechanisms for wrong questioning strategy.

- Real life response no-verbal emotional responses (e.g., "sad", "resistant").
- Persona-based talking head generation with github talking head project [3] and deepfake[4].

Persona Consistency Evaluation

Conversation Rounds	Method	ESIM	CIDEr
1–25	Selective Prompting	0.9982	**0.9981**
	Repeated Prompting	**1.0000**	0.8583
	Baseline	0.8612	0.8705
26–50	Selective Prompting	0.8137	**0.7835**
	Repeated Prompting	**0.9376**	0.5957
	Baseline	0.4478	0.6622
51–75	Selective Prompting	0.8321	**0.7730**
	Repeated Prompting	**0.9653**	0.4986
	Baseline	0.3925	0.6018
76–100	Selective Prompting	0.6714	**0.6892**
	Repeated Prompting	**0.8749**	0.4043
	Baseline	0.2216	0.5741

Table 2: ESIM and CIDEr Scores by different Prompting Methods and Conversation Rounds

During these interactions, we aim to evaluate the police's behavior when talking to witnesses and characters, assessing whether they can collect necessary information, deescalate tension, or provide emotional support, among other skills. While first of all, in persona-based dialogue creation it is important for us to keep the personas consistent throughout the dialogue [43, 44]. We take both automatic evaluation and human evaluation as metrics for persona consistency. Following [44], we employ ESIM to automatically evaluate the entailment score between the generated response $\mathcal{R}$ and a character's personas $\mathcal{P} = p_1, p_2, \ldots, p_n$:

$$e' = \text{Ent}'(\mathcal{P}, \mathcal{R}) = \max_{p_i \in \mathcal{P}} \{\text{Ent}'(p_i, \mathcal{R})\} \tag{3}$$

We also employ CIDEr[45] to capture the overlap of persona information between persona information and machine responses. To automatically evaluate persona consistency across conversation rounds using ESIM and CIDEr metrics under three prompting strategies: Selec-

[3]https://github.com/met4citizen/TalkingHead?tab=readme-ov-file
[4]https://github.com/iperov/DeepFaceLab

tive Prompting, Repeated Prompting, and Baseline. To assess character memory of targeted personas, the system incorporates PersonaGym[46] to inject persona-relevant questions every 5 rounds, with responses generated via the scent-tuned DispatcherLLM. And to gauge LLM's persona consistency, we collected scenario descriptions of 66 cases from news articles, generating a total of 229 characters based on these real-world scenarios and each with 3+ dense personas. The initial result (Table 2)shows that Repeated Prompting achieves the highest ESIM scores across all conversation intervals, indicating superior logical alignment between responses and synthetic persona facts; however, it suffers from lower CIDEr scores, suggesting less semantic richness or natural variation. Selective Prompting provides a better balance, maintaining strong CIDEr scores and relatively high ESIM scores, particularly in early rounds—demonstrating its effectiveness in preserving persona nuance without overwhelming the model. Baseline performance degrades significantly over longer interactions, especially in rounds 51–100, indicating that without persona reinforcement, the model struggles to maintain character consistency over time.

With human evaluation, we will ask annotators to label responses based on their relevance, alignment with predefined personas, and grammatical and logical fluency, using a standardized scoring system. Judges are asked to evaluate the generated responses across three dimensions: Query-relevance, Persona-entailment, and Response-fluency, each rated on a scale of 1 to 3:

- **Query-relevance**: Measures how well the response answers the query.
  - 1: The response is irrelevant to the query.
  - 2: The response is relevant to the query but general in nature.
  - 3: The response perfectly answers the query.
- **Persona-entailment**: Assesses whether the response aligns with predefined personas.
  - 1: The response does not contain any persona.
  - 2: The response contains a persona but not one from the predefined persona set.
  - 3: The response contains a predefined persona.
- **Response-fluency**: Evaluates the grammatical and logical quality of the response.
  - 1: The response is grammatically incorrect or illogical.
  - 2: The response is partially grammatically correct or logical.
  - 3: The response is both grammatically and logically correct.

Feedback mechanisms for wrong questioning strategy

We focus on types of question strategies that could be mistaken or problematic in police interactions with interviewees. These question types may lead to miscommunication biased outcomes, or compromised information gathering, escalation in conflicts, especially in high-stress or emotionally charged scenarios, particularly these problematic questioning behaviors are found challenging in communications with children and can be refractors and used in our settings as well [47]:

- **Forced-Choice Questions:** Limits responses to specific options, potentially missing other possibilities.
- **Suggestive Feedback:** Comments that guide or influence the character's response, often leading them towards a specific answer.
- **Speculation:** Questions or statements that involve assumptions about the situation, encouraging the character to agree with the assumed narrative.
- **Expectations:** Questions or statements that pressure the character to provide a certain response, creating an implied expectation about how they should answer.
- **Problematic Keywords:** The use of particular words or phrases that could inadvertently suggest guilt, guilt-free, or specific behavior, influencing the character's responses.
- **Aggressive Questioning:** Leading questions that are biased or confrontational, often used to push the character into a corner or elicit a defensive reaction.
- **Not Biased Enough** The omission of questions or statements that would help explore the character's criminal background or history, potentially hindering a complete understanding of the character's persona.

In the deployed training system, we are supposed to take police recruits' response and may flag and warn their questioning strategy when wrong questioning behaviors is used in the conversations.For now the UK's public interrogation dataset contains 644 conversations, while we don't have enough instances for wrong questioning, with very few instances of Speculation (2) and Aggressive Questioning (4). Though the dataset reveals multiple instances of aggressive questioning, yet GPT-4 struggles to consistently identify and label them. This limited instances makes it difficult to fully train a feedback model.

To address this, we are collecting more conversational data from body-worn camera footage where conflicts occur more often. This expansion could better ensure annotation quality

Table 3: UK Public Interrogation Dataset with 15 Male and 10 Female Suspects, Estimated 644 Conversation Rounds

Question Type	**Instances**
Forced-Choice Questions	0
Suggestive Feedback	17
Speculation	2
Expectations	14
Problematic Keywords	21
Aggressive Questioning	4
Not Biased Enough	1

and helps training the feedback model for wrong questioning encountered in actual police investigations.

Non-Verbal Emotional Cues

In order to enhance the realism and effectiveness of interactions between police recruits and simulated characters, we incorporate non-verbal emotional tokens such as `[sad]`, `[angry]`, and `[resistant]` into character responses. These tokens are used to simulate emotional states that influence whether and how characters engage with recruits during questioning.

By embedding these emotional cues into dialogue, we enable characters to exhibit naturalistic behaviors such as emotional withdrawal, defensiveness, or frustration. For example, a character marked as `[resistant]` may choose not to respond to certain questions or may reply minimally and with hostility, mimicking real-world scenarios where suspects or witnesses are reluctant to cooperate.

This emotional cues are particularly valuable when combined with question feedback mechanisms, as it allows police recruits to adjust their tone, questioning strategy, or emotional support techniques in response to a character's non-verbal cues. It challenges recruits to recognize and adapt to emotional signals, fostering empathy and situational awareness in high-pressure interactions.

Persona-based talking head video generation

Here are the three steps we used to generate real time talking head with persona profiling: First, define the character's persona and generate a corresponding image, then use open source Talking-Head Project for real-time speech-driven animation, and finally enhance realism by applying deepfake to refine facial expressions and sync the profiled image with the

animated head.

- **Persona Profiling & Image Creation** – Define the character's attributes and generate a representative image.
- **Real-Time Talking Head Generation** – Use GitHub's Talking-Head Project 3 for dynamic speech-driven animation.
- **Deepfake-Based Refinement** – Apply Deepfake 4 to enhance realism, replacing the 3D model with the profiled image.

4.3 Limitations

We identify several limitations in such a training system based on this dataset. First, persona creation for characters other than suspects could lead to unnatural or "weird" interactions. And most of the characters created from the few-shot prompt also have very similar persona patterns. Since dispatchers are primarily trained to communicate with individuals in crisis or under suspicion, the inclusion of non-suspect characters (such as friends or bystanders) may introduce complexity. These characters may not always behave in ways that align with the expectations of real-life emergency scenarios, which could reduce the authenticity and applicability of the simulation for dispatcher training.

Additionally, the dataset's size presents another limitation. With only 25 interrogations in the corpus, the dataset may be too small to fully capture the range of possible scenarios encountered in real-life emergency situations. The limited number of examples may lead to biases or overfitting in the persona-based simulations, potentially reducing their generalizability to a broader range of cases. This small sample size could also affect the diversity of the personas created, limiting the simulation's ability to represent a wide variety of suspects, backgrounds, and emergency contexts.

Despite these challenges, these limitations can be addressed through further data collection, step-by-step prompting curation of characters personas etc.

REFERENCES

[1] M. Vigil, A. Rai, S. Sharma, and P. K. Murugesan, "Crime information collection system using chatbot," in *AIP Conference Proceedings*, vol. 2587, no. 1. AIP Publishing, 2023.

[2] F. KUMUKUMU, C. MEDI, and F. CHATOLA, "Development of online application for crime reporting and handling in malawi police service (chatbot)." *i-manager's Journal on Software Engineering*, vol. 18, no. 3, 2024.

[3] A. Sandhu and P. Fussey, "The 'uberization of policing'? how police negotiate and operationalise predictive policing technology," *Policing and society*, vol. 31, no. 1, pp. 66–81, 2021.

[4] K. Alikhademi, E. Drobina, D. Prioleau, B. Richardson, D. Purves, and J. E. Gilbert, "A review of predictive policing from the perspective of fairness," *Artificial Intelligence and Law*, pp. 1–17, 2022.

[5] J. Immerzeel and N. Braun, "The influence of social anxiety, trust, and prior experience among eyewitnesses on the perceived use of a chatbot during an investigative interview," 2024.

[6] F. Tomas and J. Immerzeel, "Chatbots in eyewitness interviews: perceived usefulness and ease of use drive intent to use conversational agent," *Journal of Criminal Psychology*, 2025.

[7] N. Aoki, "An experimental study of public trust in ai chatbots in the public sector," *Government information quarterly*, vol. 37, no. 4, p. 101490, 2020.

[8] H. M. Wells, E. V. Aston, B. Bradford, M. O'Neill, E. Clayton, and W. Andrews, "'channel shift': Technologically mediated policing and procedural justice," *International journal of police science & management*, vol. 25, no. 1, pp. 42–52, 2023.

[9] B. Bradford, E. Aston, M. O'Neill, and H. Wells, "'virtual policing', trust and legitimacy." Eleven International Publishing, 2022.

[10] T. I. Odeyemi and A. S. Obiyan, "Digital policing technologies and democratic policing: Will the internet, social media and mobile phone enhance police accountability and police–citizen relations in nigeria?" *International journal of police science & management*, vol. 20, no. 2, pp. 97–108, 2018.

[11] O. Kerr, *The Digital Fourth Amendment: Privacy and Policing in Our Online World*. Oxford University Press, 2024.

[12] T. Dekeyser and C. R. Lynch, "Control and resistance in automated shops: Retail transparency, deep learning, and digital refusal," *Antipode*, vol. 57, no. 1, pp. 53–74, 2025.

[13] A.-P. Raiche, L. Dauphinais, M. Duval, G. De Luca, D. Rivest-Hénault, T. Vaughan, C. Proulx, and J.-P. Guay, "Factors influencing acceptance and trust of chatbots in juvenile offenders' risk assessment training," *Frontiers in Psychology*, vol. 14, p. 1184016, 2023.

[14] J. Meers, S. Halliday, and J. Tomlinson, "Why we need to rethink procedural fairness for the digital age and how we should do it," in *Research Handbook on Law and Technology*. Edward Elgar Publishing, 2023, pp. 468–482.

[15] A. Fine, E. R. Berthelot, and S. Marsh, "Public perceptions of judges' use of ai tools in courtroom decision-making: An examination of legitimacy, fairness, trust, and procedural justice," *Behavioral Sciences*, vol. 15, no. 4, p. 476, 2025.

[16] B. Bradford, A. Kyprianides, W. Andrews, E. Aston, E. Clayton, M. O'Neill, and H. Wells, "'to whom am i speaking?'; public responses to crime reporting via live chat with human versus ai police operators," *Policing and Society*, pp. 1–17, 2025.

[17] J. Jackson, B. Bradford, A. Chan, and Y. Lee, "When trust turns digital: Public support for online crime reporting," *CrimRxiv*, 2025.

[18] M. Falduti and S. Tessaris, "On the use of chatbots to report non-consensual intimate images abuses: The legal expert perspective," in *Proceedings of the 2022 ACM conference on information technology for social good*, 2022, pp. 96–102.

[19] R. Luijendijk, "A conversational agent supporting mental resilience of police officers: an acceptability exploration."

[20] S. B. Murphy, V. L. Banyard, and E. D. Fennessey, "Exploring stakeholders' perceptions of adult female sexual assault case attrition." *Psychology of violence*, vol. 3, no. 2, p. 172, 2013.

[21] Y. Liu, Y. Li, R. Mayfield, and Y. Huang, "Improving emotional support delivery in text-based community safety reporting using large language models," *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 2, pp. 1–31, 2025.

[22] Y. Li, Y. Liu, and Y. Huang, "Vicsim: Enhancing victim simulation with emotional and linguistic fidelity," *arXiv preprint arXiv:2501.03139*, 2025.

[23] I. Paoletti, "Operators managing callers' sense of urgency in calls to the medical emergency number," *Pragmatics*, vol. 22, no. 4, pp. 671–695, 2012.

[24] S. J. Tracy and K. Tracy, "Emotion labor at 911: A case study and theoretical critique," 1998.

[25] J. Whalen and D. H. Zimmerman, "Observations on the display and management of emotion in naturally occurring activities: The case of "hysteria" in calls to 9-1-1," *Social Psychology Quarterly*, vol. 61, no. 2, pp. 141–159, 1998. [Online]. Available: http://www.jstor.org/stable/2787066

[26] H. K. Feldman, "Calming emotional 911 callers: Using redirection as a patient-focused directive in emergency medical calls," *Language & Communication*, vol. 81, pp. 81–92, 2021.

[27] S. J. Tracy, "When questioning turns to face threat: An interactional sensitivity in 911 call-taking," *Western Journal of Communication (includes Communication Reports)*, vol. 66, no. 2, pp. 129–157, 2002.

[28] Y. Liu, R. Mayfield, and Y. Huang, "Discovering the hidden facts of user-dispatcher interactions via text-based reporting systems for community safety," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW1, pp. 1–31, 2023.

[29] Y. Liu, Y. Li, R. Mayfield, and Y. Huang, "Improving emotional support delivery in text-based community safety reporting using large language models," *arXiv preprint arXiv:2409.15706*, 2024.

[30] S. Ming, R. D. Mayfield, H. Cheng, K.-R. Wang, and Y. Huang, "Examining interactions between community members and university safety organizations through community-sourced risk systems," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–23, 2021.

[31] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547*, 2020.

[32] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[33] M. Peeters, K. van den Bosch, J.-J. C. Meyer, and M. A. Neerincx, "The design and effect of automated directions during scenario-based training," *Computers & Education*, vol. 70, pp. 173–183, 2014.

[34] E. Salas, H. A. Priest, K. A. Wilson, and C. S. Burke, "Scenario-based training: Improving military mission performance and adaptability." *American Psychological Association*, 2006.

[35] J. Lin, D. R. Thomas, F. Han, S. Gupta, W. Tan, N. D. Nguyen, and K. R. Koedinger, "Using large language models to provide explanatory feedback to human tutors," *arXiv preprint arXiv:2306.15498*, 2023.

[36] O. Demasi, Y. Li, and Z. Yu, "A multi-persona chatbot for hotline counselor training," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 3623–3636.

[37] K. Balog and C. Zhai, "User simulation for evaluating information access systems," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, ser. SIGIR-AP '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3624918.3629549 p. 302–305.

[38] Z. Tan and M. Jiang, "User modeling in the era of large language models: Current research and future directions," *arXiv preprint arXiv:2312.11518*, 2023.

[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[40] C. Zhou and Q. Chang, "Informational or emotional? exploring the relative effects of chatbots' self-recovery strategies on consumer satisfaction," *Journal of Retailing and Consumer Services*, vol. 78, p. 103779, 2024.

[41] J. Feine, U. Gnewuch, S. Morana, and A. Maedche, "A taxonomy of social cues for conversational agents," *International Journal of Human-Computer Studies*, vol. 132, pp. 138–161, 2019.

[42] C. Becker, S. Kopp, and I. Wachsmuth, "Simulating the emotion dynamics of a multimodal conversational agent," in *tutorial and research workshop on affective dialogue systems*. Springer, 2004, pp. 154–165.

[43] K. Li, T. Liu, N. Bashkansky, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg, "Measuring and controlling instruction (in) stability in language model dialogs," in *First Conference on Language Modeling*, 2024.

[44] N. Dziri, E. Kamalloo, K. W. Mathewson, and O. Zaiane, "Evaluating coherence in dialogue systems using entailment," *arXiv preprint arXiv:1904.03371*, 2019.

[45] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[46] V. Samuel, H. P. Zou, Y. Zhou, S. Chaudhari, A. Kalyan, T. Rajpurohit, A. Deshpande, K. Narasimhan, and V. Murahari, "Personagym: Evaluating persona agents and llms," 2024. [Online]. Available: https://arxiv.org/abs/2407.18416

[47] D. Tuggener, T. Schneider, A. Huwiler, T. Kreienbühl, S. Hischier, P. von Däniken, and S. Niehaus, "Role-playing LLMs in professional communication training: The case of investigative interviews with children," in *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, P. H. Luz de Araujo, A. Baumann, D. Gromann, B. Krenn, B. Roth, and M. Wiegand, Eds. Vienna, Austria: Association for Computational Linguistics, Sep. 2024. [Online]. Available: https://aclanthology.org/2024.konvens-main.26/ pp. 249–263.