Improving Emotional Support Delivery in Text-Based Community Safety Reporting Using Large Language Models

YIREN LIU, Informatics, University of Illinois Urbana-Champaign, USA
YERONG LI, Department of Computer Science, University of Illinois Urbana-Champaign, USA
RYAN MAYFIELD, LiveSafe Inc., USA
YUN HUANG, School of Information Sciences, University of Illinois Urbana-Champaign, USA

Emotional support is a crucial aspect of communication between community members and police dispatchers during incident reporting. However, there is a lack of understanding about how emotional support is delivered through text-based systems, especially in various non-emergency contexts. In this study, we analyzed two years of chat logs comprising 57,114 messages across 8,239 incidents from 130 higher education institutions. Our empirical findings revealed significant variations in emotional support provided by dispatchers, influenced by the type of incident, service time, and a noticeable decline in support over time across multiple organizations. To improve the consistency and quality of emotional support, we developed and implemented a fine-tuned Large Language Model (LLM), named *dispatcherLLM*, designed to suggest replies through simulating human dispatchers' languages with appropriate emotional support. We evaluated *dispatcherLLM* by comparing its generated responses to those of human dispatchers and other off-the-shelf models using real chat messages. Additionally, we conducted a human evaluation to assess the perceived effectiveness of the support provided by *dispatcherLLM*. This study not only contributes new empirical understandings of emotional support in text-based dispatch systems but also demonstrates the significant potential of generative AI in improving service delivery.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; **Empirical studies in HCI**; • **Computing methodologies → Natural language generation**.

Additional Key Words and Phrases: text-based reporting system, live chat, safety reporting, emotion classification, event argument extraction, large language models

ACM Reference Format:
Yiren Liu, Yerong Li, Ryan Mayfield, and Yun Huang. 2025. Improving Emotional Support Delivery in Text-Based Community Safety Reporting Using Large Language Models. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW114 (April 2025), 31 pages. https://doi.org/10.1145/3711012

1 Introduction

The importance of emotional support in public safety services is well-documented, particularly in voice-based emergency response systems [19, 55, 66, 76]. Studies show that delivering emotional support effectively through communication is crucial for rapidly establishing rapport and trust, which facilitates quicker and more accurate information exchange [55]. Such an empathetic approach is adopted in various emergency response practices to build trust [76] and reduce the

Authors' Contact Information: Yiren Liu, yirenl2@illinois.edu, Informatics, University of Illinois Urbana-Champaign, USA; Yerong Li, yerong2@illinois.edu, Department of Computer Science, University of Illinois Urbana-Champaign, USA; Ryan Mayfield, ryan@livesafemobile.com, LiveSafe Inc., USA; Yun Huang, yunhuang@illinois.edu, School of Information Sciences, University of Illinois Urbana-Champaign, USA.

ACM 2573-0142/2025/4-ARTCSCW114
https://doi.org/10.1145/3711012

initial distress of individuals affected by traumatic experiences [62]. These insights are aimed at enhancing service quality and improving the provision of emotional support within the safety reporting domain [67].

As community Information and Communication Technology (ICTs) have evolved, text-based safety reporting systems [1, 7, 31, 49] are increasingly used by organizations for risk management. For instance, many universities and public organizations offer mobile applications that allow their community members to submit text reports about potential risks, concerning situations, or suspicious activities [1, 7, 31]. Additionally, city police agencies have implemented text-to-911 that enable live chat with dispatchers [18, 26]. Recent works [45, 49] have demonstrated the effectiveness of text-based reporting systems, such as reducing the labor costs associated with safety service provision, enhancing user-perceived value, and fostering inclusivity by assisting marginalized groups in reporting risks. Despite their growing popularity and demonstrated benefits, there remains a significant gap in understanding the emotional dynamics and support delivery within these text-based reporting systems.

To address this gap, we conducted a comprehensive system log analysis of a text-based community safety reporting system, named LiveSafe [49], which serves over 200 higher education institutions. This analysis helped us identify patterns of emotional expressions and support delivered in text-based incident reporting. We also observed a longitudinal decline in the provision of emotional support across various organizations, which prompted the development of new technical solutions. Although recent studies have demonstrated the promise and significant implications of using AI and Large Language Models (LLMs) to provide emotional support in various application domains [12, 34, 42, 86], little is known about whether AI models could be trained to support dispatchers in providing emotional support. Therefore, building on the empirical findings and leveraging the real chat logs, we developed an LLM, called *dispatcherLLM*, fine-tuned for the domain of text-based safety incident reporting. It is designed to provide emotional support during incident handling by simulating human dispatchers' languages with appropriate emotional support. In this study, we address the following research questions:

- **RQ1**: *In what circumstances do users express negative emotions in text-based incident reports?*
- **RQ2**: *How do dispatchers provide emotional support in response to different types of incidents?*
- **RQ3**: *How can LLM-generated suggestions for dispatchers enhance the emotional support perceived by users in incident handling?*

In the remainder of this paper, we first present an overview of the methods employed for analyzing the system chat logs and the development of the *dispatcherLLM* model. We then detail the results and findings yielded from the analysis and evaluation of the proposed model. Finally, we discuss the findings and the implications for the future design of text-based reporting systems. We refer to the community members who submitted reports as "users" and the human agents from the safety organizations who handled the reports via text as "dispatchers."

Many CSCW scholars have explored community-sourced systems for public safety and risk management [5, 33, 45, 49]. Our study makes novel and significant contributions to the CSCW community by examining how recent advances in AI technology (e.g., LLMs) can be applied to enhance the communication between service providers and the public, with broader implications for AI-based community risk management systems [53]. The contribution of this study is four-fold:

- First, we performed an emotion analysis on real text-based incident reporting systems. New empirical findings showed that users' emotions tend to be more negative when reporting specific categories of incidents, such as those related to *Mental Health*. However, their messages generally become more positive through interactions with dispatchers.

- Second, we found that the delivery of emotional support by dispatchers was associated with factors related to the incident, service timing, and organizational characteristics. A surprising finding was that organizations with prolonged use of the system tended to provide less emotional support. Notably, emotional support tends to diminish during certain working hours of the day (e.g., 8 a.m. - 12 p.m. and 12 p.m. - 4 p.m.).
- Additionally, we developed and implemented an LLM fine-tuned using domain-specific chat logs to mitigate the two issues mentioned above. Our evaluation, using both automatic metrics and user surveys, demonstrated that our *dispatcherLLM* could provide more consistent and effective emotional support compared to both human dispatchers and other general LLMs (not trained with domain-specific chats) for various incidents.
- Lastly, our findings offer strong empirical and practical value and shed light on human-AI co-creation of public services [85]. Specifically, this research pioneers the use of LLM to improve emotional support in public safety services, while LLMs are being increasingly utilized across other service domains [4, 22, 48] to address the labor shortage. Given the sensitive nature of public safety, our goal is not to replace human workers with AI-based solutions. Instead, it is crucial to explore how LLMs can be leveraged to provide more consistent responses. We discuss several implications for the future design of text-based risk reporting systems, aiming to enhance service quality and alleviate dispatcher burnout.

2 Related Work

In this section, we present prior work on the importance and value of emotional support in safety incident reporting service. We then discuss existing systems and studies designing text-based risk reporting systems. Finally, we discuss recent works on AI-enabled systems providing emotional support.

2.1 The Role of Emotional Support in Existing Safety Incident Reporting Systems

In emergency response systems, the delivery of emotional support is critical, particularly due to the "emotional pain" experienced by individuals reporting traumatic or safety-related incidents [76]. This support, when delivered effectively by dispatchers, is crucial for successful information gathering, enhancing trust, and fostering cooperation from callers.

The benefits of such proficient emotional support are significant, contributing to the recovery of individuals from traumatic experiences [62]. Meanwhile, inadequate emotional support can negatively impact the dispatcher-caller relationship. Inappropriate questioning strategies, for instance, have been perceived as delaying assistance, leading to user frustration and reluctance in cooperation [19, 55]. This is further supported by findings that demonstrate that a lack of emotional support can result in callers perceiving questions as "face-threatening," thereby resisting responding to questions [66]. This body of research highlights the crucial role of emotional support in incident handling systems, not only in facilitating immediate information collection but also in establishing a positive, cooperative rapport between dispatchers and callers for effective emergency response.

However, there is currently a lack of examination regarding the delivery of emotional support within text-based reporting systems and its potential impact on the service delivery of safety organizations. Existing work only discussed the functional aspects of the conversations between users and safety administrative agents using text-based reporting systems [45], e.g., the response rates (how many of the reports were handled) and responsiveness (how quickly they were handled). The issues reported by users via mobile or web-based applications are frequently less urgent than those reported through the 911 emergency hotline, as evidenced in the study by Iriberri et al. [32]. These user-generated reports, also referred to as "tips" later in this study, typically detail less critical situations compared to the dire emergencies usually reported through the traditional 911 system. Prior studies also showed that a person's speech and voice could convey more information regarding the person's mental state and emotional experience [27, 56, 59]. Although the emotional aspect of incident reporting has been well discussed by prior studies, the previous accounts are limited to the context of emergency call-taking but not text-based reporting systems. Meanwhile, there has been a lack of empirical understanding about how people show their emotional state in text-based incident reporting and how service providers respond to different emotional states by delivering their service.

2.2 Community Risk Reporting Systems

Information and Communication Technology (ICT) has been increasingly utilized in risk management in public settings [63]. Several systems have been developed to enable the reporting of safety events and incidents by the community, including web-based forms, mobile apps, and systems that allow users to report safety concerns and risks through crowd-sourcing [8, 13, 32, 58]. Some systems also connect users with emergency service providers and automatically share their location, improving the effectiveness of getting assistance [2, 28, 35]. In addition, live-chat safety incident reporting features have been incorporated into risk management systems by various organizations [11, 49, 69]. Previous research has found that campus community safety apps with emergency texting features have a higher perceived utility by community members and a lower cost compared to traditional emergency communication systems such as blue-light emergency phone towers [57, 80]. Text-based reporting systems, such as text-to-911, have also been found to be beneficial for deaf and hard-of-hearing individuals and people with disabilities [18].

Existing work has also suggested that how technological systems are designed can greatly impact users' perception of the target tasks and even gradually shift their behaviors. The implementation and adaptation of newly introduced technological systems into an existing organization workflow have been a non-trivial problem widely discussed by existing studies. Tyre and Orlikowski [68] revealed the highly discontinuous nature of technological adaptation in organizations, which usually happens within a relatively brief window after the initial implementation of a new technology. A later study by Buchanan et al. [10] discussed the problem of decay in certain organizational changes, i.e. "initiative decay" where the gains from change are lost when new practices are abandoned. According to a study by Mendoza et al. [47], the ease with which users can access ongoing training and their ability to perceive a technology's utility are the two main factors that encourage the productive use of a newly introduced technology. Later work by Barki et al. [6] revealed that users tend to pick up adaptive behaviors in response to technological failures, which is beneficial to the organization in terms of efficiency by facilitating a better match between the system and the context in which it is being used. However, given the research regarding text-based reporting systems, existing studies do not sufficiently address the question of how users engage with agents from safety reporting organizations when utilizing these text-based systems by analyzing empirical data, particularly during conversations.

2.3 Emotional Supports with Conversational Systems

The capability to provide emotional support is vital for various conversational agents, including those involved in clinical support [14, 16, 72], customer service [22], social interactions [29, 38, 48, 78] and education [4]. Within the domain of Human-Computer Interaction, the integration of emotional dialogue takes on pivotal significance in enriching the natural dynamics of human communication. The acknowledgment of the vital role of emotions in human interaction prompts the integration of automated emotional dialogue systems, intending to bridge this gap and cultivate more intuitive and meaningful interactions between users and computers. Recently, the emergence of LLMs has led researchers to scrutinize their effectiveness in emotion classification for diverse applications. Some researchers strategize methods to elicit emotional support from LLMs, employing strategies such as prompting [40, 74] and data-augmented [84] fine-tuning. However, the utilization of emotional chatbots in the safety reporting domain remains relatively unexplored, primarily due to constraints related to accessing large datasets and annotated emotional information. In the safety reporting domain, AI-mediated conversational interaction with users can assist dispatchers' handling of negative emotions such as despair and feelings of danger and mitigate emotional labor [67, 73].

Despite the crucial role emotional chatbots could play in supporting users through challenging situations, their utilization in safety reporting remains relatively unexplored. To provide further understanding in this domain, we examined the text-based reporting process from a multi-stakeholder and longitudinal perspective that encompasses both dispatchers and users. Drawn from the findings of our analyses, we introduced and evaluated a fine-tuned LLM for the safety incident reporting domain. The insights then consequently informed our design recommendations for future AI-enabled text-based risk reporting systems.

3 Method

To address the proposed research questions, we conducted a comprehensive system log analysis of a real text-based safety incident reporting system, called LiveSafe[1], which has been used by more than 200 organizations, serving communities varying in location region and size. The chat log of each incident report consists of both users' messages and dispatchers' replies. The dataset allowed us to conduct the following analyses:

- To answer **RQ1**, what emotional states are presented in users' text reports of safety incidents, we analyzed the users' messages. Specifically, we applied emotion classification using the GoEmotions dataset [15] and the calculation of Polarity Scores. We also executed statistical analyses to identify contextual factors that were associated with users' emotions in the reports.
- To answer **RQ2**, whether dispatchers provide emotional support in their replies through the text-based reporting system, we conducted text analyses on the dispatchers' replies. We also employed statistical analyses to identify contextual factors that were associated with dispatchers' delivery of emotional support.
- To answer **RQ3**, we fine-tuned the Llama-2 model [65] to yield a *dispatcherLLM* (Language Learning Model), which can be used to suggest replies by simulating human dispatchers' emotion support languages. We further completed an evaluation by comparing our proposed *dispatcherLLM* with existing LLMs and showed the improved performance of delivery emotional support. We also presented results where participants rated their preferred responses and explained their choices.

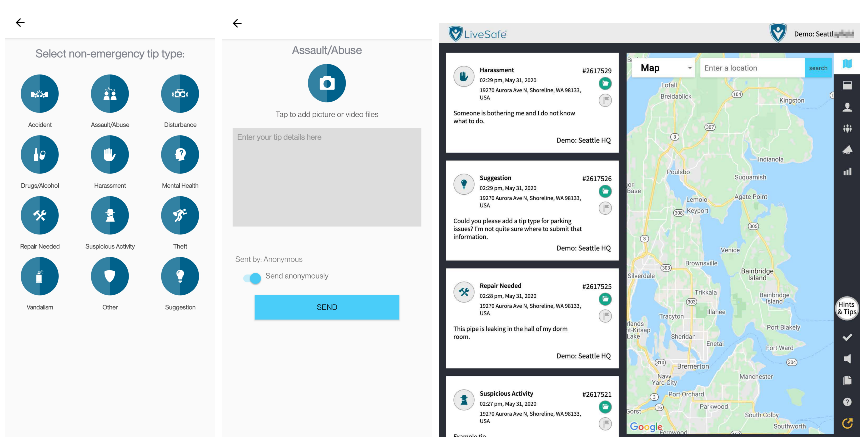
Below, We provide details of the dataset, the analysis procedure, and the method used for fine-tuning the *dispatcherLLM* model.

3.1 Dataset

The chat log was collected through the LiveSafe app across the time period between 2018 and 2019, which allows users to report certain events to a human agent via a chat service. LiveSafe is a risk management system that aims to facilitate communication between members of an organization and the safety teams responsible for managing risk [49]. This tool, which was first introduced in 2013, has been adopted by over 200 higher education institutions. Members of the community can submit information or "tips" through the LiveSafe mobile app or web portal, while safety

[1]https://www.vectorsolutions.com/solutions/vector-livesafe

organization dispatchers can respond to these tips through the LiveSafe Command Dashboard. Tips can take the form of text, photographs, video, or audio recordings and can be classified into different "tip types," which are chosen by local safety departments for each organization, as shown in Fig. 1. After a user submits a tip, dispatchers from safety departments can engage in a conversation with the user through the Command Dashboard, as shown in Fig. 1c).



(a) Tip Categories (b) Submitting Tip (c) Dispatcher's Dashboard

Fig. 1. The user interface of the LiveSafe mobile app; a) The user needs to select a tip category after choosing to submit a tip; b) After the tip category is chosen, the user can submit text and image descriptions and connect to a human agent; c) The dispatcher can connect and chat with the user using the Command Dashboard.

3.1.1 Data ethics. Ethical considerations pertaining to data were taken into account before and throughout the entire research process due to the potential for incident reporting logs to contain private and sensitive information. Several steps have been taken to ensure that data ethics, such as anonymization, data minimization, and consent, were properly addressed.

First, the data was carefully anonymized by the organization before being accessible to the researchers for conducting the log analysis. The organizations and users were represented using anonymous IDs, and all mentions about geolocations, times and person/organization names in the chat log text were detected and replaced with tags (e.g., [LOCATION], [NAME]). As an additional measure of anonymization, a random selection of universities was further removed entirely from the dataset by LiveSafe. The dataset was fully anonymized.

Second, the researchers of this study did not actively collect the data, nor did they have access to any identifiable user information. The data was provided by the LiveSafe organization to researchers for analysis. An NDA (non-disclosure agreement) was established between the researchers and the LiveSafe organization. The study examined information from organizations and individual users who had consented to the use of anonymized and aggregated data for system improvement research.

Third, our research strictly adhered to the principle of data minimization, limiting our data collection to only what was essential for the study. All personal and sensitive information had been removed, and identifiable information, including user, geolocation, and organization information, had been removed or anonymized prior to researchers gaining access to the data.

3.1.2 Data cleaning and pre-processing. To reduce the noise within the chat log data, we first cleaned up the data by removing "Test" tips and irrelevant/ambiguous categories, which include incidents from categories including *SafeRide; 911 / Call; Broadcast Message; Scavenger Hunt / Test; Operational Procedure Log; Broadcast Check-in Drill; Broadcast Check-in; Request Security Presence / Walk-through; Other; Misc.* We only kept tips after 2018 (between 2018-01-01 and 2019-12-31) and chats with more than 2 utterances (one conversation turn), since we want to investigate chats with at least one turn of information-collecting conversation. Finally, for each tip we performed language detection using *langdetect* [61] to identify messages in English. We identified 16.2% (1,588 out of 9,827) of the tips to be non-English, and removed them to simplify further analysis.

After performing cleaning, the resulting dialogues are categorized into different tip categories, which individuals selected before initiating a chat conversation. The dataset consists of 8,239 incidents, involving 18 different incident categories, where the top 10 categories of incidents with highest tip counts are *Noise Disturbance* (N=2,282), *Suspicious Activity* (N=1,700), *Emergency Message* (N=1,133), *Drugs/Alcohol* (N=1,031), *Facilities/Maintenance* (N=593), *Harassment/Abuse* (N=455), *Accident/Traffic/Parking* (N=348), *Theft/Lost Item* (N=269), *Mental Health* (N=171) and *Vandalism/Damage* (N=90). Sensitive information (names, places, times, etc) in the dataset was masked for privacy. It is worth noting the impacts this might have had on the annotation process and the model performance, as it is not always clear what the masked value was (e.g. "The incident must have occurred sometime between [TIME] and [TIME] which is when I got out of work.").

The dataset provides abundant conversation logs for further analysis. To distinguish it from the later annotated sample, we refer to this dataset as all data, containing a total of 8,239 conversations and 57,114 utterances.

3.2 Detecting Emotions Involved in the Chats

In order to gain a deeper understanding of the emotional dynamics during conversations between users and dispatchers, we use an ML-based method to classify and quantify the emotions of the chat utterances. In this section, we discuss the methods used in this study to 1) classify the emotion of each chat utterance; 2) quantify and aggregate emotion negativity on a conversational level; and 3) identify dispatchers' delivery of emotional support.

3.2.1 Classifying emotions with RoBERTa and identifying emotional support. Following the approach of [45], we used a RoBERTa-based model to perform emotion classification to identify whether dispatchers provided emotional support [70]. The RoBERTa model was finetuned using the GoEmotions dataset introduced by Demszky et al. [15]. The dataset contains 58K Reddit comments annotated with 28 types of fine-grained emotions (including *neutral*). We refer to the classification of positive/neutral/negative emotions as described in the GoEmotions dataset. The fine-tuned model was able to classify text into the 28 fine-grained emotion classes, reaching a satisfying overall accuracy of 93.5% when evaluated on the GoEmotions Dataset.

To further identify the emotional support within dispatchers' utterances, we utilize the emotion classification model obtained above. We employed a similar method to identify emotional support as used in Khanpour et al. [37]'s work within the domain of online health communities. After applying the emotional classification model over all dispatchers' chat utterances (N = 33,718), we consider the utterances classified under the emotions, including "caring," "love," "sadness," "remorse," and "grief," as providing emotional support according to the definition of Weber et al. [75]. For example, by applying the above method, we identified the following chats where the dispatcher showed emotional support to an incident labeled as *Emergency Message*, where a dispatcher reassured the caller by texting, "*I'm with you and officer [PERSON] and officer [PERSON] will be right there I promise.*", which was classified as caring by the model. In a different scenario, the dispatcher texted "*Ok. I am sorry to hear that happened to you,*" when the user reported a *Theft / Lost Item* incident, which was identified by the model as sadness.

3.2.2 Measuring conversational emotion polarity. To further measure the emotional polarity of users throughout conversations, we quantify and aggregate the negativity of users' utterance emotions by conversation specifically for the context of safety reporting. We identified a number of heuristics to better capture the traits of the safety reporting domain after qualitatively analyzing the chat logs. Among these heuristics are: 1) The majority of users have bad emotions when they first start a chat, and dispatchers usually fix or settle them gradually throughout conversations. 2) If a user still has negative emotions at the end of the session, that is more concerning.

Drawing from existing literature [23, 82], we use a measurement metric for quantifying the negativity of users' utterances throughout a single reporting conversation by calculating the relative frequency of users' utterances with negative emotions. We define negative emotions as in the work of Demszky et al. [15], which consists of fine-grained emotions including *anger*, *annoyance*, *disappointment*, *disapproval*, *disgust*, *embarrassment*, *fear*, *grief*, *nervousness*, *remorse* and *sadness*. We modified the emotion polarity score metric with an increasing weight toward the end of each conversation based on our heuristics, which can be denoted as:

$$S_n = \frac{\sum_i^N s_i e^i}{\sum_i^N e^i} \tag{1}$$

where n denotes the n^{th} conversation with a total count N of user utterances. s_i refers to the emotional polarity of the $i^t h$ user utterance, i.e. -1 for negative emotions and 0 for neutral and positive emotions. The resulting polarity score S_n represents the user's level of negative emotion during the n^{th} conversation. For example, with user utterances with emotions [fear, confusion, neutral, curiosity, gratitude], s will be [-1, -1, 0, 0, 0]. The resulting polarity score is denoted as $S \in [-1, 0]$, indicating how negative the user's overall emotion is throughout the reporting conversation.

3.3 Measuring Information Collection through Event Argument Extraction

In order to further analyze how dispatchers collect incident-related information from users during conversations, we introduced a BERT-based model to automatically extract event argument information taking the conversation utterance text as input. In the following section, we first introduce how an event ontology was introduced for safety reporting purposes. Then, we describe the model used for automatic event argument extraction, which was fine-tuned and evaluated using our proposed event ontology.

3.3.1 Event ontology. To further understand how dispatchers collect incident-related information from users during conversations, we refined the event argument schema based on the Automatic Content Extraction (ACE) dataset [71] and adapted it to the domain of safety incident reporting. The ACE dataset is an event extraction task dataset widely used by prior studies [41, 43, 77]. The dataset in question comprises annotations pertaining to event and entity information obtained from news articles. While the dataset does partially encompass events that are relevant to the domain of safety, such as *Attack* with arguments including *Attacker*, *Target*, *Instrument*, *Time*, and *Place*, the original ACE dataset schema falls short in its ability to directly annotate incidents related to safety due to its coarse-grained nature. To address this issue, we have modified the ACE event schema and subsequently employed it in the analysis of LiveSafe chat logs. Based on the initial ACE event argument schema, an exploratory annotation process was undertaken by the researchers. During the annotation process, we identified a set of event-related arguments, including *Attacker*, *Target*, *Location*, *Weapon*, *Start Time*, *End Time*, and *Target Object*. These arguments were linked to the event handle, constituting a vital component of the event information. The complete ontology and definition can be found in Appendix A.

3.3.2 Model. To capture the event information from the conversations between users and dispatchers, we established an ML-based model to perform event argument extraction. The model is based on the safety reporting event ontology introduced in prior work by Liu et al. [45]. Our event argument extraction model aims to extract certain information about a safety incident given the dialogue history. In order to achieve this, we have identified and selected key fields that are related to the categories of interest, as in the ontology obtained in 3.3.1, and can be extracted as spans of text from the dialogue history. Our model design adheres to the zero-shot setting of STARC [21], which is a state-of-the-art method that can be applied to new domains without additional fine-tuning.

The system design involves the choice of questions for each field, the QA model selection, and the decoding of QA outputs into event argument extraction outputs. Several phrasings of natural language questions were compared for each slot. For example, for the "object stolen" slot on theft, the questions "What was stolen?" and "What object was stolen?" were compared, and the second phrasing scored higher. For every new utterance in the dialogue, we predict a new dialogue state consisting of a list of predicted values for each pre-defined slot. The dialogue history until the utterance is passed to the QA model, along with one question for each slot. Valid answers (controlled by the hyper-parameters of a minimum score and maximum length) are appended to an ongoing list of answers for each slot.

3.4 *DispatcherLLM*: A LLM for Generating Scenario-Based Dispatcher Responses

To explore the feasibility of improving both informational and emotional support of incident report handling using AI, we fine-tuned an LLM on safety reporting chat logs from *LiveSafe*. The model's generation output is evaluated using both automatic metrics and human evaluation with community members.

3.4.1 Instruct-tuning with chat log. A Llama-2 model [65] was fine-tuned and evaluated using the same set of chat logs analyzed in this study. We used the 7B instruction-tuned version of the Llama-2 model released by Meta. We chose the instruct-tuned Llama-2 7B model because of its open-sourced nature and ability to be replicated and validated in various application scenarios. Meanwhile, the instruct-tuned Llama-2 model has been aligned to handle user instructions and deliver emotional support for general applications [30]. In this study, we further fine-tuned the model to adapt to the task space of incident report handling while utilizing the model's learned ability to provide emotional support. The model is fine-tuned on a training partition of 31,359 dispatcher utterances. A validation set consisting of 10,453 dispatcher utterances was used to assess the generalization of the trained model. We refer to the resulting model as *dispatcherLLM*. Since the instruct-tuned model should be capable of accommodating various incident scenarios, we generate a summary of incident scenarios using GPT-3.5 for each chat log. The generated summaries were then incorporated as scenario descriptions during the fine-tuning process. During supervised fine-tuning, we fine-tuned the model to generate each dispatcher response given the incident scenario summary and dialogue history. Specifically, dialogue history and incident summary are pre-filled as pre-formatted prompts for Llama-2 and then aligned with the dispatcher's utterance by each turn. In the inference phase, our system retrieves the most similar document from the training set, based on both the summary and dialogue history, in order to enhance the *dispatcherLLM*'s responses. An example of the constructed prompt and dispatcher responses generated by the model can be found at A.2.

3.4.2 Human evaluation. To further evaluate the delivery of effective emotional support, we conducted a survey study with community members from a higher education institution. We recruited 17 undergraduate students from a U.S.-based university community and designed a survey to collect feedback by asking them to compare responses generated by the instruct-tuned model with both responses generated by GPT-3.5 and human dispatchers' responses. The participants are instructed to compare the outputs from GPT-3.5 and *dispatcherLLM* with human dispatchers' responses and choose the better response based on different scenarios randomly sampled from the test set. We also collected users' qualitative feedback and ratings on the emotional support provided by the two models' generated responses. The findings of the survey study are presented in 4.3.2. The survey details and included samples can be found in the appendix (A.3 and A.4). This study aimed to gather human feedback on the generated emotional support to inform future research on model enhancement and practical applications. It is not intended to be exhaustive or conclusive.

4 Findings

4.1 Users' Emotions Vary by Incident Type and Reporting Stage (RQ1)

We first examine the emotional dynamics within the conversation between users and dispatchers. In order to conduct a more comprehensive analysis, we base our analysis on the results obtained from the emotion detection model described in 3.2. Using the measurement introduced in 3.2.2, we calculated the emotional polarity score for all incident reporting conversations. The resulting polarity scores ranged between -0.75 and 0. The results showed that while most tips (N=6,623, 82.7%) contained no negative emotions from users' utterances, tips with extreme polarity scores less than or equal to -0.5 (N=262, 3.3%) were also found. To further understand what led to users' negative emotions, we conducted analysis to identify the different contextual factors, including tip category, anonymity, years in use and tip volumes, associated with users' emotional statuses.

4.1.1 Users expressed more negative emotions for certain incident types . For most tip categories, the first messages from users were often neutral in emotion. This suggests that users may initially approach the dispatcher with a calm and collected demeanor in the majority of incident types reported. The results also show that users tend to express more gratitude for incidents that are less urgent in nature, such as facilities/maintenance, noise disturbance, theft/lost item, and vandalism/damage. This may be because these types of incidents are typically less severe and do not require immediate action, allowing the users to feel more grateful for the assistance provided by the dispatcher. In contrast, for incidents related to mental health and injury/medical, users tend to express more caring emotions. This may be because these types of incidents are often more serious and require a more compassionate response from the dispatcher. The higher level of caring emotions in these situations may reflect the users' concern for their own well-being or the well-being of others involved in the incident.

As shown in Figure 2, it is also found that the average polarity scores differ across different incident categories. A one-way ANOVA indicated a significant disparity across incident categories ($F(17, 8221) = [4.65], p < 0.001^{***}$). Incidents that are more urgent in nature or related to personal safety tend to incur more exhibited negative emotions from users ($t(722) = [-2.13], p = 0.034^{*}$), e.g. *Harassment / Abuse* ($M = -0.06, SD = 0.13$) vs. *Theft/Lost Item* ($M = -0.04, SD = 0.11$). Meanwhile, *Mental Health* ($M = -0.07, SD = 0.14$) related incidents have also been found to have more negative emotions in users' utterances during reporting.

To further examine the factors associated with users' negative emotions, we conducted a linear regression analysis over whether users' first message was emotionally negative. As shown in Table 1, we found that users' negative emotions are associated with certain categories of events, such as *Emergency Message* and *Mental Health*. The results of the regression revealed no association between users' negative emotions and organizations' years in use of the LiveSafe system, and the average number of tips received per year by the organization. The findings have shown that users' negative emotions only varied by the factors related to the nature of incidents and the design

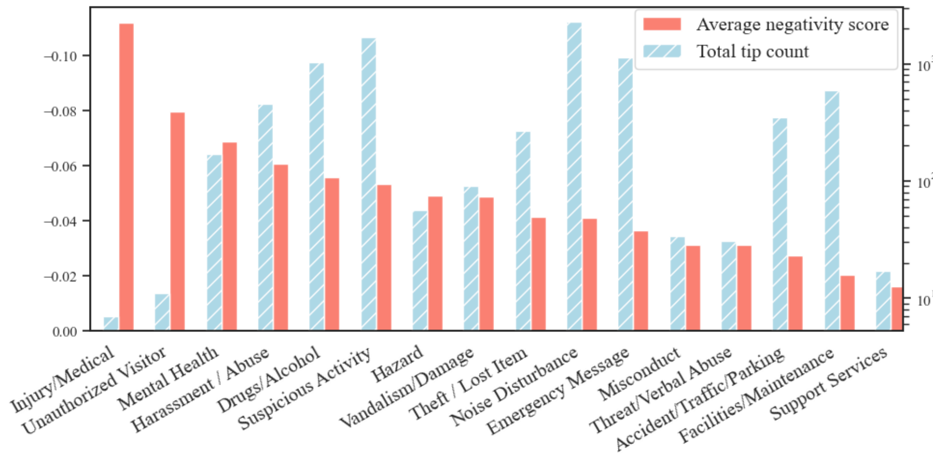


Fig. 2. Average negativity score and total tip count by tip category. ANOVA revealed a significant difference ($F(17, 8221) = [4.65], p < 0.001^{***}$) in user's emotional polarity across incident categories during reporting.

choices of the reporting systems, but not by characteristics of the organizations, which is reflected through the years in use since implementation (experience) and average tips per year (usage and scale of the system).

4.1.2 Users express more positive emotions with dispatchers involved during conversations. During the reporting conversations, users first provide an overall description of the incidents. Dispatchers then join the conversations to further collect more information and provide assistance to address users' concerns. To further understand users' emotion changes during this process, we conducted a sentiment analysis based on the emotion classification results described in 3.2. Each conversation is divided into three stages — the initiation stage (first 1/3 of utterances), the information-gathering stage (middle 1/3 of utterances), and the elaboration stage (last 1/3 of utterances). As shown in the results of the emotion analysis in Figure 3, we discovered an increase in users' positive sentiment as conversations proceeded. A Chi-Square Goodness of Fit Test showed that the sentiment distributions exhibited significant differences among the three stages of conversations ($X^2(4, 8239) = [1160.34], p < 0.001^{***}$). This is reflected in a consistent increase in the ratio of utterances with positive sentiment ($t = [-31.369], p < 0.001^{***}$), and a decrease in ratios of utterances with both neutral and negative sentiment. The decrease in users' neutral utterances implied the process of users providing incident-related information during the earlier stage of conversations, which later led to a higher proportion of positive utterances as dispatchers provided assistance after information was collected. Although a decrease in negative sentiment is observed, the magnitude of the change is not as obvious, which might imply a higher difficulty for dispatchers to address users' negativity in sentiment or a lack of focus in providing emotional assistance during incident report handling.

Summary (RQ1): In terms of users' emotions conveyed in their incident report messages, our findings showed that when reporting specific types of incidents, such as *Mental Health* and *Emergency Message*, users were more likely to start with negative emotions, as indicated in Table 1. However, their emotions tended to become less negative towards the end of their interactions with the dispatchers, as shown in Fig. 3.

	Model 1		Model 2		Model 3	
	Coeff.	**(S.E.)**	**Coeff.**	**(S.E.)**	**Coeff.**	**(S.E.)**
Event category						
Suspicious activity	REF					
Accident/Traffic/Parking	-0.17	(0.35)	-0.15	(0.41)	-0.15	(0.40)
Contact Mall/Corporate/Property Security	**-38.51**	**(0.24)**	**-17.97****	**(0.48)**	**-13.93****	**(0.48)**
Drugs/Alcohol	0.02	(0.23)	-0.01	(0.22)	-0.02	(0.23)
Emergency Message	**0.81*****	**(0.18)**	**0.81*****	**(0.25)**	**0.79*****	**(0.25)**
Facilities/Maintenance	0.30	(0.24)	0.31	(0.22)	0.29	(0.22)
Harassment/Abuse	0.30	(0.27)	0.28	(0.24)	0.27	(0.24)
Hazard	0.08	(0.73)	0.08	(0.81)	0.01	(0.81)
Injury/Medical	1.51	(1.09)	1.55	(0.97)	1.46	(0.97)
Mental Health	**1.97*****	**(0.24)**	**1.95*****	**(0.26)**	**1.96*****	**(0.26)**
Misconduct	**-16.09**	**(0.67)**	**-23.88***	**(0.50)**	**-23.85***	**(0.50)**
Noise Disturbance	-0.28	(0.20)	-0.33	(0.24)	-0.31	(0.24)
Support Services	**1.74****	**(0.66)**	**1.76.**	**(0.95)**	**1.71.**	**(0.94)**
Suspicious/Unattended Package	**-14.80**	**(0.86)**	**-15.38****	**(0.82)**	**-18.64****	**(0.82)**
Theft/Lost Item	**0.88****	**(0.27)**	**0.85****	**(0.27)**	**0.87****	**(0.27)**
Threat/Verbal Abuse	**-1.90**	**(2.49)**	**-59.56***	**(0.33)**	**-31.56***	**(0.33)**
Unauthorized Visitor	1.16	(1.06)	1.25	(0.93)	1.19	(0.93)
Vandalism/Damage	0.01	(0.60)	-0.02	(0.74)	-0.05	(0.75)
Anonymous	-0.39	(0.31)	-0.40	(0.33)	-0.21	(0.14)
Time of day						
4 a.m. - 8 a.m.	-0.41.	(0.19)	-0.41.	(0.16)	-0.40	(0.33)
8 a.m. - 12 p.m.	0.14	(0.18)	0.13	(0.19)	-0.41.	(0.16)
12 p.m. - 4 p.m.	-0.03	(0.20)	-0.03	(0.18)	0.14	(0.19)
8 p.m. - 12 a.m.	-0.12	(0.17)	-0.12	(0.15)	-0.04	(0.18)
12 a.m. - 4 a.m.	0.00	(0.00)	0.00	(0.00)	-0.13	(0.16)
chat length	-0.19	(0.12)	-0.20	(0.14)	0.00	(0.00)
Org. years in use			0.07	(0.05)	0.08	(0.05)
Org. tips received per year					0.00*	(0.00)

. $p < 0.05$; * $p < 0.01$; ** $p < 0.005$; *** $p < 0.001$ (two-tail); S.E. clustered by organization

Table 1. The results of a linear regression present the factors that are associated with users' emotional negativity (N = 8,239). The higher the coefficient, the more likely users were to express negative emotions (e.g., for mental health and emergency incidents); Models 1, 2, and 3 gradually introduced new variables (anonymity, years in use, and tip volume) to demonstrate robustness.

4.2 Factors Associated with the Delivery of Dispatchers' Emotional Support (RQ2)

Whether dispatchers decided to provide emotional support to users during reporting conversations depends on many factors. Since this is often not included as a standard procedure in safety departments' playbooks, it is vital to understand what factors were associated with dispatchers' decision making in providing emotional support in practice. To examine this, we conducted a logistic regression analysis over the relationship between whether a dispatcher provided emotional support and a range of potential factors involved in the decision-making process.

4.2.1 Tip-related factors. As shown in Table 2, the delivery of dispatchers' emotional support varied by several tip-related factors, e.g., tip categories, anonymity of the tips, and emotional polarity of the users' first messages. An inconsistency of emotional support across different tip categories was revealed by the results of the logistic regression. For example, users reporting incidents under the categories of *Harassment/Abuse* and *Mental Health* were found to be more likely to receive

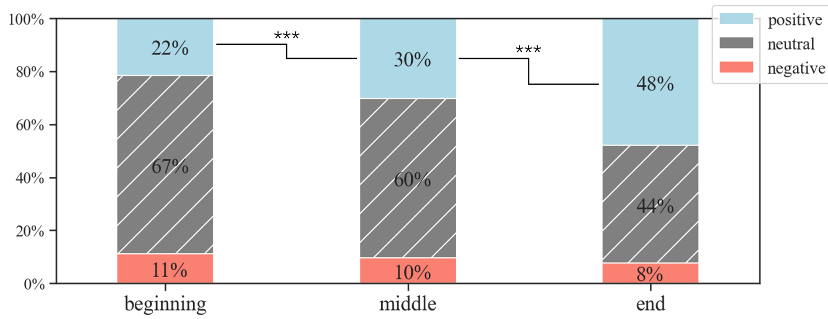


Fig. 3. Change in the sentiment ratio of users' utterances throughout conversations: ***beginning*** refers to the first 1/3 of each conversation, ***middle*** refers to the 2/3 of each conversation, and ***end*** refers to the last 1/3 of each conversation. The t-tests indicate a significant increase in users' positive sentiment as their conversations with dispatchers progressed.

dispatchers' emotional support compared to *Drugs/Alcohol* and *Noise Disturbance*. Meanwhile, tips where users reported anonymously were found less likely to be provided with emotional support. We also found that users with a stronger negative emotion display in their messages were more likely to receive emotional support from dispatchers.

4.2.2 Service-related factors. In order to gain deeper insight into the distribution of reported tips across various time periods, a Chi-Square Goodness of Fit Test was conducted to test the difference in the distribution of tips throughout different segments of the day. The results of the test indicated that there is a significant difference in the distribution of user-reported incident types among the different time periods of the day ($X^2(30, 8239) = [1245.21], p < 0.001^{***}$). More specifically, higher volumes of incidents were being reported during periods of *Afternoon*, *Evening*, *Night*, and *Late Night*. The increase in volume is especially observable during the *Night* with a drastic increase in incidents related to *Noise Disturbance* and *Drugs/Alcohol*. An increased amount of *Harassment/Abuse* incidents was also observed during *Afternoon*, *Evening*, and *Night*, while significantly smaller than the amount of other incidents such as *Noise Disturbance* and *Drugs/Alcohol*.

The results also showed that two service-related factors were associated with the delivery of emotional support, as shown in Table 2. The results from Model 1 of Table 2 also showed additional inconsistency in emotional support based on when the report happened with respect to the time of day. Generally, we found that dispatchers tended to provide less emotional support during normal working hours during the day (i.e., 8 a.m. - 12 p.m. and 12 p.m. - 4 p.m.). Since there might be a difference in tip volumes of different tip categories reported during each time period of the day, it might affect dispatchers' behavior and strategies when handling users' incident reports. Due to the different volumes of incidents during each time period, dispatchers on each shift might experience burnout at varied levels. The findings suggested that shifts with higher incident report volumes could lead to greater dispatchers' burnout (e.g., *Afternoon*, *Evening* and *Night*), thus decreasing the willingness for dispatchers to provide emotional support. The findings also revealed that dispatchers' differences in the delivery of emotional support might be affected by the volume of different reported tip categories at their corresponding shifts of the day. For example, dispatchers during *Night* shifts might tend to deliver less emotional support since increased reports of *Drugs/Alcohol* incidents. As shown by the logistic regression analysis results in Table 2, this effect impacts not only the emotional support delivery of the increased tip category, but also other incident reports under the same time period/shift.

	Model 1		Model 2		Model 3	
	Coeff.	**(S.E.)**	**Coeff.**	**(S.E.)**	**Coeff.**	**(S.E.)**
Event category						
Suspicious activity	REF					
Accident/Traffic/Parking	-0.32	(0.20)	-0.30	(0.20)	-0.27	(0.19)
Contact Mall/Corporate/Property Security	0.22	(1.10)	0.13	(1.13)	0.18	(1.13)
Drugs/Alcohol	**-0.52.**	**(0.16)**	**-0.47.**	**(0.16)**	**-0.47.**	**(0.16)**
Emergency Message	**0.64*****	**(0.12)**	**0.63*****	**(0.12)**	**0.64*****	**(0.12)**
Facilities/Maintenance	0.08	(0.21)	0.11	(0.21)	0.14	(0.21)
Harassment/Abuse	**0.64*****	**(0.18)**	**0.68*****	**(0.19)**	**0.67*****	**(0.18)**
Hazard	-0.69	(0.42)	-0.58	(0.44)	-0.58	(0.45)
Injury/Medical	0.50	(1.17)	0.41	(1.18)	0.37	(1.14)
Mental Health	**0.54.**	**(0.21)**	**0.58.**	**(0.20)**	**0.57.**	**(0.21)**
Misconduct	**0.98***	**(0.33)**	**0.88.**	**(0.35)**	**0.91.**	**(0.36)**
Noise Disturbance	**-0.42.**	**(0.19)**	**-0.45***	**(0.16)**	**-0.43***	**(0.16)**
Support Services	0.06	(0.60)	0.10	(0.65)	0.14	(0.64)
Suspicious/Unattended Package	**-16.90***	**(0.76)**	**-19.74***	**(0.77)**	**-18.00***	**(0.77)**
Theft/Lost Item	**0.44.**	**(0.19)**	**0.47***	**(0.18)**	**0.49***	**(0.18)**
Threat/Verbal Abuse	**1.17****	**(0.39)**	**1.10****	**(0.42)**	**1.13****	**(0.41)**
Unauthorized Visitor	0.11	(0.93)	-0.05	(0.95)	-0.11	(1.01)
Vandalism/Damage	**-1.15.**	**(0.51)**	**-1.04.**	**(0.52)**	**-1.03.**	**(0.52)**
Anonymous	**-0.38*****	**(0.09)**	**-0.38*****	**(0.09)**	**-0.40*****	**(0.09)**
Time of day						
4 a.m. - 8 a.m.	REF					
8 a.m. - 12 p.m.	**-0.65*****	**(0.25)**	**-0.66*****	**(0.23)**	**-0.66*****	**(0.23)**
12 p.m. - 4 p.m.	**-0.51*****	**(0.20)**	**-0.53*****	**(0.19)**	**-0.53*****	**(0.19)**
8 p.m. - 12 a.m.	-0.34	(0.20)	-0.35	(0.19)	-0.35	(0.19)
12 a.m. - 4 a.m.	-0.32	(0.23)	-0.32	(0.22)	-0.32	(0.22)
4 a.m. - 8 a.m.	-0.12	(0.20)	-0.12	(0.19)	-0.14	(0.19)
Org. years in use			**-0.13*****	**(0.03)**	**-0.13*****	**(0.03)**
Org. Tips received per year			0.00	(0.00)	0.00	(0.00)
User's emotional polarity					**-1.13*****	**(0.33)**

. $p < 0.05$; * $p < 0.01$; ** $p < 0.005$; *** $p < 0.001$ (two-tail); S.E. clustered by organization

Table 2. Results of a logistic regression analysis present the factors that are associated with emotional support delivered for incident reports (N = 8,239). For example, dispatchers were more likely to provide emotional support when users reported Harassment/Abuse incidents and less likely when users reported Suspicious/Unattended Package incidents. Meanwhile, dispatchers were less likely to provide emotional support during normal working hours between 8 am to 4 pm. Models 1, 2, and 3 gradually introduced new variables (years in use, tip volume, and users' emotion polarity score) to demonstrate the robustness.

4.2.3 Organizational Factors. Also, we found that the likelihood of dispatchers providing emotional support to users was negatively associated with the amount of time that passed since organizations implemented their incident reporting systems. This might suggest that over time, dispatchers become more focused on resolving incidents and less likely to provide emotional support, which further indicates a possible decay in dispatchers' awareness of the delivery of emotional support when handling users' incident reports. This effect was also found to be consistent regardless of the scale of each organization, which is measured by the average amount of tips received by each organization per year in both Model 2 and Model 3. The findings implied that the potential temporal decay in dispatchers' emotional support might be generalizable and not limited to organizations with certain sizes and scales. Additionally, we also found that when the absolute value of users' polarity score is higher, as shown in Model 3 of Table 2, which indicates that users are more likely to have negative emotions remaining by the end of conversations, dispatchers tend to provide more emotional support.

Our analysis of conversation logs from a text-based safety incident reporting system revealed a decline in emotional support provided by dispatchers across different tip categories. However, we found that this decline was more pronounced for incidents that were more urgent, such as *Emergency Messages*, *Harassment / Abuse*, and *Suspicious Activity*. On the other hand, this decline in emotional support was less noticeable for less-urgent events, such as theft or lost items. Additionally, our analysis showed that organizations with higher age/experience tend to provide less emotional support. While technology can be beneficial in terms of providing informational support (e.g., collecting information from users), our findings suggest that it may also lead to a degradation in emotional support from the first year of implementation. In an attempt to establish a systematic practice and positive word-of-mouth among users, organizations might incur an above-average level of service during the beginning of the implementation, while gradually reducing to a baseline during the years to come.

Summary (RQ2): To understand what contextual factors were associated with the dispatchers' delivery of emotional support, we conducted a logistic regression analysis of dispatchers' messages. We identified factors related to emotional support delivery from three different aspects: tip-related factors, service-related factors, and organizational factors. From a tip-related perspective, dispatchers' likelihood to provide emotional support to users during reporting conversations was found to be associated with tip category, anonymity, and users' emotional polarity. For example, dispatchers tended to provide more emotional support for incidents under the categories of *Harassment / Abuse* and *Mental Health*, but less for incidents related to *Drugs / Alcohol* and *Noise Disturbance*. From a service-related perspective, dispatchers were found to provide less emotional support during normal working hours during the day. From an organizational perspective, a significant negative association was also found between the likelihood of providing emotional support and the number of years the organization had used the safety reporting system, suggesting a decline of emotional support delivery after the initial period of system usage.

4.3 DispatcherLLM Providing Informational and Emotional Support (RQ3)

To investigate the potential of using large language models to improve the consistency of emotional support delivery, we introduced a domain-adapted LLM using the LiveSafe chat log dataset, and evaluated its effectiveness by comparing it with other LLMs.

4.3.1 Outperforming Popular LLMs for Overall Responses. After training the *dispatcherLLM* model following the method presented in section 3.4, we first evaluated its performance in handling incident reports. For example, Fig. 4 shows a real chat between a user and a human dispatcher,

Table 3. *dispatcherLLM* refers to our fine-tuned Llama 2 with retrieval-augmented dialogue generation. Our *dispatcherLLM* outperforms the original Llama-2 and GPT-3.5 on the generated utterances across nine incident types that were used to train the model.

Comparison of models over all incidents		
Model	**ROUGE**	**BERTScore**
dispatcherLLM	**19.89**	**48.66**
Llama-2	6.20	34.72
GPT-3.5	6.06	40.09

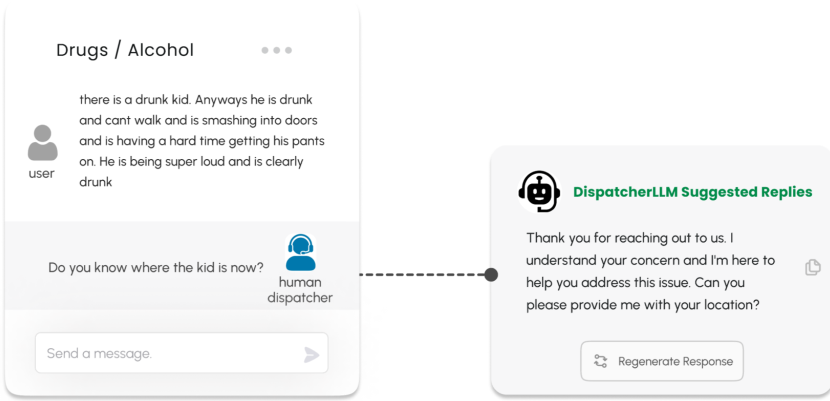


Fig. 4. Our *dispatcherLLM* is used to automatically generate responses, given a user message. The left side of the figure shows an actual chat log between a user and a dispatcher, regarding a *Drugs / Alcohol* incident. On the right, the figure displays a response generated by our *dispatcherLLM*. This illustration serves not only to compare the two types of responses but also to demonstrate a potential use case scenario in which *dispatcherLLM* suggests responses to human dispatchers, thereby enhancing the quality of their replies.

Table 4. 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^{*}

when the user reported a suspicious activity. In the log, the human dispatcher only asked for location in a straightforward tone, whereas *dispatcherLLM* provided emotional support (*empathy display* according to [45]) before asking for location information. We then calculated the similarity between the suggested responses and those of human dispatchers' responses in the system log. We then compare these scores across 10,453 incidents against those generated by two popular LLMs, i.e., Llama-2 and GPT-3.5. Two commonly used metrics for evaluating language generation, ROUGE and BERTscore, [20, 50, 81] are used to show the performance. As shown in Table 3, *dispatcherLLM* yielded much better results than the other two models.

4.3.2 Improved Emotional Support from dispatcherLLM. We further compared the emotional support rates between human dispatchers and *dispatcherLLM*. We used the same method described in section 3.2.1 to detect emotional support in the responses generated by *dispatcherLLM*. The support rate was defined by the ratio of messages providing emotional support from human dispatchers or *dispatcherLLM*. We compared the emotional support rates from two perspectives: incident category and time of service.

First, we ran paired t-tests to assess the differences in emotional support between humans and *dispatcherLLM*. Table 4 presents the average support rate for different incident types. The results show that *dispatcherLLM* significantly increased emotional support in responses for six types of incidents, including: *Suspicious Activity*, *Drugs / Alcohol*, *Emergency Message*, *Harassment / Abuse*, *Noise Disturbance*, and *Theft / Lost Item*. On the other hand, human dispatchers showed a higher support rate in the *Mental Health* category. Referring to the findings in RQ1 (Table 1) and RQ2 (Table 2), it was observed that users did not express strong emotions in their messages when reporting *Accident / Traffic / Parking* incidents. In contrast, users' emotional states were significantly more negative when reporting Mental Health incidents. This polarity in emotional states likely influenced the performance of *dispatcherLLM*.

The second notable improvement concerned the temporal consistency is *dispatcherLLM*'s delivery of emotional support. A Levene's Test of Variances ($F = 780.27, p < 0.001^{***}$) showed that the LLM ($Z = 0.51$) displayed a more consistent level of positive emotions throughout different hours of the day than humans ($Z = 0.54$). There is an interesting deficiency in emotional support provided by human dispatchers during daytime hours. Humans tend to become more emotionally supportive between 9:00 PM and 3:00 AM, suggesting a shift in their capacity for emotional engagement during nighttime hours. In comparison, *dispatcherLLM*'s responses exhibited a more evenly distributed pattern of emotional positivity throughout the day. While *dispatcherLLM*'s emotional support remained consistent, the human dispatcher's ability to offer emotional assistance experienced a distinct temporal variation, as shown in Table 2. This result holds practical value as it suggests that *dispatcherLLM* potentially offers not only consistent availability but also maintains a stable emotional response, regardless of the time of service.

4.3.3 Users' Perceptions of the Responses Generated by LLMs. To further understand how users perceive the emotional support in the responses generated by *dispatcherLLM*, we conducted a survey study with human evaluators.

During the study, the participants were given randomly sampled messages of incident reporting and asked to comment on whether emotional support should be provided in the example contexts and provide an explanation of their choice; if a participant responded with yes, they were further asked which LLM's (*dispatcherLLM* or GPT-3.5) response could be used to improve the dispatcher's original reply, or if neither of them would be helpful. We also asked the participants to rate the appropriate level of emotional support. The major findings are highlighted below.

First, participants thought *dispatcherLLM* generated more human-like and *"direct"* responses while GPT-3.5's responses were too lengthy (P10). We conducted a t-test showing a significant difference in the chat length between generated responses from human dispatchers and GPT-3.5 ($t = [-4.478], p = 0.011^{*}$), while no significant difference in chat length was identified between human dispatchers and *dispatcherLLM*, which indicated a higher similarity in the linguistic style between human dispatcher and *dispatcherLLM* than GPT-3.5.

Second, participants agreed that in situations such as assault, abuse, and harassment, increased emotional support is essential. For example, Figure 5 showed a random example incident of *Harassment/Abuse*. When rating whether the response provided an appropriate level of emotional support, participants gave *dispatcherLLM* a significantly higher score over GPT-3.5 ($t = [2.135], p = 0.041^{*}$). They preferred the responses of *dispatcherLLM*, because they found it provided both emotional support and *"ensured that the appropriate actions are taken in a timely manner ..."* (P10). Several participants liked the conciseness of *dispatcherLLM*'s response: *"... it is concise and gives the appropriate amount of response with emotional support and call to action."* (P7).

Third, even when emotional support was not perceived as critical, participants still found the generated responses helpful in handling the incidents. For example, when reviewing a noise disturbance case, two-thirds of the participants thought "a noise complaint probably doesn't require extensive emotional aid" (P13). This comment aligns with the results shown in Table 1—where users expressed fewer negative emotions in their reports, and Table 2—where dispatchers provided minimal emotional support. The lack of users' need for emotional support offered a potential explanation for why there was no significant difference in participants' perceived level of support between *dispatcherLLM* ($M = 3.53, SD = 0.94$) and GPT-3.5 ($M = 3.35, SD = 1.22$). Nonetheless, more than half of the participants thought the *dispatcherLLM* responses would be helpful in handling the *Noise Disturbance* incidents.

Overall, participants' perspectives on how AI-provided emotional support varied across different types of incidents. More than 80% of the participants agreed that AI needs to provide support in certain circumstances, particularly those involving mental health and emotional distress, e.g., "anytime you sense distress in mental health scenarios" (P1) and "emergencies involving mental health or emotional crises need special care" (P3). In a case where a mental health report was sent, 88.2% participants preferred the generated response from GPT-3.5 ($t = [8.865], p < 0.001^{***}$). As one participant explained, the *"longer reply is showing that they truly do care and will do anything to support this person going through such a hard time"* (P1). Considering the limited emotional support provided by human dispatchers during *Mental Health* reporting, as shown in Table 2, we may further improve LLM's emotional support by augmenting the training data from domain experts such as mental health counseling to fine-tune the *dispatcherLLM* in future work, because the GPT-3.5 responses may lack professional expertise or not be sufficient to address the users' real need.

Summary (RQ3): The evaluation of *dispatcherLLM* showcased its feasibility for assisting in the handling of safety incident reports. First, the generated responses of *dispatcherLLM* were more similar to those of human dispatchers than those generated by the off-the-shelf model and GPT-3.5. The model was also observed to deliver greater emotional support than human dispatchers in two aspects: across various incident types and throughout different service hours. According to feedback from participants, they generally preferred *dispatcherLLM*'s responses for they are more human-like, providing both emotional support and relevant instructions. When emotional support is not deemed critical, participants still found the LLM-generated responses to be helpful in improving human dispatchers' incident report handling. It is important to note that their preferences for AI models also varied by incident type. In *Mental Health* scenarios, participants favored longer and more empathetic responses. Future work should comprehensively evaluate these models in diverse contexts to offer more in-depth guidance. Human dispatchers should also be cautious about the over-reliance on the LLM-generated responses.

5 Discussion

In this section, we discuss the findings of this study and their connections to prior work and implications for future designs in the field of community safety risk reporting systems.

5.1 Temporal Dynamics of Emotional Support in Text-Based Safety Reporting

Users' negative emotions during incident reporting (RQ1) Existing studies have pointed out that users can exhibit negative emotions during traditional emergency incident reporting and can cause a negative impact on the handling process if not appropriately addressed by dispatchers

User's First Message
My wife was reading at the Nobel library in a single room this afternoon. A guy stormed into her room and grabbed her hand. She was on the #nd floor. She is really scared and shocked by this incident. That guy was laughing while in her room. He has curly hair and was wearing a hoody. She couldn't give any other information regarding his appearance. Please kindly look into it. Thank you.

Human Dispatcher	GPT-3.5	*dispatcherLLM*
Thank you for contacting the [ACRONYM] Police. Do you know how long ago this happened?	I'm so sorry to hear about this terrifying experience. Thank you for reaching out to us. It's important that we take immediate action to ensure your wife's safety and apprehend the individual responsible. I want to make sure your wife is safe right now. Can you please confirm her current location and provide me with any other pertinent details? Additionally, I want you to know that you can stay on the line with me so that I can offer support and guidance as we address this situation together. Let's work together to keep your wife safe and ensure that this individual is held accountable for their actions.	Thank you for reporting this incident to us. I'm so sorry to hear that your wife had this frightening experience. I'm here to help and support you through this. Can you please tell me more about the location of the library? Is it a public library? And do you know the time of the incident?

Fig. 5. A randomly selected *Harassment/Abuse* incident was showcased in the user survey. The follow-up question generated by *dispatcherLLM* was perceived by participants to be more contextualized and specific than GPT-3.5's response.

[19, 76]. The negativity in users' emotions, per our findings of **RQ1**, was also found in the reporting scenarios when using text-based reporting systems. A gradual decrease in both users' negative emotions and neutral utterances was found during conversations with dispatchers. This finding implied that two different types of support were provided by dispatchers to users during incident reporting sessions — informational and emotional support, which has been discussed in prior research related to online chat-based healthcare consultation [64]. Within the context of text-based safety reporting, informational support refers to the process of both information collection and providing assistance, including dispatching officers, instructing, and providing additional information and resources. In terms of emotional support, the behavior of dispatchers accords with that of traditional voice-based incident reporting [19]. The effect of dispatchers' emotional support delivery can be reflected through the changes in the negative emotions of users as dispatchers attempt to calm users.

Our findings of **RQ1** analysis revealed an association of users' overall emotion with safety incidents' internal characteristics, i.e. tip categories. Incidents that are less severe or life-threatening in nature tend to trigger fewer negative emotions in users' responses. Another factor associated with users' emotional polarity is anonymity, as anonymous users tend to be more negatively polarized when making an incident report. This corresponds to findings of prior studies revealing anonymity, when offered as an option in online counseling systems [24], has the advantage of eliciting more truthful disclosure from users. However, no difference in users' emotional negativity was found across organizations with respect to the experience since the implementation of the safety reporting system and the number of tips received per year. This hints that users' initial emotion polarity only depends on the categories of incidents and users' choices in whether to report anonymously, as opposed to the characteristics of organizations. This indicates the consistency in levels of users' negativity to be expected by dispatchers regardless of their belonging organizations. On the other hand, higher learning institutions, which comprise the organizations where our analyzed data were collected, have a higher turnover rate of community members with the change caused by newly admitted students and graduates mostly in an annual cycle compared to other organizations and communities, which renders our findings to be more generalizable for communities with more frequent flows of members.

Organizations' temporal decay in emotional support delivery (RQ2) Although users exhibit consistent negative emotional displays across organizations, dispatchers were found to provide different levels of emotional support regarding how long organizations have implemented the system. Temporally, organizations were found to have an elevated level of emotional support delivery during the first year of reporting system implementation. The ratio of provided emotional support gradually decreased to a baseline as organizations became more experienced in working with the system. This finding corresponds to a similar effect as "initiative decay" discussed in Buchanan et al. [10]'s work. This revealed two insights about the organizational adaptation of safety reporting systems: 1) providing a high level of emotional support during the initial implementation of a safety reporting system may help to establish positive word-of-mouth and encourage ongoing use of the system; 2) as organizations become more experienced with using the system, the level of emotional support provided may decline, potentially due to a lack of awareness of the ongoing need for this type of support. As the incentive for providing high-quality service with emotional support is reduced after the initial implementation period, it may be important for organizations to regularly assess the level of emotional support being provided to users of the safety reporting system, and to actively make efforts to maintain or increase this support as needed. This could involve providing training or resources for dispatchers to offer emotional support effectively, or implementing systems to monitor and track the level of support being provided. Ensuring that users continue to receive adequate emotional support can help to improve the overall effectiveness of the safety reporting system and increase user satisfaction.

Our findings also identified a difference in the nature of incidents reported by users by different time periods of the day. Dispatchers during certain time periods might suffer more from burnouts than others, depending on the distribution of incidents reported by users. For instance, incidents related to *Noise Disturbance* and *Drugs/Alcohol* were found to have a significant increase in numbers during the night shift (8 p.m. to 12 a.m.), which are often deemed to be more repetitive in terms of the procedures required to handle them. Additionally, more incidents related to *Harassment/Abuse* were found to have been reported by community members during the night shift. This change in the composition of incident types caused an increase in both task volumes, repetitiveness and dispatchers' emotional cost required to handle the reports, leading to potentially higher emotional burnouts when compared to other shifts. This corresponds with our regression analysis results where even though the tip categories were controlled, there exists a persistent negative impact on emotional support delivery during the higher-volume shifts.

Improved emotional support from *dispatcherLLM* (RQ3) Prior studies have discussed how AI can be used to provide emotional support in various domains such as mental health counseling [16] and customer services [22, 29]. Our study made the first attempt in the incident reporting domain by exploring the feasibility of using LLMs to provide both informational and emotional support. The findings from RQ3 demonstrated that LLM-based chat models are perceived as useful for improving responses from human dispatchers. According to the RQ2 results in Table 2, *dispatcherLLM* may be applied to support human dispatchers across various times of the day when they have psychological burnout [67, 73]. The AI's support can significantly enhance the consistency of the responses. For many incident categories, our developed *dispatcherLLM* outperformed human dispatchers, except for *Mental Health.* This underperformance might be partially attributed to the relative scarcity of training data in these specific categories. It is also possible that the human dispatcher chat logs used to fine-tune the *dispatcherLLM* lack the necessary incident cases where deep expertise in emotional support is exemplified. Future studies should systematically evaluate LLMs' emotional support. Methods such as data augmentation [54, 83] and self-training [60] can be utilized to further incorporate expertise from domains such as mental health counseling into the model. Similarly, the AI's support can help human dispatchers reflect more critically when handling diverse incidents.

5.2 Design Implications

Improving report handling with optimized workflow and AI agents Our findings provided an example to support AI-enabled chatbots, as shown in Fig. 4, where LLMs can be leveraged to assist human dispatchers with providing emotional support during safety-related incidents. This support may be particularly useful when a wide range of reports are submitted by a large amount of community members, because dispatchers may save time from typing by adopting part of the generated responses. Also, we found that both the volumes and categories among incidents reported by users differed across different time periods of a day, potentially causing varied pressure and burnouts for dispatchers on different work shifts [76]. Mixed reporting handling regardless of their nature in complexity and need for emotional support could lead to a degradation in dispatchers' service level in terms of emotional support delivery, as shown in the findings of 4.2. For high-volume shifts where multiple dispatchers are present to provide support, dispatchers can be assigned dedicated categories of incidents based on the complexity and levels of emotional support needed. For example, playbooks and standard procedures for simpler and repetitive incidents can be created to increase handling efficiency [3] so that a higher volume of incidents can be handled by a single dispatcher with a minimal negative impact on emotional support delivery, while incidents needing more attention for emotional support can be allocated to dispatchers with lower workload to mitigate burnouts. Certain categories of incidents, such as *Noise Disturbance* and *Theft / Lost Item*, with relatively less complex and repetitive procedures, can also benefit from partial automation using LLM-based agents. This can include the application of automatic information extraction [51] incorporating methods described in 3.3.2 of this work, and contextualized dialogue response generation with emotional support similar to the model introduced in this study [44, 46, 79]. Designs should also consider assisting humans in providing emotional support without significantly increasing their cost of action, by offloading the handling of less complex incidents to LLMs. Methods such as automated response suggestions [17] and escalation protocols for human intervention [36] can be employed to facilitate a less intrusive human-AI collaboration process. Nonetheless, how organizations integrate the proposed approach into their existing workflow requires longitudinal studies for thorough evaluation.

Designing for anonymity in incident reporting In a safety incident reporting scenario, previous study [49] has suggested that the feature of allowing users to remain anonymous during reporting conversations plays a significant role in reducing users' concerns and helping users to provide more input to their submitted tips, which corresponds with the victimization theory [39] as anonymity reduces users' perceived cost of reporting. Similar to the utility-cost model of reporting [52], dispatchers' decision-making process in providing emotional support can be conceptualized as a trade-off between the utility (e.g., resolving an urgent safety issue) and cost (e.g., time spent and emotional labor invoked [25, 67]) of providing emotional support to users. Our finding implies the possibility that the display of users' anonymity status might potentially negatively impact dispatchers' evaluation of the validity of the reported incidents, which provides insights about improving the design of future reporting systems to better ensure the validity of reports and aligning users' and dispatchers' perception of anonymity. For example, safety administration might provide separate reporting channels for anonymous and non-anonymous reports with different levels of expected response time acknowledged by both stakeholders. Allowing asynchronous interaction between users and dispatchers leads to better communication and follow-up, compared to traditional communication channels (e.g., phone calls).

6 Limitations and Future Work

Although the findings in this study uncover unique insights about the role of emotions in safety reporting interaction between users and dispatchers, most analyses were conducted from a macro perspective. To further understand the effect of user emotion on dispatchers' decision-making process, more case-by-case qualitative analyses should be conducted to provide a deeper understanding of how the interactions took place. In the data pre-processing stage, we excluded any non-English tips. While this simplifies the analysis, it's possible that these tips could contain valuable information related to risk reporting. As a potential next step, analyses should be conducted from a multi-lingual perspective to examine the findings over speaking populations. During the evaluation of the *dispatcherLLM* generated responses, community members served as surrogate human evaluators instead of actual victims, which could limit the empirical value of the results. We are planning to run more comprehensive evaluation with professional dispatchers and actual victims by carefully addressing recruitment challenges and ethical concerns. The study also provides implications for mitigating dispatchers' burnout, which is unable to be observed directly from the dataset itself. Further study should be conducted to provide more solid evidence about this inference, such as conducting interviews or surveys of the dispatcher population. When fine-tuning the LLM, all sampled chat logs were utilized for training and evaluation. Future work should further explore improving the data quality by filtering the chat logs based on criteria such as quality of emotional support to improve the model responses.

7 Conclusion

In this study, we found that emotion plays a significant role in text-based reporting systems for community safety, with users' emotions varying significantly according to tip categories. Inconsistency in the emotional support provided by dispatchers was also identified, with organizations with more experience using the reporting system tending to provide less emotional support. We also piloted using LLM to facilitate safety incident report handling and identified its potential to provide better and more consistent emotional support. Our findings offer new empirical evidence for understanding the patterns of users' emotions and organizations' emotional support from different perspectives and provide insights for the design of future text-based risk reporting systems using AI techniques including LLM.

Acknowledgments

This research was conducted using OpenAI credits provided through the OpenAI Researcher Access Program, which provided access to the GPT-3.5 model. Additionally, results presented in this paper were obtained using computational resources provided through allocation CIS230059 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program [9], which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

[1] Syed Ishtiaque Ahmed, Steven J. Jackson, Nova Ahmed, Hasan Shahid Ferdous, Md. Rashidujjaman Rifat, A.S.M Rizvi, Shamir Ahmed, and Rifat Sabbir Mansur. 2014. Protibadi: A Platform for Fighting Sexual Harassment in Urban

Bangladesh. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. ACM, New York, NY, USA, 2695–2704. doi:10.1145/2556288.2557376

[2] Syed Ishtiaque Ahmed, Steven J Jackson, Nova Ahmed, Hasan Shahid Ferdous, Md Rashidujjaman Rifat, ASM Rizvi, Shamir Ahmed, and Rifat Sabbir Mansur. 2014. Protibadi: A platform for fighting sexual harassment in urban Bangladesh. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2695–2704.

[3] Zeynep Aksin, Mor Armony, and Vijay Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management* 16, 6 (2007), 665–688.

[4] Farhan Ali, Doris Choy, Shanti Divaharan, Hui Yong Tay, and Wenli Chen. 2023. Supporting self-directed learning and self-assessment using TeacherGAIA, a generative AI chatbot application: Learning approaches and prompt engineering. *Learning: Research and Practice* 9, 2 (2023), 135–147.

[5] Adriana Alvarado Garcia and Christopher A Le Dantec. 2018. Quotidian report: Grassroots data practices to address public safety. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–18.

[6] Henri Barki, Ryad Titah, and Céline Boffo. 2007. Information system use–related activity: an expanded behavioral conceptualization of individual-level information system use. *Information systems research* 18, 2 (2007), 173–192.

[7] Rena Bivens and Amy Adele Hasinoff. 2018. Rape: is there an app for that? An empirical analysis of the features of anti-rape apps. *Information, Communication & Society* 21, 8 (2018), 1050–1067. doi:10.1080/1369118X.2017.1309444 arXiv:https://doi.org/10.1080/1369118X.2017.1309444

[8] Jan Blom, Divya Viswanathan, Mirjana Spasojevic, Janet Go, Karthik Acharya, and Robert Ahonius. 2010. Fear and the city: role of mobile services in harnessing safety and security in urban use contexts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1841–1850.

[9] Timothy J Boerner, Stephen Deems, Thomas R Furlani, Shelley L Knuth, and John Towns. 2023. Access: Advancing innovation: Nsf's advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and Experience in Advanced Research Computing*. 173–176.

[10] David Buchanan, Louise Fitzgerald, Diane Ketley, Rose Gollop, Jane Louise Jones, Sharon Saint Lamont, Annette Neath, and Elaine Whitby. 2005. No going back: A review of the literature on sustaining organizational change. *International Journal of Management Reviews* 7, 3 (2005), 189–205.

[11] CatapultEMS. 2022. WeTip Anonymous Reporting System. https://www.wetip.com/. Online; accessed 06 July 2022.

[12] Neo Christopher Chung, George Dyer, and Lennart Brocki. 2023. Challenges of large language models for mental health counseling. *arXiv preprint arXiv:2311.13857* (2023).

[13] Irena Pletikosa Cvijikj, Cristina Kadar, Bogdan Ivan, and Yiea-Funk Te. 2015. Towards a crowdsourcing approach for crime prevention. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. 1367–1372.

[14] Karl Daher, Jacky Casas, Omar Abou Khaled, and Elena Mugellini. 2020. Empathic chatbot response for medical assistance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–3.

[15] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547* (2020).

[16] Kerstin Denecke, Sayan Vaaheesan, and Aaganya Arulnathan. 2020. A mental health chatbot for regulating emotions (SERMO)-concept and usability test. *IEEE Transactions on Emerging Topics in Computing* 9, 3 (2020), 1170–1182.

[17] Giulia Di Fede, Davide Rocchesso, Steven P Dow, and Salvatore Andolina. 2022. The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Proceedings of the 14th Conference on Creativity and Cognition*. 623–627.

[18] Federal Communications Commission. 2022. Text to 911: What You Need To Know. https://www.fcc.gov/consumers/guides/what-you-need-know-about-text-911. Online; accessed 11 July 2022.

[19] Heidi Kevoe Feldman. 2021. Calming emotional 911 callers: Using redirection as a patient-focused directive in emergency medical calls. *Language & Communication* 81 (2021), 81–92.

[20] Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021*. 807–819.

[21] Shuyang Gao, Sanchit Agarwal, Tagyoung Chung, Di Jin, and Dilek Hakkani-Tür. 2020. From Machine Reading Comprehension to Dialogue State Tracking: Bridging the Gap. *CoRR* abs/2004.05827 (2020). arXiv:2004.05827 https://arxiv.org/abs/2004.05827

[22] Katja Gelbrich, Julia Hagel, and Chiara Orsingher. 2021. Emotional support from a digital assistant in technology-mediated services: Effects on customer satisfaction and behavioral persistence. *International Journal of Research in Marketing* 38, 1 (2021), 176–193.

[23] Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. *Advances in Neural Information Processing Systems* 32 (2019).

[24] Annika Gieselmann and Reinhard Pietrowsky. 2016. Treating procrastination chat-based versus face-to-face: An RCT evaluating the role of self-disclosure and perceived counselor's characteristics. *Computers in Human Behavior* 54 (2016), 444–452.

[25] Sarah E Golding, Claire Horsfield, Annette Davies, Bernadette Egan, Martyn Jones, Mary Raleigh, Patricia Schofield, Allison Squires, Kath Start, Tom Quinn, et al. 2017. Exploring the psychological health of emergency dispatch centre operatives: a systematic review and narrative synthesis. *PeerJ* 5 (2017), e3735.

[26] Rob Grace and Sierra Sinor. 2021. How to text 911: A content analysis of text-to-911 public education information. In *The 39th ACM International Conference on Design of Communication*. 135–141.

[27] Judith A Hall and Marianne Schmid Mast. 2007. Sources of accuracy in the empathic accuracy paradigm. *Emotion* 7, 2 (2007), 438.

[28] Md Monsur Hossain, Moushumi Sharmin, and Shameem Ahmed. 2018. Bangladesh emergency services: a mobile application to provide 911-like service in Bangladesh. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. 1–11.

[29] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch your heart: A tone-aware chatbot for customer care on social media. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[30] Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656* (2023).

[31] Mary Ann E Ignaco. 2019. Development of mobile application for incident reporting. In *Proceedings of the 20th Annual SIG Conference on Information Technology Education*. 162–162.

[32] Alicia Iriberri, Gondy Leroy, and Nathan Garrett. 2006. Reporting on-campus crime online: User intention to use. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, Vol. 4. IEEE, 82a–82a.

[33] Mahmood Jasim, Pooya Khaloo, Somin Wadhwa, Amy X Zhang, Ali Sarvghad, and Narges Mahyar. 2021. CommunityClick: Capturing and reporting community feedback from town halls to improve inclusivity. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–32.

[34] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[35] P Kalyanchakravarthy, T Lakshmi, R Rupavathi, S Krishnadilip, and P Lakshmankumar. 2014. Android Based Safety Triggering Application. *International Journal Of Computer Science and Information Technologies* 5, 1 (2014).

[36] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence.. In *IJCAI*. 4070–4073.

[37] Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2018. Identifying emotional support in online health communities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[38] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2022. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society* (2022), 14614448221142007.

[39] Dan A Lewis. 2017. *Fear of crime: Incivility and the production of a social problem*. Routledge.

[40] Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760* (2023).

[41] Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. *arXiv preprint arXiv:2104.05919* (2021).

[42] Yanran Li, Ke Li, Hongke Ning, Xiaoqiang Xia, Yalong Guo, Chen Wei, Jianwei Cui, and Bin Wang. 2021. Towards an online empathetic chatbot with emotion causes. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2041–2045.

[43] Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7999–8009.

[44] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards Emotional Support Dialog Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 3469–3483.

[45] Yiren Liu, Ryan Mayfield, and Yun Huang. 2023. Discovering the Hidden Facts of User-Dispatcher Interactions via Text-based Reporting Systems for Community Safety. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–31.

[46] Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion* 64 (2020), 50–70.
[47] Antonette Mendoza, Jennie Carroll, Linda Stern, et al. 2010. Software appropriation over time: from adoption to stabilization and beyond. *Australasian Journal of Information Systems* 16, 2 (2010).
[48] Jingbo Meng, Minjin Rheu, Yue Zhang, Yue Dai, and Wei Peng. 2023. Mediated Social Support for Distress Reduction: AI Chatbots vs. Human. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–25.
[49] Shufan Ming, Ryan DW Mayfield, Haocong Cheng, Ke-Rou Wang, and Yun Huang. 2021. Examining interactions between community members and university safety organizations through community-sourced risk systems. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.
[50] Zou Minghui, Pan Rui, Zhang Sai, and Zhang Xiaowang. 2022. Using Extracted Emotion Cause to Improve Content-Relevance for Empathetic Conversation Generation. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*. 811–823.
[51] Rashid Minhas, Camilla Elphick, and Julia Shaw. 2022. Protecting victim and witness statement: examining the effectiveness of a chatbot that uses artificial intelligence and a cognitive interview. *AI & SOCIETY* 37, 1 (2022), 265–281.
[52] Samuel L Myers. 1980. Why are Crimes Underreported? What is the Crime Rate? Does it" Really" Matter? *Social Science Quarterly* 61, 1 (1980), 23–43.
[53] Otuu Obinna Ogbonnia, Joseph Henry Anajemba, Aguboshim Felix Chkwuma, and Deepak Sahoo. 2023. Trust-Based Classification in Community Policing: A Systematic Review. In *2023 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 1–8.
[54] Yanting Pan, Yixuan Tang, and Yuchen Niu. 2023. An Empathetic User-Centric Chatbot for Emotional Support. *arXiv preprint arXiv:2311.09271* (2023).
[55] Isabella Paoletti. 2012. Operators managing callers' sense of urgency in calls to the medical emergency number. *Pragmatics* 22, 4 (2012), 671–695.
[56] Steven Pinker and Paul Bloom. 1990. Natural language and natural selection. *Behavioral and brain sciences* 13, 4 (1990), 707–727.
[57] Brian A Reaves. 2015. *Campus law enforcement, 2011-12*. US Department of Justice, Office of Justice Programs, Bureau of Justice.
[58] Aderonke Busayo Sakpere, Anne VDM Kayem, and Thabo Ndlovu. 2015. A usable and secure crime reporting system for technology resource constrained context. In *2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops*. IEEE, 424–429.
[59] Klaus R Scherer, Rainer Banse, and Harald G Wallbott. 2001. Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-cultural psychology* 32, 1 (2001), 76–92.
[60] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*. 194–205.
[61] Nakatani Shuyo. 2010. Language Detection Library for Java. http://code.google.com/p/language-detection/
[62] Marit Sijbrandij, Rebecca Horn, Rebecca Esliker, Fiona O'may, Relinde Reiffers, Leontien Ruttenberg, Kimberly Stam, Joop de Jong, and Alastair Ager. 2020. The effect of psychological first aid training on knowledge and understanding about psychosocial support principles: a cluster-randomized controlled trial. *International journal of environmental research and public health* 17, 2 (2020), 484.
[63] Duško Sivčević, Ivan Košanin, Slobodan Nedeljković, Vojkan Nikolić, Kristijan Kuk, and Srđan Nogo. 2020. Possibilities of used intelligence based agents in instant messaging on e-government services. In *2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH)*. IEEE, 1–5.
[64] Hongying Tan and Mengling Yan. 2020. Physician-user interaction and users' perceived service quality: evidence from Chinese mobile healthcare consultation. *Information Technology & People* (2020).
[65] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
[66] Sarah J Tracy. 2002. When questioning turns to face threat: An interactional sensitivity in 911 call-taking. *Western Journal of Communication (includes Communication Reports)* 66, 2 (2002), 129–157.
[67] Sarah J Tracy and Karen Tracy. 1998. Emotion labor at 911: A case study and theoretical critique. (1998).
[68] Marcie J Tyre and Wanda J Orlikowski. 1994. Windows of opportunity: Temporal patterns of technological adaptation in organizations. *Organization science* 5, 1 (1994), 98–118.
[69] University of Illinois Chicago. 2022. UIC SAFE App Office of Preparedness and Response University of Illinois Chicago. https://ready.uic.edu/toolkit/uic-safe-app/. Online; accessed 06 July 2022.
[70] Alan Vaux, Sharon Riedel, and Doreen Stewart. 1987. Modes of social support: The social support behaviors (SS-B) scale. *American Journal of Community Psychology* 15, 2 (1987), 209–232.

[71] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia* 57 (2006), 45.

[72] Liuping Wang, Dakuo Wang, Feng Tian, Zhenhui Peng, Xiangmin Fan, Zhan Zhang, Mo Yu, Xiaojuan Ma, and Hongan Wang. 2021. Cass: Towards building a social-support chatbot for online health community. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–31.

[73] Phoenix Chi Wang. 2022. The 'desired state of mind': Emotional labor and the hidden cost of symbolic power in 911 emergency response in the US. *Ethnography* (2022), 14661381221145353.

[74] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology* 17 (2023), 18344909231213958.

[75] Keith Weber, Aaron Johnson, and Michael Corrigan. 2004. Communcating emotional support and its relationship to feelings of being understood, trust, and self-disclosure. *Communication research reports* 21, 3 (2004), 316–323.

[76] Jack Whalen and Don H. Zimmerman. 1998. 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* 61, 2 (1998), 141–159. http://www.jstor.org/stable/2787066

[77] Congying Xia, Chenwei Zhang, Tao Yang, Yaliang Li, Nan Du, Xian Wu, Wei Fan, Fenglong Ma, and Philip Yu. 2019. Multi-grained named entity recognition. *arXiv preprint arXiv:1906.08449* (2019).

[78] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 3506–3510.

[79] Xiaohan Xu, Xuying Meng, and Yequan Wang. 2022. PoKE: Prior Knowledge Enhanced Emotional Support Conversation with Latent Variable. *arXiv preprint arXiv:2210.12640* (2022).

[80] Manako Yabe. 2017. Cost-benefit evaluation: Students', faculty's, and staff's willingness to pay for a campus safety app. *Journal of Criminal Justice Education* 28, 2 (2017), 207–221.

[81] Ao Zhang, Shaojuan Wu, Xiaowang Zhang, Shizhan Chen, Yuchun Shu, and Zhiyong Feng. 2020. EmoEM: Emotional Expression in a Multi-turn Dialogue Model. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 496–501.

[82] Changli Zhang, Daniel Zeng, Jiexun Li, Fei-Yue Wang, and Wanli Zuo. 2009. Sentiment analysis of Chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology* 60, 12 (2009), 2474–2487.

[83] Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2022. AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:258588110

[84] Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*. 1552–1568.

[85] Qingxiao Zheng, Zhongwei Xu, Abhinav Choudhary, Yuting Chen, Yongming Li, and Yun Huang. 2023. Synergizing human-AI agency: a guide of 23 heuristics for service co-creation with LLM-based agents. *arXiv preprint arXiv:2310.15065* (2023).

[86] Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584* (2023).

A APPENDIX

A.1 Definitions of the Ontology

Definition of Event Entities		
Entity	**SubEntity**	**Definition**
Person	—	A single person or a group of people
Location	—	A location denoted as a point such as in a postal system or abstract coordinates
Weapon	—	The primary method/instrument used by the offender which causes harm
Time	—	The time an incident occurred
Object	—	An inanimate object involved in the incident
Description	Person-Age	The age of the person
	Person-Race	The racial description of the person
	Person-Appearance	The physical appearance of the person
	Person-Clothing	The clothing worn by the person
	Person-Sex	The person's biological sex description
	Person-Action	An action or activity carried out by the person
	Person-Name	The name of the person
	Person-Movement	The movement or change in location of the person
	Location-Description	The descriptive information about the location
Contact	PhoneNumber	A phone number
	Email	An email address

Definition of Event Arguments	
Argument	**Definition**
ATTACKER	The attacking/instigating agent
TARGET	The target of the offense (including unintended targets)
WEAPON	The weapon used in the offense
START-TIME	When the incident starts
END-TIME	When the incident ends
PLACE	Where the incident takes place
TARGET-OBJECT	The target object of the offense (e.g., vehicle, stolen items, etc.)

Definition of Dispatchers' Intents	
Intent	**Definition**
Thank	Expressing gratitude to the user
ConfirmSendOfficer	Confirming an officer has been dispatched
NotifyOthersInCharge	Notifying other responsible personels/parties about the incident
AskMeetOfficer	Asking the user to meet an officer offline
AskToCall	Asking the user to call an officer/the organization
AskToVisit	Asking the user to visit the organization's office
AskForDetail	Asking the user for additional detail related to the incident (annotated with the event argument the dispatcher asks about, i.e., ATTACKER, TARGET, WEAPON, START-TIME, END-TIME, PLACE, TARGET-OBJECT)

A.2 Examples of Dispatcher Responses Generated by the Chat Model

Fine Tuning	RAG
A chat between an individual reporting a safety concern to the local police department and a dispatcher from the police department. The dispatcher gives helpful and detailed guidance and instructions on how to proceed. The dispatcher is also supposed to give necessary emotional support to the user.	A chat between an individual reporting a safety concern to the local police department and a dispatcher from the police department. The dispatcher gives helpful and detailed guidance and instructions on how to proceed. The dispatcher is also supposed to give necessary emotional support to the user.
	- Scenario: A user is reporting neighbors in the UV Suites for smoking weed indoors. The user lives on the terrace floor, and the neighbors, located in the #### building, are possibly in rooms #### #### or ####. The user declines the offer to meet with officers. **Round 1:** **User: Hi. I live in the uv suites and I'm reporting my neighbors for smoking weed indoors. I live on the terrace floor where the trash chute is and the people smoking are all the way at [DATE]. The smell is strong and it's quite disturbing.** **Dispatcher: [ACRONYM]. Do you know the room number? Which building?** **Round 2** **User: The #### building. It's either rooms #### #### or ####** **Dispatcher: Thank you so very much, I will send officers out. Would you like to meet with them? It's not required but I did want to offer.** **Round 3** **User: No thanks**
- Scenario: A user reported a strong odor in a location, specifying it was in [GPE]. However, the user does not know the cause of the odor and expresses feeling very nervous about it. Round 3: User: Reeks of [GPE] in [GPE] Dispatcher: Do you have any descriptions on the cars? Round 2 User: I[ORG] not quite sure, I[ORG] sorry. I went to the bathroom and instantly smelled it in the hallway/bathroom. Dispatcher: Which floor?	- Scenario: A user reported a strong odor in a location, specifying it was in [GPE]. However, the user does not know the cause of the odor and expresses feeling very nervous about it. Round 1: User: Reeks of [GPE] in [GPE] Dispatcher: Hello, where at in [GPE] is the odor coming from? Round 2 User: I[ORG] not quite sure, I[ORG] sorry. I went to the bathroom and instantly smelled it in the hallway/bathroom. Dispatcher: Which floor?

Fig. 6. In the supervised fine-tuning, a prompted dialogue history along with the GPT-3.5 generated summary is aligned with the dispatcher's output. In evaluation, a similar document in the training set, based on the dialogue history along with its summary, is retrieved to promote a more human-like response from the dispatcher bot.

A.3 Survey Incident Case Samples

User's Initial Message
[ORG] is in the [ORG] area of [ORG], loud music and talking. Can you send officers to check out? Thanks

Human Dispatcher	**GPT-3.5**	***dispatcherLLM***
Thank you for contacting [ORG]. Do you know how many people?	Hello, thank you for reporting this concern. I will definitely send officers to check out the situation. Can you please provide me with more details about the exact location and any other relevant information? It's important to stay calm and ensure your own safety while we handle this. Thank you for bringing this to our attention. We will do our best to address the situation promptly.	Hello, thank you for reporting this safety concern to us. Can you please provide me with more details about the loud music and talking that you are hearing?

Fig. 7. A randomly selected *Noise Disturbance* incident sample.

User's Initial Message
My wife was reading at the Nobel library in a single room this afternoon. A guy stormed into her room and grabbed her hand. She was on the #nd floor. She is really scared and shocked by this incident. That guy was laughing while in her room. He has curly hair and was wearing a hoody. She couldn't give any other information regarding his appearance. Please kindly look into it. Thank you.

Human Dispatcher	**GPT-3.5**	***dispatcherLLM***
Thank you for contacting the [ACRONYM] Police. Do you know how long ago this happened?	I'm so sorry to hear about this terrifying experience. Thank you for reaching out to us. It's important that we take immediate action to ensure your wife's safety and apprehend the individual responsible. I want to make sure your wife is safe right now. Can you please confirm her current location and provide me with any other pertinent details? Additionally, I want you to know that you can stay on the line with me so that I can offer support and guidance as we address this situation together. Let's work together to keep your wife safe and ensure that this individual is held accountable for their actions.	Thank you for reporting this incident to us. I'm so sorry to hear that your wife had this frightening experience. I'm here to help and support you through this. Can you please tell me more about the location of the library? Is it a public library? And do you know the time of the incident?

Fig. 8. A randomly selected *Harassment / Abuse* incident sample.

User's Initial Messages
User: [DATE] wasn't a good day. I've dealt with Depression before for [DATE] and there used to be many thoughts of ending it. As well as an "almost" attempt. I was doing fine and better, recently..[DATE]. It all just came out of nowhere and I'm just in a state of mind I don't want to be in. The thoughts came back [DATE]. Dispatcher: Thank you for using LiveSafe. We would like to help. Where are you located or can you provide a phone number where we can have a counselor call you? Dispatcher: Are you able to give us a call at (###) ###-####? User: I would rather not talk over the phone if you don't mind User: I just finished my class at [ORG] library and was biking back home that's why didn't pick up. I apologize for that. Dispatcher: That's understandable. I'm having an officer come up to text with you. Dispatcher: Hello this is [PERSON]. [PERSON] User: Hello how are you User: Thank you Dispatcher: I am concerned and would like to help you Dispatcher: Are you willing to meetup? Dispatcher: I would like to provide you with resources to help User: I just got back home. And I would rather not in this weather. I've already hurt myself this morning and I wouldn't want my bike to slip and I fall hurt myself again

Human Dispatcher	**GPT-3.5**	***dispatcherLLM***
Okay no problem	That's completely understandable, [PERSON]. I'm here to listen and support you in any way that I can. It's ok if you need some time to think about what you want to say. Just know that you're not alone in this, and there are people who care about you and want to help. Let's take things one step at a time.	I understand and I'm here for you. I want you to know that you are not alone and you are doing great.

Fig. 9. A randomly selected *Mental Health* incident sample.

A.4 Survey Questions

For each incident case sample, we ask the following survey questions:

- Which model's suggestion do you think can be used by the human dispatcher to improve their response? (A, B, or Neither)
- If you select model A or B, please explain what part of the model's suggestion improves the human dispatcher's response.
- In the context of this (each) incident type, do you think the dispatcher should provide emotional support to the user?
- On a scale of 1 to 5, to what extent do you think Model A's response provides an appropriate level of emotional support?
- On a scale of 1 to 5, to what extent do you think Model B's response provides an appropriate level of emotional support?

After reviewing the incident samples, we ask each participant the following questions:

- How do you think AI handles incidents with different levels of urgency (e.g., noise disturbance vs. gunshot)?
- Do you think AI should provide more emotional support for certain types of incidents? If so, what are the types of incidents?

Received January 2024; revised July 2024; accepted October 2024