# Designing a Human-in-the-Loop AI System for Incident Consolidation and Severity-Aware Triage in Emergency Call Centers

**İsa Utku Dursunoğlu**
Department of Computer Science and Engineering, Sabancı University, Istanbul, Turkey
utku.dursunoglu@sabanciuniv.edu

**Meliha Zeynep Demirtaş**
Department of Industrial Engineering, Sabancı University, Istanbul, Turkey
zeynep.demirtas@sabanciuniv.edu

**Barbaros Yahya**
Department of Computer Science and Engineering, Sabancı University, Istanbul, Turkey
barbaros.yahya@sabanciuniv.edu

**Adahan Yiğitol**
Department of Computer Science and Engineering, Sabancı University, Istanbul, Turkey
adahan.yigitol@sabanciuniv.edu

**Selim Balcısoy**
Department of Computer Science and Engineering, Sabancı University, Istanbul, Turkey
selim.balcisoy@sabanciuniv.edu

**ABSTRACT**

Emergency call centers experience severe cognitive and operational overload during large-scale crisis events, driven by surges in call volume, redundant incident reports, and incomplete information. While artificial intelligence offers opportunities to support emergency response operations, fully automated decision-making remains inappropriate in high-stakes, time-critical contexts. This paper presents ongoing work on the design and evaluation of a human-in-the-loop AI system that supports emergency call operators through incident consolidation, severity-aware triage, and real-time geospatial situational awareness. The system integrates speech-to-text transcription, natural language processing, probabilistic severity modeling, and spatial-semantic clustering to assist operators in identifying, prioritizing, and contextualizing incoming emergency reports while preserving human oversight. We describe the system architecture, key design decisions, and an initial technical evaluation using a synthetic but operationally grounded emergency call dataset. Preliminary results demonstrate promising classification performance and calibrated confidence estimates under simulated surge conditions. Ongoing work focuses on comparative baselines, user-centered evaluation, and field-oriented validation.

**Keywords**

Human-in-the-Loop AI, Emergency Triage, Incident Consolidation, Crisis Informatics, Situational Awareness.

## INTRODUCTION

Emergency call centers play a critical role in crisis response, acting as the primary interface between affected populations and emergency services. During large-scale incidents such as earthquakes, floods, or industrial accidents, call centers face abrupt surges in call volume, high levels of redundancy as multiple callers report the same event, and substantial uncertainty in both location and severity of incidents. These conditions place

significant cognitive demands on call operators, who must rapidly triage information, maintain situational awareness, and coordinate downstream response actions under time pressure.

Prior research in crisis informatics and emergency management has highlighted the limits of purely manual call-taking and dispatch workflows under extreme load, as well as the risks associated with over-automation in safety-critical environments. In response, there is growing interest in human-centered decision support systems that augment, rather than replace, professional judgment. Such systems aim to reduce information overload, surface relevant patterns, and support sense-making while ensuring that accountability and control remain with human operators.

This paper reports on work in progress regarding the design and early evaluation of an AI-mediated decision support system for emergency call centers. The system was designed around three guiding principles: (1) human-in-the-loop operation, ensuring that AI outputs inform but do not dictate decisions; (2) incident-level consolidation, addressing redundancy by aggregating spatially, temporally, and semantically related calls; and (3) uncertainty-aware triage, providing calibrated severity estimates to support prioritization under incomplete information.

The contributions of this work are threefold. First, we present the architecture and design rationale of a modular, real-time system that integrates transcription, natural language understanding, probabilistic modeling, and geospatial visualization for crisis response contexts. Second, we report results from an initial technical evaluation using simulated surge scenarios to assess classification performance, confidence calibration, and system behavior under load. Third, we outline key challenges and directions for ongoing validation, including baseline comparisons and user-centered studies with emergency response professionals.

## PROBLEM CONTEXT AND RESEARCH OBJECTIVES

### Problem Context: Emergency Call Handling Under Crisis Conditions

Emergency call centers represent a critical information gateway during disasters. In the immediate aftermath of large-scale, multi-hazard events—such as earthquakes followed by fires, structural collapses, and medical emergencies—call centers are confronted with extreme surges in call volume within very short time frames. Operators must rapidly interpret caller narratives that are often incomplete, emotionally charged, and spatially ambiguous, while simultaneously prioritizing life-threatening situations and coordinating response resources (Svensson & Pesämaa,2018).

A persistent challenge in such contexts is that emergency calls are inherently unstructured. Callers frequently lack precise address information, use colloquial or landmark-based descriptions, and report events from partial or uncertain perspectives. Moreover, during mass emergencies, multiple callers often report the same underlying incident independently, resulting in substantial redundancy. Human operators must therefore perform several cognitively demanding tasks at once: extracting location information, assessing severity, identifying duplicates, and maintaining a coherent mental model of the evolving situation.

Existing emergency management systems provide limited support for these tasks. Most systems are designed to record calls, support manual classification, and forward information downstream, but they offer minimal assistance in interpreting ambiguous language, consolidating redundant reports, or contextualizing incidents across geographic and hazard dimensions. Under surge conditions, this leads to increased cognitive load, slower decision-making, and a higher risk of misprioritization or resource misallocation. These limitations have been widely recognized in crisis informatics research as fundamental bottlenecks in disaster response operations (Palen & Anderson, 2016; Kaufhold et al., 2020).

While artificial intelligence has been proposed as a means to improve disaster response, many AI-driven approaches focus either on post-hoc analysis (e.g., damage assessment from imagery or social media) or on narrowly scoped automation tasks. In emergency call center environments, fully automating decision-making processes is neither feasible nor ethically acceptable. Therefore, this research investigates methods to transform large volumes of unstructured caller data into actionable situational awareness, supporting operators during multi-hazard disasters.

### Research Objectives

This research aims to design, implement, and validate an AI-driven decision-support system to assist emergency operators during multi-hazard crises. The system is designed explicitly to reduce cognitive burden, improve

information organization, and enhance situational awareness instead of making dispatching decisions automatically. This research aims to solve the identified challenges with the following methodology:

Objective 1: Support Interpretation of Ambiguous Location Descriptions. The descriptions of the locations provided by callers during the incident are often limited, partial, informal, or based on landmark references. The initial objective is to help operators by extracting and resolving approximate geographic locations from unstructured caller narratives, which allows an incident to be spatially contextualized despite the lack of direct address data.

Objective 2: Assist Severity-Aware Triage Across Multiple Hazard Types. Ranging from life-threatening events to informative assessments, emergency reports can fall into a wide variety of urgent emergencies. The second aim is to help operators differentiate among different levels of urgency relative to various hazard types, facilitating prioritization while preserving conservative, interpretable decision logic appropriate for life-critical situations. This type of urgency-level scaling will occur during the triage layer (Lanka, 2025).

Objective 3: Reduce Redundancy Through Incident Consolidation. In disasters, different calls often reference the same event. The third objective is to detect and aggregate these duplicate or extremely similar reports that could result in increased information overload and support operators in maintaining a better operational picture while also not attempting to hide potentially critical new information.
Objective 4: Provide Real-Time, Geospatially Grounded Situational Awareness. Effective disaster response involves a shared knowledge of where incidents are happening, how severely they are occurring, and how they relate spatially. The fourth objective is to visualize aggregated incident information in real time within a geospatial context to support operator sense-making, as situations may change rapidly.

Objective 5: Evaluate System Behavior Under Surge Conditions. As opposed to serving merely as a requirement for validating the preceding objectives, this last objective is an operational objective in its own right. Its focus will be on the systematic assessment of the system's ability to retain responsiveness, stability, and efficiency in addressing information overload during periods of simulated surges in calls.

## RELATED WORK

### AI Support for Emergency Call Handling and Triage

Researchers have created machine learning and Natural Language Processing (NLP) algorithms that help extract structured information for emergency calls and aid in prioritization of responses. Early applications involved monitoring 911 calls to facilitate the surveillance of public health through automated systems (Haas et al., 2011). Costa et al. (2023) investigate the use of AI in the transcription of emergency calls and classification of calls into defined categories, as well as and the challenges of deploying AI solutions to practice in real-life situations. Attiah and Kalkatawi (2025) propose a two-phase emergency call assistant pipeline that converts audio into text and then implements machine learning-based categorization to help the case taker prioritize emergencies. In the field of public safety, Atherley (2024) describes an emerging proof-of-concept "intelligent caller assist" project that is in development to improve the processing of emergency calls from police through NLP-based triaging of 911 calls at the 911 processing station. Qualitative studies on telephone-based decision-support systems indicate that emergency call handling is a complex socio-technical process, supporting the use of AI systems that assist rather than replace human judgment (Pope et al., 2017).

### Speech Recognition as the Front-End of Emergency NLP

Due to emergency information being communicated via voice under stress, several systems use automatic speech recognition (ASR) to facilitate the downstream application of NLP. Whisper is a popular multilingual ASR model, trained on large-scale web data, that is robust to transcription across languages and noise conditions (Radford et al., 2022). Nevertheless, occasional "hallucinated" insertions found in sensitive settings confirm the importance of conservative interfaces, confidence signaling, and human verification when using ASR output in high-stakes decision workflows.

### Location Extraction from Unstructured Crisis Data

Middleton et al. (2018) developed a comprehensive framework for addressing the core issues of geoparsing. Early applications used traditional statistical methods augmented by external knowledge bases (Al-Olimat et al., 2018). New approaches using Large Language Models (LLMs) and Long Short-Term Memory networks were introduced

by Eligüzel et al. (2022). The advancements of LLMs also brought new methods for handling location extraction, such as fine-tuning the T5 model (Dahlan & Yuangyai, 2024). Nevertheless, all these studies derive their data source from social media platforms, not actual emergency calls. Deciphering the emergency calls present a set of challenges: transcripts that come from automatic speech recognition can have poor punctuation, phonetic errors, and reflect the repetitive nature of spoken speech caused by panic. This work addresses this limitation by adapting BERT-based location extraction specifically for the noisy, unpunctuated, and informal linguistic patterns found in transcribed emergency calls.

### Automated Severity Scoring and Multi-Modal Fusion

In the clinical context, Williams et al. (2024) and Seo et al. (2025) have shown that although LLMs are able to make near human-level sensitivity on triage priority assessment, when the environmental context is missing they tend to show bias towards over-triage and assume the worst-case scenario. Hughes and Clark (2025) utilized visual-semantic models to effectively filter the actionable disaster content, whereas Hanny et al. (2025) introduced a "GeoAI" framework that weights the social media content based on spatiotemporal features.
However, these methodologies rely on geometric closeness rather than the estimated intensity of physical hazard to assess urgency level. Our work focuses on Disaster and Emergency Management Presidency's (AFAD's) seismic risk parameters to provide a better estimate of the situation.

### Physics-Informed Triage and Systemic Robustness

Ma et al. (2025) proposed using seismic data to evaluate the earthquake damage reports and applied a post-hoc validation of AI models using this data. Our methodology utilizes the PGA values from AFAD as a real-time deterministic constraint while making inference, advancing the foregoing work. Hong et al. (2025) employ dynamic fusion of multiple model outputs for consistency. Our orchestrator LLM acts as a fusion engine on context to find the most suitable action by combining risk scores, geolocation, and user conversation details.

## METHODOLOGY

### Methodological Approach

In this section, we provide details on the dataset and break down how modules worked and interacted. Rather than optimizing individual algorithms in isolation, the focus was on how computational components interacted with human operators to support decision-making under conditions of uncertainty, time pressure, and information overload. The system was developed iteratively to address four recurring challenges observed in emergency call center operations during disasters: (1) Interpretation of ambiguous, unstructured caller narratives; (2) Differentiation of urgency across heterogeneous hazard types; (3) Management of redundant or overlapping reports; and (4) Maintenance of coherent situational awareness during surge conditions.

To address these challenges, the system integrated multiple analytical modules into a single operational pipeline. Each module supported a distinct cognitive function commonly performed by human operators, allowing computational support to align with existing workflows rather than replace them. The architecture emphasized transparency, modularity, and conservative decision logic to ensure that outputs remained interpretable for human oversight. Other design choices for the design can be an automated and autonomous dispatch system that utilizes an LLM. While the automated approach would most probably have the least latency in terms of theory, prior research indicates that this type of system is prone to over-triaging biases and lacks ethical responsibility in situations involving the preservation of lives. Thus, we have chosen our hybrid design approach.

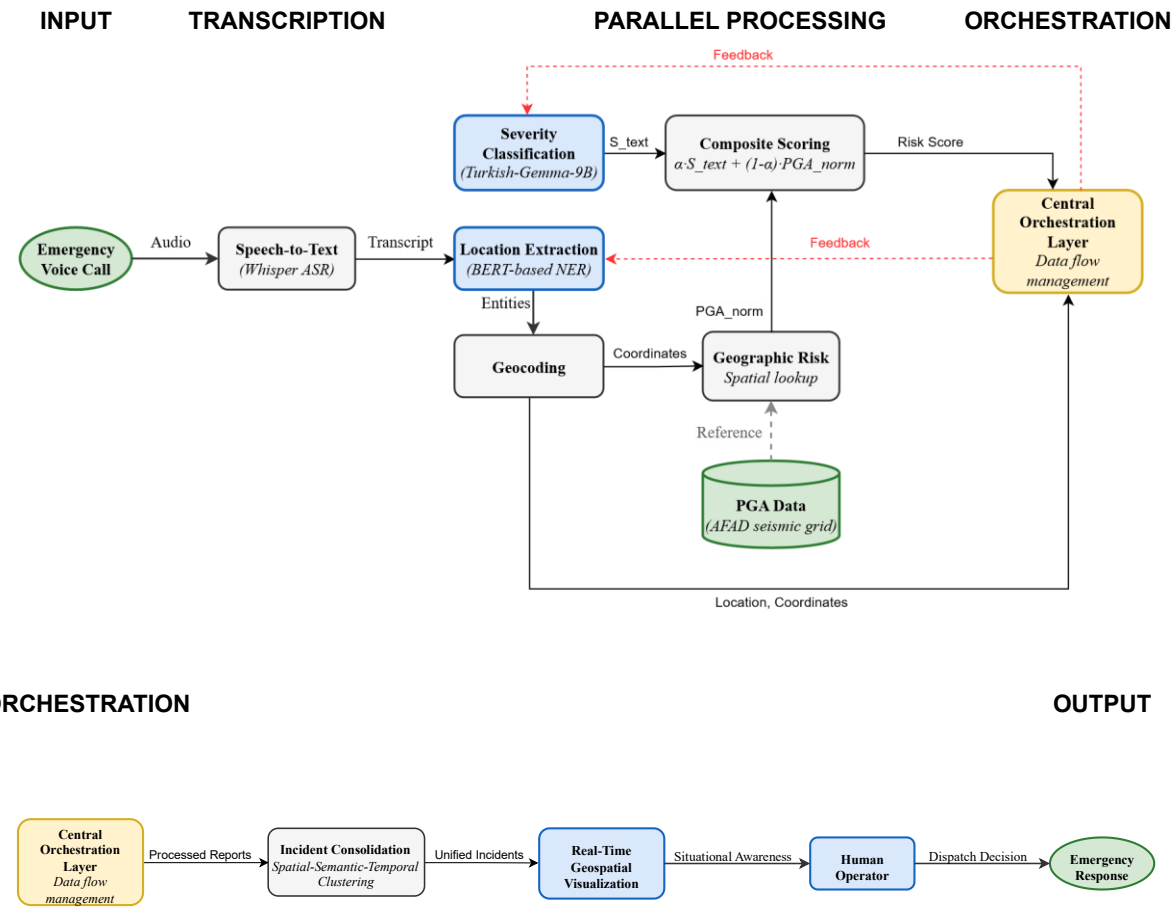


**Figure 1. System Architecture Overview. The pipeline processes inputs (top) through the central orchestration layer to generate actionable outputs (bottom).**

### Data Generation and Privacy Protection

Due to privacy concerns regarding real emergency calls, we employed robust data synthesis. A dataset of 2,421 examples, modeled after actual transcripts, was created to simulate operational conditions. To evaluate robustness in noisy environments, the dataset includes variations such as panic-induced speech breaks and incomplete sentences. Importantly, all processing was done based on privacy-preserving principles, without retaining any personally identifiable information (PII) in working memory.

### Interpretation of Caller Narratives and Location Inference

Realizing that emergency alerts are typically received over voice calls, a multi-modal ingestion layer has been integrated into this system. The audio inputs are processed with a Whisper model from OpenAI for converting audio inputs into text inputs within a near-real-time framework so that they can be fed into the processing pipeline (Radford et al., 2022). The integration of this module is beneficial for a seamless workflow from a voice description of an emergency provided by a sensor node to a structured piece of intelligence. Recent research about using ASR systems for speech recognition in emergency services highlighted that Whisper architectures offer superior robustness in handling the high background noise, emotional speech patterns, and diverse accents characteristic of distress calls compared to traditional ASR systems (Minulescu & Toma, 2025).

Emergency callers rarely provide precise or standardized location information, particularly during crises marked by stress. Instead, descriptions often rely on landmarks, neighborhoods, or relational phrases (e.g., "near the metro," "behind my building"). As noted in the recent research on emergency log analysis, such unstructured narratives pose significant challenges for generic NLP models, necessitating specialized pipelines to resolve spatial ambiguities (Thuestad & Grutle, 2023). To support operators in spatially contextualizing such reports, the system performed a structured interpretation of caller narratives to infer approximate geographic locations using a BERT-based Named Entity Recognition (NER) pipeline fine-tuned for the Turkish language to extract Location

(LOC) and Organization (ORG) entities (Devlin et al., 2019; Eligüzel et al., 2022). This design choice is supported by the research that domain-specific fine-tuning significantly improved entity extraction performance in morphologically rich languages like Turkish during disaster contexts (Eligüzel et al., 2022).

This process involved identifying place-related references within caller text and resolving them through a combination of language-based interpretation and rule-guided geographic querying via the OpenStreetMap (OSM) Nominatim API.[1] The methodology prioritized robustness over precision: when exact coordinates could not be reliably inferred, the system favored district-level or neighborhood-level localization rather than producing potentially misleading point estimates. A hierarchical fallback resolution strategy was employed, progressively relaxing constraints when initial interpretations failed, thereby maintaining operational usefulness even under high ambiguity.

To ensure independent and reliable operation, the system reordered extracted entities—placing building names before district names—to optimize the search process. In cases where the high-level AI dependencies did not work, the system activated the rule-based cleaner to eliminate the conversation "noise" words, including "lütfen" and "nerede." Moreover, all geographical searches were performed under the constraints of a 10-second time-out and were bounded within the confines of a view box to avoid processing delays during mass disasters. All derived locations were accompanied by confidence measures and the traces of the derived resolution to give the operators an insight into the derivation of the location estimate at any particular time. This design aligns with crisis informatics principles, emphasizing transparency and explainability in decision-support systems.

### Severity-Aware Triage Across Multiple Hazard Types

Emergency calls during disasters span a wide range of urgency levels and hazard categories, including medical emergencies, structural damage, fires, and informational reports. Supporting triage in this context requires sensitivity to linguistic cues, hazard semantics, and situational context. The system assisted triage by analyzing caller narratives to identify indicators of urgency and categorizing reports into three broad priority levels: immediate response required, delayed or secondary response, and informational support. Rather than producing binary decisions, the system generated structured assessments that included confidence levels and extracted indicators (e.g., references to entrapment, loss of consciousness, or fire spread).

To quantify urgency, the system computed a text-based severity score by combining classification confidence with predefined base severity weights assigned to each category. Table 1 presents these severity mappings, which reflect operational priorities in emergency response contexts.

**Table 1. Base Severity Weights for Priority Classes**

| Priority Class | Base Severity | Operational Interpretation |
|---|---|---|
| URGENT_RESPONSE | 1.0 | Immediate life-threatening situations |
| SECONDARY_RESPONSE | 0.6 | Delayed response; non-critical injuries |
| SUPPORT_INFO | 0.2 | Informational requests; no immediate danger |

The text-based severity score was computed as the product of classification confidence and the corresponding base severity weight, yielding a normalized indicator suitable for integration with environmental risk factors. To contextualize triage decisions spatially, the methodology incorporated region-specific hazard information, allowing severity assessments to reflect both reported conditions and environmental risk factors. Importantly, environmental context modified but did not override—clear linguistic indicators of immediate danger. This conservative weighting strategy ensured that computational assessments remained aligned with ethical requirements for human oversight in emergency response.

### Geographic Risk Integration and Composite Scoring

The system incorporated Peak Ground Acceleration (PGA) data from AFAD to include seismic hazard context in prioritization decisions. The 14,036 grid points in the PGA dataset, covering seismically active areas, had values that indicated expected ground acceleration under a 10% chance of exceedance in 50 years. Using caller-reported or inferred coordinates, nearest-neighbor spatial lookup was used to calculate geographic risk. PGA values were

[1] https://nominatim.org/

classified based on the severity levels shown in Table 2 after being normalized against a maximum threshold ($PGA_{max} = 0.8g$).

**Table 2. PGA-Based Geographic Risk Categories**

| PGA Range (g) | Risk Category | Interpretation |
|---|---|---|
| ≥ 0.6 | VERY_HIGH | Severe structural damage expected |
| 0.4−0.6 | HIGH | Significant damage likely |
| 0.2−0.4 | MODERATE | Moderate damage possible |
| < 0.2 | LOW | Minor damage expected |

A risk formula was used to combine text-based severity with geographic hazard context. The linear formulation weighted both components explicitly:

$$Risklinear = \alpha \times Stext + (1-\alpha) \times PGAnorm \quad (1)$$

where $S_{text}$ represents the text-based severity score, $PGA_{norm}$ stands for the normalized ground acceleration, and $\alpha = 0.7$ prioritizes caller-reported conditions over environmental context. The reason for choosing this particular value of $\alpha = 0.7$ is that we purposely tried to stress the importance of the linguistic severity more than normalizing the geographic risks. In such a way, we will avoid situations when clear signals of approaching mortality threat are undervalued because of relatively safe surroundings.

### Incident Consolidation and Redundancy Management

A defining characteristic of large-scale emergencies was the prevalence of redundant reporting: multiple callers independently described the same incident from different perspectives. While such redundancy could provide corroboration, it also imposed a significant cognitive burden on operators who had to recognize overlaps and mentally consolidate information. To address this challenge, the system implemented an incident consolidation mechanism that identified reports likely referring to the same underlying event. Consolidation was based on spatial proximity, semantic similarity of reported conditions, and temporal alignment. When similarity exceeded defined thresholds, reports were grouped under a single evolving incident representation rather than presented as separate entries. Crucially, consolidation did not discard information. New reports contributed additional detail to existing incidents, updating severity indicators or contextual notes as appropriate. This approach preserved informational richness while reducing clutter, supporting operator sense-making during high-volume situations.

### Orchestration and Real-Time Processing Under Surge Conditions

Disaster scenarios are characterized by concurrency. The system is therefore designed to process multiple reports in parallel, coordinating interpretation, triage, consolidation, and visualization without blocking or queue-induced delays. A central orchestration layer managed data flow between modules, ensuring that partial outputs (e.g., unresolved locations) did not stall downstream processes. This layer also supported simulation of surge conditions by injecting concurrent reports at configurable rates, enabling stress testing of system behavior under realistic disaster loads. By maintaining modular separation between analytical functions, the system supported graceful degradation. If one component experienced delay or uncertainty, other components continued operating, reflecting the resilience requirements emphasized in crisis response research.

### Real-Time Visualization and Operator Interaction

To support situational awareness, processed incidents were displayed on a real-time geospatial interface. The interface displayed consolidated incidents, priority levels, temporal progression, and contextual indicators in a manner designed to support rapid operator comprehension. Visualization choices emphasized clarity over complexity: color-coded severity levels, minimal text summaries, and spatial clustering allowed operators to identify emerging hotspots and critical incidents at a glance. Detailed information remained accessible on demand, enabling deeper inspection without overwhelming the primary operational view. The visualization was updated continuously as new reports arrived or existing incidents evolved, supporting dynamic sense-making on rapidly changing environments.

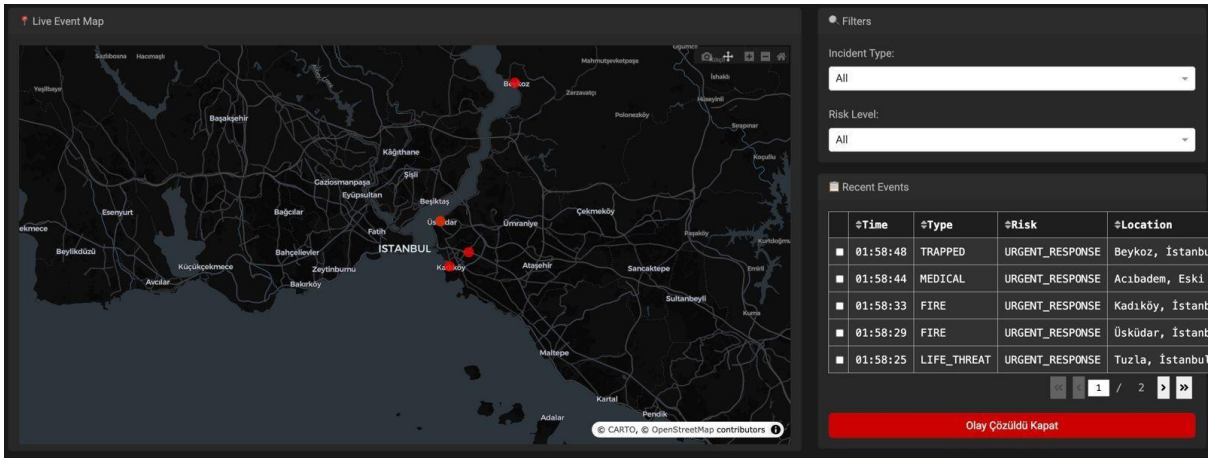


**Figure 2. Real-Time Geospatial Visualization Interface**

## EVALUATION STRATEGY

The evaluation methodology focused on operational behavior rather than algorithmic optimization alone. System performance was assessed through functional validation, simulated surge scenarios, and stress testing under concurrent report loads. Key evaluation dimensions included responsiveness, stability, effectiveness of incident consolidation, and clarity of information presentation. Rather than positioning accuracy metrics as endpoints, the evaluation examined whether the system supports improved organization of emergency information during crisis conditions. This aligns with ISCRAM's emphasis on practical utility, human-centered design, and resilience in real-world emergency response settings.

## RESULTS AND DISCUSSION

### Experimental Setup

The system was evaluated through a combination of functional validation and simulated surge scenarios designed to emulate the information dynamics of large-scale disasters. Table 3 summarizes the experimental configuration.

**Table 3. Experimental Configuration**

| Component | Specification |
|---|---|
| Base Model | Turkish-Gemma-9B (ytu-ce-cosmos) |
| Fine-tuning Method | QLoRA (4-bit NF4 quantization) |
| LoRA Configuration | r=16, $\alpha$=32, dropout=0.1 |
| Training Data | 2,421 annotated emergency calls |
| Validation / Test Split | 303 / 303 (stratified 80/10/10) |
| Optimizer | AdamW (LR=$5\times10^{-5}$, weight decay=0.05) |
| Training Epochs | 3 |
| Class Weighting | Balanced (sklearn compute_class_weight) |
| Hardware | NVIDIA Tesla T4 (16GB) |

This section presents the results of system evaluation and discusses their implications from a crisis informatics perspective. Rather than focusing solely on algorithmic performance, the evaluation examines how the system behaved under conditions representative of emergency response operations, particularly high call volume, ambiguity, and redundancy. The results are interpreted in terms of their contribution to operator workload reduction, situational awareness, and operational resilience during multi-hazard crises.

### System Behavior Under Operational Conditions

Evaluation scenarios included simultaneous reports with varying urgency, missing locations, and redundant calls referring to the same incident. Across these conditions, the system demonstrated stable real-time behavior. Incoming reports were processed without blocking, and intermediate uncertainties (e.g., unresolved locations or ambiguous severity cues) did not prevent other components from operating. This resilience is particularly important in crisis settings, where partial or degraded information is the norm rather than the exception. From an operational standpoint, the system's ability to continue functioning despite ambiguity supports continuous situational awareness rather than forcing operators to wait for "clean" data before acting.

Simulated surge conditions were used to assess system responsiveness and stability when processing many concurrent reports. Under these conditions, the system maintained near real-time updates to the operational display, with no observed cascading delays or system-level bottlenecks. The orchestration design allowed different analytical functions to proceed independently, enabling partial results to be surfaced even when some components experienced uncertainty. This behavior reflected principles of graceful degradation, which are critical in crisis response systems where complete information is rarely available. According to the results from the "stress test," the system could support continuous situational awareness during periods of intense information influx, a core requirement for effective emergency management.

### Interpretation of Ambiguous Location Information

A key challenge in emergency call handling is the interpretation of vague or informal location descriptions. Evaluation results indicate that the system was generally able to spatially contextualize reports even when callers provided incomplete or landmark-based references. In cases where precise geolocation was not possible, district-level or neighborhood-level localization was achieved, allowing incidents to be placed meaningfully within the operational map. Importantly, the system avoided presenting overconfident or misleading point-level locations when uncertainty was high. Instead, it surfaced approximate spatial contexts accompanied by indicators of confidence. This behavior aligns with crisis informatics recommendations that emphasize the value of usable location information over illusory precision. For operators, having a reliable approximate location is often more actionable than an exact but potentially incorrect coordinate.

### Support for Severity-Aware Triage

The system's triage support was evaluated in terms of its ability to distinguish between broadly different urgency levels across heterogeneous hazard types. Results showed that reports describing immediate life-threatening conditions were consistently identified and surfaced as high priority, while informational or low-urgency reports were separated accordingly. Table 4 presents per-class classification performance on the held-out test set (n=303). The fine-tuned model achieved a Macro F1 score of 0.8205 with a 95% bootstrap confidence interval of [0.7737, 0.8623].

**Table 4. Classification Performance on Test Set**

| Priority Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| URGENT_RESPONSE | 0.8846 | 0.8846 | 0.8846 | 130 |
| SECONDARY_RESPONSE | 0.5938 | 0.7308 | 0.6552 | 52 |
| SUPPORT_INFO | 0.9725 | 0.8760 | 0.9217 | 121 |
| Macro Average | 0.8169 | 0.8305 | 0.8205 | 303 |

The SECONDARY_RESPONSE class exhibited lower precision (0.5938), attributable to class imbalance (17.2% of training data) and the inherent ambiguity of intermediate-severity reports. However, the elevated recall (0.7308) indicated that the system erred toward capturing potential emergencies rather than dismissing them, a desirable property in crisis contexts. A notable outcome was the system's handling of intermediate cases, reports that did not clearly indicate imminent danger but still required attention. These were surfaced in a manner that preserved their visibility without competing directly with critical incidents. From an operator perspective, this differentiation supported more nuanced prioritization and reduced the risk that less urgent reports obscure critical ones during surge conditions.

Prioritization was further aided by the incorporation of environmental hazard context without overpowering obvious linguistic cues of urgency. This cautious relationship between caller narratives and contextual data was a deliberate design decision in line with operational and ethical emergency response requirements. Expected Calibration Error (ECE) was used to evaluate the model's calibration, and the result was 0.1048, indicating a moderate calibration appropriate for decision-support applications where outputs supplement human judgment rather than replace it. The key performance metrics are compiled in Table 5.

**Table 5. Summary of Model Performance**

| Metric | Value |
|---|---|
| Macro F1 | 0.8205 |
| Weighted F1 | 0.8601 |
| Accuracy | 0.8548 |
| Mean Confidence | 0.9596 |
| ECE (Calibration) | 0.1048 |
| 95% CI (Macro F1) | [0.7737, 0.8623] |

### Incident Consolidation and Redundancy Reduction

One of the most significant operational effects observed during evaluation was the reduction of redundant information presented to operators. In the simulation of disaster scenarios, multiple reports frequently referred to the same incident from different callers. The system's incident consolidation mechanism successfully grouped such reports into unified incident representations, updating them as new information became available. This consolidation reduced the number of discrete items requiring operator attention, helping to maintain a clearer operational picture. Rather than scanning long lists of similar reports, operators could focus on evolving incidents with aggregated context. This "aggregated context" approach aligned with digital health optimization strategies which also use textual feature extraction to streamline emergency service support (Attiah & Kalkatawi, 2025). Importantly, consolidation did not suppress new information; additional reports enriched existing incidents by providing supplementary details specific to the incident rather than being discarded. From a crisis informatics perspective, this functionality directly addressed a well-documented source of cognitive overload in emergency response: the need to manually recognize and correlate overlapping information streams. The results suggested that automated consolidation could meaningfully support operator sensemaking during high-volume events.

### Real-Time Visualization and Operator Sensemaking

The real-time visualization interface played a central role in translating processed information into actionable situational awareness. Evaluation showed that consolidated incidents, priority indicators, and spatial distribution enabled rapid identification of critical developments. By emphasizing clarity and minimalism in the primary view, the dashboard supported quick scanning and prioritization. More detailed information remained accessible on demand, allowing operators to investigate specific incidents without overwhelming the main display. The dynamic updating of incidents as new reports supported an evolving understanding of the situation, rather than a static snapshot of a single incident, which was particularly important in fast-moving disaster contexts. This dynamic movement of incidents could be observed through the map on the main display of the dashboard, which made it easier to see the overall view shown in Figure 2 and focus on a single incident by selecting it, as shown in Figure 3.

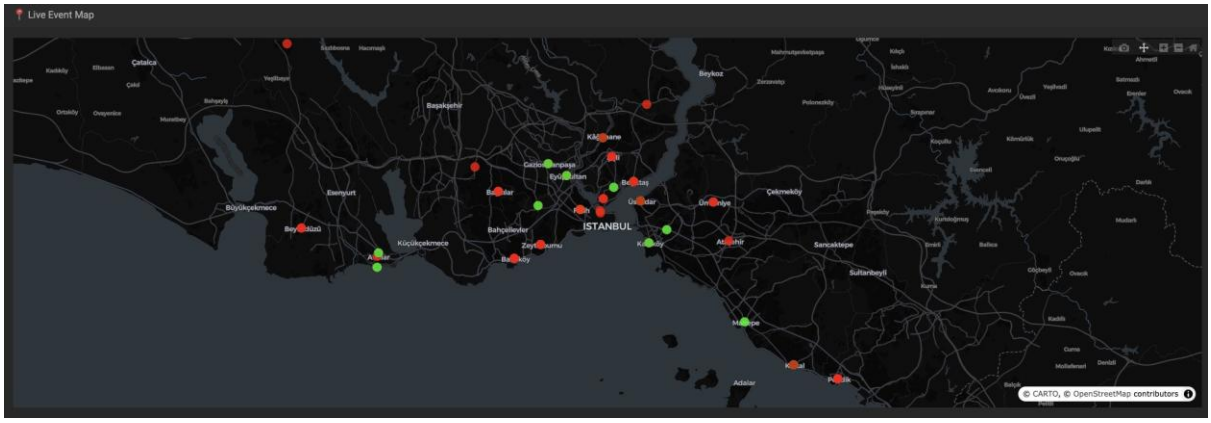


**Figure 3. Single Incident Selection View**

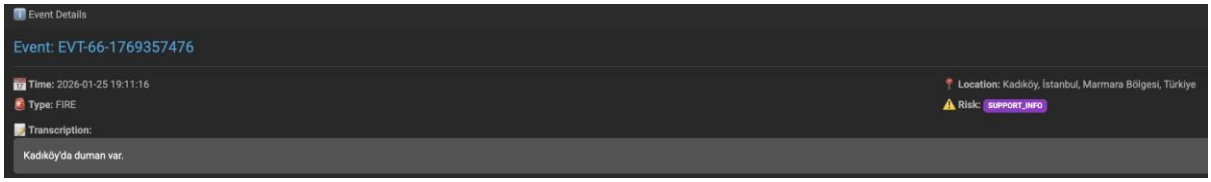


**Figure 4. Incident Detail View**

### Discussion: Implications for Crisis Informatics

Taken together, the results suggest that the primary value of the system lies not in isolated technical accuracy, but in its ability to restructure information flows during emergency response. By assisting with interpretation, triage, and consolidation, the system helps transform fragmented and redundant reports into a more coherent and understandable operational picture. Several implications emerge for crisis informatics research:

- Information organization is as critical as information accuracy. Reducing redundancy and improving coherence have a crucial effect on operator effectiveness during crises.
- Human-in-the-loop design remains essential. The system's conservative, explainable outputs support human judgment rather than replacing it, aligning with ethical and operational realities of emergency response.
- Resilience to ambiguity and overload is a key evaluation criterion. Systems that continue functioning under uncertainty and surge conditions provide greater practical value than those optimized for ideal inputs.

## ETHICAL CONSIDERATIONS AND LIMITATIONS

Strict adherence to ethical safeguards is required when deploying AI in life-critical domains. AI is only used as a decision-support tool in this system's "human-in-the-loop" design, not as an independent agent (Domfeh & Dancy, 2025). To mitigate the risk of algorithmic bias against vulnerable populations—such as callers with heavy accents or those unable to provide precise addresses—the system employs cascading fallback heuristics that prioritize broad location safety over precise but potentially erroneous pinning. Environmental risk factors can raise an incident's priority, but they are structurally prohibited from superseding unambiguous textual evidence of a life-threatening emergency due to the classification model's conservative scoring logic. This guarantees that the system will continue to be transparent, understandable, and compliant with the humanitarian requirement to do no harm.

While the system's architecture incorporates strict ethical safeguards, the current empirical validation has significant limitations that constrain the strength of our early conclusions. First, the evaluation relies entirely on a synthetic dataset of 2,421 annotated emergency calls. Although rigorously designed to simulate operational surge conditions, synthetic data cannot fully capture the linguistic variability, stress-induced phonetic errors, and complex noise profiles of live call center environments. Furthermore, the present study lacks baseline comparisons

to human emergency operators or alternative fully automated systems. Finally, the absence of a user-centered evaluation with actual emergency operators means the system's true impact on cognitive load, situational awareness, and operational efficiency has yet to be empirically validated in a real-world setting.

## CONCLUSION AND FUTURE WORK

This paper presented ongoing work on the design and initial evaluation of a human-in-the-loop AI system intended to support emergency call center operations during large-scale crisis events. The proposed system integrates speech transcription, natural language processing, probabilistic severity estimation, and spatial-semantic incident consolidation to assist operators in triaging incoming calls and maintaining situational awareness under conditions of high uncertainty and surge demand. The results reported in this study represent an early-stage technical assessment rather than a completed operational validation. While preliminary experiments using an operationally grounded synthetic dataset indicate promising classification performance and well-calibrated confidence estimates, these findings should be interpreted as evidence of feasibility rather than proof of effectiveness in real-world emergency settings. In particular, the current evaluation does not yet quantify the system's impact on operator workload, decision quality, or response outcomes, nor does it include comparisons against alternative consolidation or triage baselines.

Several limitations, therefore, remain. The dataset used in this study, while designed to reflect realistic emergency call characteristics, does not capture the full variability of live call center environments. Additionally, the absence of user-centered evaluation with emergency professionals limits conclusions about usability, trust, and integration into existing workflows. Dealing with these constraints of validity is the major concern for our ongoing research. In future, emphasis will be on conducting empirical studies by testing our system with anonymized data from actual emergency calls. Also, there will be a greater stress on conducting user-centric evaluations with the help of live emergency operators. Such user-centric evaluations will be conducted to thoroughly test how usable our system is and how much it can enhance trust and efficiency of the operators in real-life circumstances. Finally, technical evaluations will include conducting comparisons with baseline methods that include manual processes and traditional rule-based approaches to emergency triage.

## REFERENCES

Al-Olimat, H. S., Thirunarayan, K., Shalin, V. L., & Sheth, A. (2018). Location name extraction from targeted text streams using gazetteer-based statistical language models. *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, 1986–1997. https://doi.org/10.48550/arXiv.1708.03105

Atherley, L. T. (2025). Intelligent risk management: natural language processing real-time triage of police calls for service. *Police Practice and Research*, *26*(6), 654–672. https://doi.org/10.1080/15614263.2024.2388210

Attiah, A., & Kalkatawi, M. (2025). AI-powered smart emergency services support for 9-1-1 call handlers using textual features and SVM model for digital health optimization. *Frontiers in Big Data*, *8*, Article 1594062. https://doi.org/10.3389/fdata.2025.1594062

Costa, D. B., Pinna, F. C. A., Joiner, A. P., Rice, B., Souza, J. V. P., Gabella, J. L., Andrade, L., Vissoci, J. R. N., & Néto, J. C. (2023). Ai-based approach for transcribing and classifying unstructured emergency call data: A methodological proposal. *PLOS Digital Health*, *2*(12), e0000406. https://doi.org/10.1371/journal.pdig.0000406

Dahlan, A. F., & Yuangyai, C. (2024). T5-based named entity recognition for social media: A case study for location extraction. *2024 IEEE International Conference on Industry 4.0 (IAICT)*, 354–359. https://doi.org/10.1109/IAICT62357.2024.10617592

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 4171-4186.

Domfeh, E. A., & Dancy, C. L. (2025). Human-ai use patterns for decision-making in disaster scenarios: A systematic review. *arXiv preprint arXiv:2509.12034*. https://doi.org/10.48550/arXiv.2509.12034

Eligüzel, N., Çetinkaya, C., & Dereli, T. (2022). Application of named entity recognition on tweets during earthquake disaster: A deep learning-based approach. *Soft Computing*, *26*(1), 395–421. https://doi.org/10.1007/s00500-021-06370-4

Haas, A. J., Gibbons, D., Dangel, C., & Allgeier, S. (2011). Automated surveillance of 911 call data for detection of possible water contamination incidents. *International Journal of Health Geographics*, *10*, 22. https://doi.org/10.1186/1476-072X-10-22

Hanny, D., Schmidt, S., Gandhi, S., Granitzer, M., & Resch, B. (2025). A multimodal geoai approach to combining text with spatiotemporal features for enhanced relevance classification of social media posts in disaster response. *Big Earth Data*, 1–45. https://doi.org/10.1080/20964471.2025.2572140

Hong, L., Song, X., Anik, A. S., & Frias-Martinez, V. (2025). Dynamic fusion of large language models for crisis communication. *Proceedings of the International ISCRAM Conference*. https://doi.org/10.59297/nqysjq45

Hughes, A. L., & Clark, H. (2025). Seeing the storm: Leveraging multimodal llms for disaster social media video filtering. *Proceedings of the International ISCRAM Conference*. https://doi.org/10.59297/f9bnkx60

Kaufhold, M. A., Rupp, N., Reuter, C., & Habdank, M. (2020). Mitigating information overload in social media during conflicts and crises: design and evaluation of a cross-platform alerting system. *Behaviour & Information Technology*, 39(3), 319-342. https://doi.org/10.1080/0144929X.2019.1620334

Lanka, S. (2025). Architectural patterns for AI-enabled triage and crisis prediction systems in public health platforms. *International Journal of Research and Applied Innovations*, *8*(1), 11648–11662. https://doi.org/10.15662/IJRAI.2025.0801003

Ma, Z., Li, L., Li, J., Hua, W., Liu, J., Feng, Q., & Miura, Y. (2025). A multimodal, multilingual, and multidimensional pipeline for fine-grained crowdsourcing earthquake damage evaluation. *arXiv preprint arXiv:2506.03360*. https://doi.org/10.48550/arXiv.2506.03360

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, Y. (2018). Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems*, *36*(4), Article 40. https://doi.org/10.1145/3202662

Minulescu, D.-E., & Toma, S¸.-A. (2025). Whisper based speech recognition for emergency services. *2025 17th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 1–6. https://doi.org/10.1109/ECAI65401.2025.11095458

Palen, L., & Anderson, K. M. (2016). Crisis informatics—New data for extraordinary times. *Science*, 353(6296), 224-225. DOI:10.1126/science.aag2579

Pope, C., Turnbull, J., Jones, J., Prichard, J., Rowsell, A., & Halford, S. (2017). Has the NHS 111 urgent care telephone service been a success? case study and secondary data analysis in england. *BMJ Open*, *7*(5), e014815. https://doi.org/10.1136/bmjopen-2016-014815

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*. https://doi.org/10.48550/arXiv.2212.04356

Seo, J. W., Park, S. J., Kim, Y. J., Kim, J. Y., Kim, K. G., & Yoon, Y. H. (2025). Artificial intelligence for severity triage based on conversations in an emergency department in korea. *Scientific reports*, *15*(1), 16870. https://doi.org/10.1038/s41598-025-99874-0

Svensson, M., & Pesämaa, O. (2018). How Does a Caller's Anger, Fear and Sadness Affect Operators' Decisions in Emergency Calls?. *International Review of Social Psychology*, 31(1), 7. https://doi.org/10.5334/irsp.89

Thuestad, J. A., & Grutle, Ø. (2023). *Speech-to-text models to transcribe emergency calls* [Master's thesis, The University of Bergen]. https://hdl.handle.net/11250/3083251

Williams, C. Y. K., Zack, T., Miao, B. Y., Sushil, M., Wang, M., Kornblith, A. E., & Butte, A. J. (2024). Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA network open*, *7*(5), e248895. https://doi.org/10.1001/jamanetworkopen.2024.8895