Using Large Language Models to Enhance Emergency Response Systems: A Context Awareness Approach

Mohamed Khalil Brik
The American University in Cairo
mohamedkhalil.brik@aucegypt.edu

Dr. Amr El Mougy
The American University in Cairo
amr.elmougy@aucegypt.edu

Dr. Sherif Aly
The American University in Cairo
sgamal@aucegypt.edu

Abstract

Classical emergency response systems, such as 911, are not designed to process multiple emergency calls in populated areas consistently and efficiently. Human operators work long hours handling exhausting calls, which puts them under intense mental pressure. Therefore, these critical systems are prone to human errors. In this research, we propose a complete system that uses Artificial Intelligence to support human operators in responding to 911 calls. Our novel approach uses sentiment analysis on the caller's voice to determine their age group and the urgency level of the call. We also use a Named Entity Recognition model to derive the Caller's Name and the Emergency Address. Our Classifer Model determines the Emergency Type: Medical, Mental Health, Crime, etc. We have also trained Modular Large Language Models to handle each type of emergency. The LLMs are trained on official manuals from official public safety organizations. Our system is aware of the context of the emergency and uses information about the emergency, such as the address and the name of the caller, to give customized recommendations. The system evaluates the operator's performance and gives break reminders.

I Introduction

Emergency Response (ER) is the most critical stage in the emergency management process. A direct coordinated intervention after the incident can be the reason for saving lives and protecting public safety. Information systems to improve emergency response have existed since the 1960's. These systems enhance incident location tracking, resource allocation, and communication between responders and impacted individuals. However, these systems overlook the need to support the responder while handling the emergency. Once they answer the emergency call, the responders are overwhelmed with the vast amount of information they are required to simultaneously. As a result, they might fail to make the right decisions and eventually cause harm to individuals involved in the scene. This research introduces three novel solutions to emergency response systems to better support the responders.

First, we will use sentiment analysis on the call to determine the stress and anger levels of the responder and the individual (s) in the emergency. These metrics are an informative tool for the responder to evaluate the caller's case. It is also a metric for a classifier system that will suggest to the responder when they need to take a break before handling the next call. This step is crucial because many responders overlook taking breaks, which decreases the efficiency of their response.

Second, we will use a classifier to decide the type(s) of the emergency. After proper classification, the relevant modular LLMs are invoked to handle the call. The LLMs will have access to the call transcript in real time. Based on each LLM's specialty, they will give the responder recommendations through notes, questions, procedures, and resources. If the call involves multiple emergency types, a reconciler LLM will receive the recommendations of the modular LLMs and produce the final combined output.

Third, if the caller's name or national ID is correctly mentioned, our system will look up their medical records to inform the responder about aspects relevant to the emergency, such as physical disabilities, previous surgeries, allergies, and the emergency contact person if needed. The invoked LLMs will access this information to give the responder better context-aware recommendations.

The rest of this paper will provide further back-

ground information on our approach and the technologies in place. Next, we will explain our system architecture. Later, we will present our methodology for developing this system and how to achieve integration with existing emergency response systems. In the results section, we will elaborate on the quantitative and qualitative outcomes we reached from this research. Finally, in the last two sections, we will discuss our contributions and provide further research directions.

II Background and Motivation

1 History of Emergency Response Systems

The first use of a national emergency telephone number was in the United Kingdom in 1937, when people dialed the 999 number [5]. In 1968, the Alabama Telephone Company introduced the 911 emergency contact number for the first time in the United States. The first 911 call was made in Haleyville, Alabama, by Speaker of the House Rankin Fite and answered by U.S. Representative Tom Bevill [4]. Enhanced 911 (E911) systems started between the 1980s and the 1990s. E911 allowed new features, such as automatic location identification of the call, which improved this line's response time [4]. Since the early 2000s, 911 dispatchers have started to use computer-aided dispatch (CAD) to record EMS, police, and fire services logs. CAD started to send messages to the dispatched via a mobile data terminal (MDT), which stores and retrieves data (i.e., radio logs, field interviews, client information, schedules, ...). Today, with the Artificial Intelligence revolution, Emergency Response Systems should use this new technology to achieve enhancements that otherwise could not be accomplished with classical technologies.

2 Challenges of Emergency Response Systems

Modern Emergency Response Systems face multiple challenges.

First, responders' response time to answer an urgent call is the decisive factor of life and death in many critical scenarios. In metropolitan cities with high crime rates, ERSs receive hundreds of calls, which overwhelm the human responders and decrease the overall quality of their service. In 2024, officers in New York responded to 911 calls on average in 15 minutes and 23 seconds. This time is the duration of the response from the call until the officers are on the ground at the scene [6].

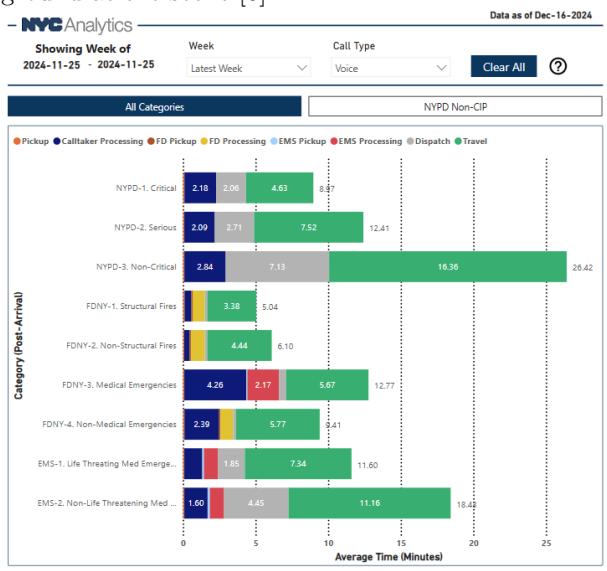

Figure 1: End-to-End Response Time in NYC

Another problem with ERSs is the quality of response. The responder evaluates the call's urgency and decides whether to dispatch the right help. However, responders sometimes fail to take critical cases seriously, which leads to catastrophic results on the ground. A 2021 study highlighted that only 65% of mental health case responders correctly classified during the initial call compared to the classification of cleared call types after the on-the-ground responders investigated the case [7].

The quality of the response depends on the responder's training and well-being. Emergency responders need to stay clear-minded and calm as they talk to people in the middle of domestic violence situations, fires, car accidents, and even murders. The job necessitates multitasking, problem-solving, and facing heart-breaking situations while maintaining professionalism. A study shows that most dispatchers don't get sufficient support for their mental health, making them vulnerable to burnout and mental illnesses such as depression and post-traumatic stress disorder [8].

3 Large Language Models

Large Language Models (LLMs) such as GPT3.5 and Bard are Artificial Intelligence-based systems designed to understand, generate, and manipulate human language. LLMs are built on deep learning architectures like transformers and are trained on vast amounts of text data from various sources such as the internet, books, articles, etc. The variety of data LLMs are trained on enables them to predict and generate creative texts based on human prompts. LLMs use cases cover a wide range of tasks such as transla-

tion, summarization, question answering, programming, and classification. OpenAI announced that ChatGPT's weekly users have grown to 200 million [3]. This large number of users shows the rising dependency on LLMs in different disciplines. This large number of users also allows the model to improve its quality of response. The accurate output of LLMs depends on the quality of the data used in the training phase and the prompt given by humans, a field known as Prompt Engineering. Today, LLMs face challenges such as alignment, training dataset quality, and the cost of training [2].

III Methodology

1 System Overview

In this research, we suggest the below architecture to integrate Large Language Models and Machine Learning Classification techniques into the Emergency response systems. Our approach starts with performing a speech-to-text conversion of the call in real time. Then, we use a Named Entity Recognition Model to extract the Name of the caller and/or their national ID, the address of the emergency, and the names of any people involved in the scene. Consequently, we use a classifier-based Large Language Model that analyses the phrases of the emergency seeker to determine the type of emergency the responder is dealing with. When the type of emergency is specified, the corresponding Modular Large Language Model will be invoked to analyze the call. All the corresponding LLMs will be invoked if there are multiple types of emergencies, such as medical and crime emergencies. At the end, a reconciler LLM will generate the final recommendations to the responder. A sentiment analysis module will continuously analyze the voice of both the caller and the responder to generate the emotions score, including joy, anger, stress, confidence, etc. These scores will help the responder to evaluate the case and provide and dispatch the correct units. We will also use these scores to assess the responder's efficiency in handling the calls and remind them when to take a break to handle further calls as efficiently as possible. The name and/or national ID of the caller extracted from the Named Entity Recognition model will be used to search the caller's medical information in health records databases. The assistant LLM will use the medical condition information, such as medications used, allergies, previous surgeries, and mobility issues, to give personalized recommendations to the responder. We will also use the address of the incident to study the surrounding neighborhood of the location. The system will flag critical locations relevant to the emergency type, such as gas stations, schools, or nuclear facilities.

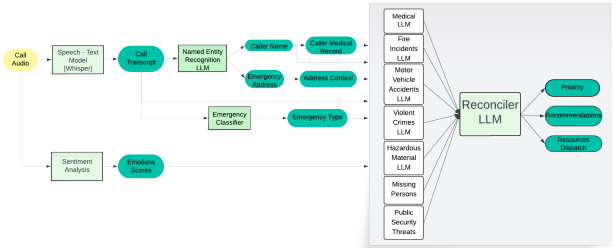

Figure 2: System Architecture

2 Call Transcription

Large Language Models only operate on text. As a result, for our models to analyze the call, we need to convert the speech to text in real time. For this purpose, we will be using the Whisper model by Open AI. This model's choice has many advantages for our use case. First, Whisper provides better robustness in transcription compared to previous models. It makes 55.2% less eros than state-of-the-art LibriSpeech [9]. Second, Whisper provides multi-lingual speech recognition and translation for over 50 languages [9]. The translation is critical to improving the communication between the responders and the callers, who may find it challenging to communicate in English, especially during an emergency. Third, Whisper proved to be very robust with a voice that has additive noise [9]. This feature is significant in emergency calls that are often made in loud environments where the responder alone cannot hear critical information from the caller.

3 Named Entity Recognition

The Named Entity Recognition base model we used is the "bert-large-cased-finetuned-conll03-english" [15]. We used this model to extract the "caller's name" and the "emergency address." The system will use the caller's name to extract the medical information from the medical database we simulated in our implementation. The medical LLM will use the extracted data to generate personalized recommendations on the caller's case.

The system performs a search on the address of the urgency to study the nearby buildings. For instance, a building type is related to an emergency, such as a fire emergency near a gas station. In that case, that location will be flagged on a map to the responder and used by the LLM to give recommendations accordingly. In the results section we will show the

accuracy of using this model.

4 Emergency Classification

This table explains the different types of emergencies with the vehicles needed to dispatch.

Table 1: Different Emergency Types

Type	Description	Units
Medical	Heart attacks, strokes, injuries, or unconscious individuals.	Ambulance, Advanced Life Support Unit
Fire Incidents	Structural fires, vehicle fires, wildfires, or uncontrolled flames.	Fire Engine, Ladder Truck, Battalion Chief.
Motor Vehicle Accidents	Collisions involving vehicles	Police Patrol Unit, Ambulance, Fire Engine or Rescue Unit.
Violent Crimes	Physical attacks, robberies, or active shooter situations.	Police Patrol Unit, SWAT Team, Ambulance.
Burglaries	Unauthorized entry into buildings intending to commit theft or other crimes.	Police Patrol Unit, K-9 Unit.
Hazardous Material Incidents	Release of dangerous substances with health or environmental risks.	HazMat Unit, Fire Engine, Ambulance.
Domestic Disturbances	Conflicts within a household involving violence or threats.	Police Patrol Unit, Crisis Intervention Team.
Missing Persons	Reports of unaccounted individuals.	Police Patrol Unit, Search and Rescue.
Suspicious Packages/Bomb Threats	Unattended items suspected to be dangerous or direct threats of explosives.	Police Patrol Unit, Bomb Squad, Fire Engine and Ambulance.

5 Large Language Model Trainings

Large Language Models (LLMs) have general-purpose Natural Language Processing capabilities that users apply to a variety of text tasks. LLMs are biased because they are large language models [10]. They are not expert systems. There are two main approaches to specialize LLMs: Fine-tuning and Prompt Engineering.

Fine-tuning involves training the model on task-specific or domain-specific data to adjust its weights for that particular task. The foundational model is trained on field-related datasets and cannot process inputs from other fields. Fine-tuning is a computationally expensive process requiring specialized hardware like GPUs or TPUs and a long training period [12].

Prompt Engineering designs input prompts to wrap the model's behavior around specific tasks. However, it does not modify the model's weights or underlying structure. Prompt Engineering has a low cost, and no extra training or resources are needed [11].

In this research, we will use the Prompt Engineering approach to create large language models specialized for responding to specific types of emergencies.

6 Medical Large Language Model

We used the Llama 2 model by the Meta team(13) for the medical Large Language Model. In Figure 4, you will find the prompt we engineered to use this model as a Medical Emergency specialist. We also used a general Dispatcher Training Manual developed by the Sebastopol Police Department (14). This manual provides general information on ethics, criminal justice in the US, standard law enforcement abbreviations, and essential dispatch functions that human responders need to know. Besides, we collected a dataset of medical allergies and conditions with descriptions for each (15). This dataset will allow our medical model to identify any medical state described by the caller and provide the standard procedures to follow.

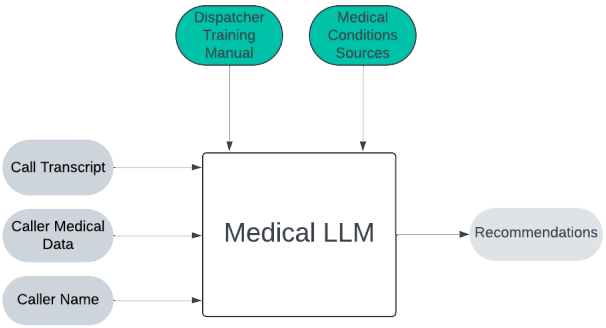

Figure 3: Medical LLM

You are a 911 medical assistant designed to help emergency responders by providing accurate, fast, and precise guidance during emergency medical calls. Your role is to assist in assessing symptoms, offering life-saving interventions, and providing instructions to both callers and responders in real-time. You must ensure that your responses are calm, concise, and rooted in established medical protocols for emergency situations.

- Provide step-by-step instructions for emergency medical procedures (e.g., CPR, choking, bleeding control).
- Offer real-time assessments of symptoms based on the information provided by the caller, including triaging severity.
- Provide guidance on stabilizing a patient until paramedics or medical teams arrive.
- Ensure clear and empathetic communication, especially when dealing with panicked or distressed callers.
- Be ready to assist pediatric, adult, and elderly patients in emergency medical scenarios.

Use the following two sources to guide your responses and provide accurate medical information:

- Dispatcher Training Manual
- Medical Data Set

Figure 4: Medical LLM Prompt Engineering

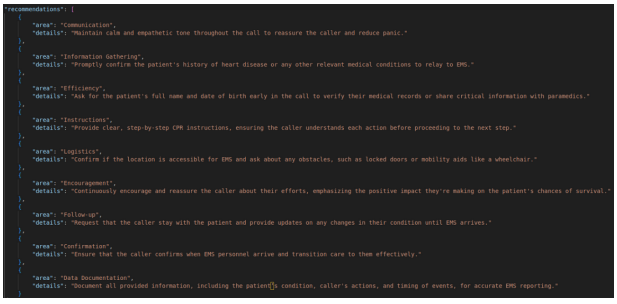

Figure 5: Medical LLM Sample Output

IV Results and Analysis

1 Named Entity Recognition Results

We have tested the NER model on a real 911 call dataset. The total number of calls is 518, covering a diverse range of emergencies. Table 2 summarizes the results of the entire dataset. Not all real calls include the caller's name and the emergency's address. Our model is only able to detect names and addresses that are mentioned in the call transcript. Out of the 518 calls, 76.5% of the calls included the address that our model extracted, and 48.3% of the calls had names. The "Both" field shows that about 39% of the calls included the name and the address, while the remaining 61% either included the name only, the address only, or neither. These results prove that our model can extract the data when mentioned. However, the emergency responder must ask the caller about their name and address to use the full capabilities of our system. Figure 5 shows the sample output of the results.

Field	Non-null Count	Null Count	Non-null %	Null %
Name	250	268	48.3%	51.7%
Address	397	121	76.5%	23.5%
Both	203	74	39.2%	14.3%

Table 2: NER Results

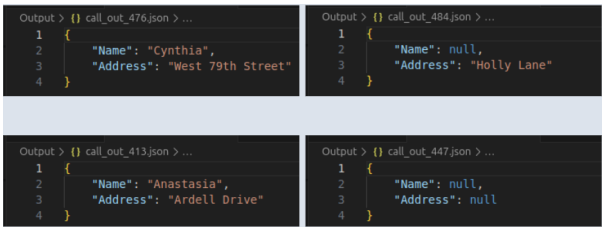

Figure 6: Named Entity Recognition Sample Results

2 Software Interface for Emergency Responders

Our approach requires an upgrade to the software emergency responders use to process the different emergencies. The software we developed, shown in the figure below, will include a map that will flag the location of the emergency and the relevant surrounding buildings as well as the available units to handle the emergency. We also included a table that summarizes the medical information for the caller.

The dispatch bottom allows the responder to quickly choose which emergency team is suitable for the emergency. The LLM recommendations part will give notes, questions, procedures, and which vehicles to dispatch to the human emergency reposnder.

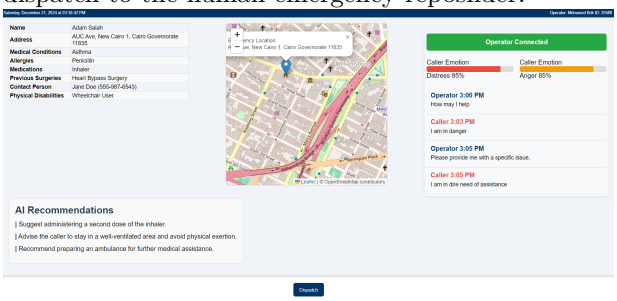

Figure 7: Software for Emergency Responders

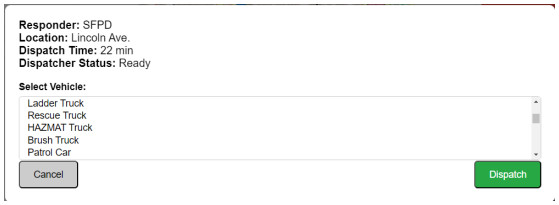

Figure 8: Dispatch Vehicles

V Discussions and Future Work

This part of the paper will provide key points that further research should investigate.

First, we used prompt engineering to train the models. The other alternative is to use Fine-tuning on task-specific datasets. Fine-tuning could provide better accuracy than Prompt Engineering. The same dataset we used for the medical LLM could be used to retrain LLMs and update the weights of the neural networks. A comparative study is needed to study the difference between using both approaches across the different types of emergencies, specific LLMs. Our modular LLMs' advantage is that each Model is trained, tested, and deployed independently. As a result, we could have a system with LLMs trained using Fine tuning and others with Prompt Engineering. The modular approach allows researchers to choose different training methods and different foundational models for distinct emergencies.

Second, we need to investigate privacy issues associated with using LLMs to access citizens' private information, such as medical records. The Model needs to use personal information to provide personalized recommendations, improving the accuracy of the respondents' support. Privacy protection could be achieved using Role-Based Access Control (RBAC). The Model will only access the caller's medical information in a medical emergency. Furthermore, the system must comply with legal standards such as the Health Insurance Portability and Accountability Act(HIPAA). This act aims to protect the privacy and security of individuals' health information. HIPAA sets national standards for protecting health information, applying to healthcare providers and businesses handling sensitive data.

Third, the system should handle significant catastrophic events like earthquakes differently than case-by-case incidents. It should cluster the calls related to the same event and generate useful information for emergency responders to understand the heavily impacted locations and the resources to deploy.

VI Conclusion

In conclusion, in this paper, we used Large Language Models to enhance emergency response systems. We started by specifying the problems of current emergency response systems. We then introduced Large Language Models as a new technology that will revolutionize how we design software systems. We also developed a system that uses classical algorithms such as search integrated with LLM abilities to give tailored recommendations to emergency responders in real-time to increase their efficiency in handling a diverse range of emergencies while prioritizing responders' mental health. Further work is needed to add other features and use different types of LLMs to improve the overall performance of emergency response systems.

References

[1] Shahrah, A. Y., Alfawareh, H. M., Farea, M. M., & Thabtah, F. (2017). Emergency response systems: Proceedings of the second international conference on internet of things, data and cloud computing. *ACM Other Conferences*. https://doi.org/10.1145/3018896.3056778

[2] Naveed, H., et al. (2024). A comprehensive overview of large language models. *Preprint submitted to Elsevier*. https://arxiv.org/pdf/2307.06435

[3] Reuters, "OpenAI says ChatGPT's weekly users have grown to 200 million," 2024. https://www.reuters.com/technology/artificial-intelligence/openai-says-chatgpts-weekly-users-have-grown-200-mill

[4] Iredell County, NC. (n.d.). History of 9-1-1. *History of 9-1-1 — Iredell County, NC*. Retrieved December 19, 2024, from https://www.iredellcountync.gov/1768/History-of-9-1-1

[5] BT PLC, "BT PLC News Article," 2007. https://web.archive.org/web/20090114072201/http://www.btplc.com/News/Articles/ShowArticle.cfm?ArticleID=6e55cb12-8c0c-417f-b68c-6a7f62b1d8c8

[6] McCarthy, C., & Bhole, A. (2024, September 16). New Yorkers are waiting the longest in decades for cops to respond to 911 calls, crimes. *New York Post*. https://shorturl.at/ByYGX

[7] Simpson, R., & Orosco, C. (2021, December 8). Re-assessing measurement error in police calls for service: Classifications of events by dispatchers and officers. *PLOS ONE*. https://pmc.ncbi.nlm.nih.gov/articles/PMC8654180/?utm_source=chatgpt.com

[8] Robb-Dover, K. (2024, November 11). Mental health care for 911 dispatchers. *FHE Health*. https://fherehab.com/learning/dispatcher-mental-health-care

[9] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (n.d.). Robust speech recognition via large-scale weak supervision. *OpenAI*. https://cdn.openai.com/papers/whisper.pdf

[10] Resnik, P. (2024). Large language models are biased because they are large language models. *arXiv*. https://doi.org/10.48550/arXiv.2406.13138

[11] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024, February 5). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv*. https://arxiv.org/abs/2402.07927#

[12] Fu, Z., Yang, H., So, A. M.-C., Lam, W., Bing, L., & Collier, N. (2022). On the effectiveness of parameter-efficient fine-tuning. *arXiv*. https://arxiv.org/abs/2211.15583

[13] Touvron, H., Martin, L., Stone, K., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *GenAI, Meta*. https://arxiv.org/pdf/2307.09288

[14] Sebastopol Police Department. (2022, April 28). *Dispatcher training manual* (Chief of Police Ronald Nelson). https://www.cityofsebastopol.gov/wp-content/uploads/2023/06/Dispatch-Training-Manual-1.pdf

[15] DBMDZ. (2024). *BERT large cased fine-tuned on CoNLL-03 English*. Hugging Face. https://huggingface.co/dbmdz/bert-large-cased-finetuned-conll03-english