CLEAR-Command: Coordinated Listening, Extraction, and Analysis for Emergency Response with Large Language Models

Achref Doula[*1], **Bela Bohlender**[*1], **Max Mühlhäuser**[1], **Alejandro Sanchez Guinea**[2]
[1]Technical University of Darmstadt, Germany
[2]NTT Data, Luxembourg
doula@tk.tu-darmstadt.de, bela@bohlender.email,
max@informatik.tu-darmstadt.de, alejandro.guinea@global.ntt

Abstract

Effective communication is vital in emergency response scenarios, where clarity and speed can save lives. Traditional systems often struggle under the chaotic conditions of real-world emergencies, leading to breakdowns in communication and task management. In this paper we introduce the **CLEAR** (**C**oordinated **L**istening, **E**xtraction, and **A**nalysis for Emergency **R**esponse)-Command system, which leverages Large Language Models (LLMs) to enhance emergency communications. CLEAR-Command automates the transcription, summarization, and task extraction from live radio communications of emergency first responders using the OpenAI Whisper API for transcription and Chat-GPT 4 for summarization and task extraction. We decided for Chat-GPT 4 after conducting an expert pre-study that showed it to be the most accurate LLM in terms of task extraction for our case. To evaluate our system, we conducted a user study with 13 participants. Our results show that CLEAR-Command significantly outperforms traditional radio communication in terms of clarity, trust, and correctness of task extraction. The link to a live demo website of our system is https://clear-command.vercel.app. The video demonstrating our system can be found on https://youtu.be/ZF3HMMUEq9o. All project details are presented in our Gitlab page https://gitlab.com/achref.d/clear-command.

1 Introduction

In search and rescue operations (SAR) during emergency scenarios, the efficiency of communication among first responders, such as firefighters and medical intervention teams, is critical to managing incidents effectively (Shittu et al., 2018). Critical situations are often marked by high pressure and the need for rapid decision-making, where clear and concise communication can mean the difference between life and death. However, the complex nature of such environments can lead to communication breakdowns, overwhelming the responders with fragmented and sometimes redundant information (Saoutal et al., 2014; Willms et al., 2019). The primary challenge in these settings is maintaining situational awareness while managing a high volume of communications, often under severe time constraints (Chehade et al., 2020). Traditional emergency communication systems often fall short in the face of complex and evolving scenarios (Menold et al., 2015), where information overload can lead to critical details being missed or miscommunication, which may cause critical incidents and mistakes (Chałupnik and Atkins, 2020).

The recent advancements in natural language processing, particularly the development of large language models (LLM) and multimodal large language models (MLLM), offer substantial promise for enhancing emergency communication systems (Doumanas et al., 2024). State-of-the-art LLMs and MLLMs are adept at transcribing spoken communication (Amodei et al., 2016), analyzing complex textual data, and systematically extracting actionable plans (Song et al., 2023; Wang et al., 2022). Despite concerns regarding the potential for model-generated hallucinations and inaccurate predictions, the incorporation of context-aware processing capabilities assists these models in adapting effectively to the fluctuating dynamics of complex environments (Ji et al., 2023; Chun et al., 2023). Furthermore, the integration of human oversight within human-in-the-loop systems along with uncertainty quantification approaches enhance the reliability of the outputs while concurrently mitigating risks inherent in autonomous decision-making systems (Han et al., 2024; Quach et al., 2023).

In this paper, we introduce the **CLEAR** (**C**oordinated **L**istening, **E**xtraction, and **A**nalysis for Emergency **R**esponse)-Command system,

[*] Authors contributed equally to this work.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), pages 20–28
April 30, 2025 ©2025 Association for Computational Linguistics

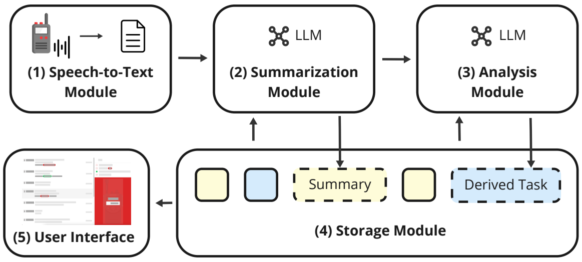


Figure 1: **CLEAR-Command System Architecture.**. The system is composed of a (1) Speech-to-Text Module for transcribing the audio transmissions, an (2) Summarization Module for summarizing the transcriptions, a (3) Analysis Module for task extraction, a (4) Storage Module for providing historical context, and a (5) user interface for the firefighter to access and interact with the system.

which leverages the recent advancements in LLMs to assist SAR teams organizing their communications. CLEAR-Command automates the organization and summarization of vocal communications, transforming complex streams of speech data into structured, actionable tasks. CLEAR-Command consists of a speech-to-text transcription model that transforms spoken communications into text and an LLM that summarizes the relevant parts of the communication and extracts the tasks and their status. This allows emergency first responders to effectively monitor the progression of tasks and provides them with a dynamic, real-time overview of task allocation and execution status. The specific LLM used by CLEAR-Command is GPT-4. We decided on this, based on a pre-study with 4 experts that assessed the quality of text transcription, summarization, and task extraction that different LLMs showed when dealing with emergency scenarios. To evaluate our system, we conducted a user study with 13 participants. Our results show that CLEAR-Command significantly outperforms traditional radio communication in terms of clarity, trust, and correctness of task extraction.

The link to a live demo website of our system is `https://clear-command.vercel.app`. The video demonstrating our system can be found on `https://youtu.be/ZF3HMMUEq9o`. All project details are presented in our Gitlab page `https://gitlab.com/achref.d/clear-command`.

2 Related Work

2.1 Information Extraction

The evolution of information extraction marks a significant transition from rule-based methodologies (Chiticariu et al., 2013; Li et al., 2011) to deep learning frameworks (Veyseh et al., 2022; Liu et al., 2020; Ren et al., 2023). Recent advancements leverage large language models to identify relevant entities and actions within unstructured text (Nakshatri et al., 2023; Cai et al., 2023; Gestrin et al., 2024; Kirk et al., 2024) for tasks ranging from named-entity recognition to robot task extraction from textual input. In this work, we leverage LLMs to extract tasks from transcribed radio communications in ASR missions.

2.2 Speech-to-Text Transcription

Early approaches to Voice-to-text transcription primarily relied on Hidden Markov Models (HMMs) combined with Gaussian Mixture Models (GMMs) (Cui and Gong, 2007) and feedforward neural networks. With the advent of deep learning, more sophisticated models such as Recurrent Neural Networks (Amodei et al., 2016; Saon et al., 2021; Hori et al., 2018), Long Short-Term Memory networks (Soltau et al., 2016; Zhang and Lu, 2018), and Transformer-based models (Kim et al., 2022; Dong et al., 2018; Gulati et al., 2020) have become prevalent. Shifting towards cloud-hosted API services, modern speech recognition solutions harness these services to deliver scalable and integrative capabilities. Cloud-hosted API services, such as OpenAI Whisper, Google Cloud Speech-to-Text, and IBM Watson Speech-to-Text, provide entry points for real-time speech recognition that scales and integrates with various applications. In this work, we use the OpenAI Whisper API to transcribe the radio communications to text, to be processed by LLMs for summarization and task extraction.

2.3 Text Summarization

Text summarization has evolved significantly with the adoption of deep learning, transitioning from traditional extractive techniques that select key sentences (Goularte et al., 2019; Tang et al., 2020) to more sophisticated abstractive methods that generate new sentences (Gerani et al., 2014; Shen et al., 2023), capturing deeper nuances of the original texts. Transformer models have been pivotal in this advancement, learning from large datasets to produce contextually relevant and fluent summaries (Pilault et al., 2020). Reinforcement learning has further refined these models, enhancing their adaptability and accuracy based on user feedback (Ramamurthy et al.). Additionally, the emergence of domain-specific models, trained on specialized datasets, ensures that summaries in fields like medicine or emergencies are not only precise but also practically applicable, handling domain-specific terminologies and styles effectively (Otal and Canbaz, 2024; Thirunavukarasu et al., 2023).

3 CLEAR-Command System

The CLEAR-Command system architecture is composed of 4 modules and a user interface, as illustrated in Figure 1. The Speech-to-Text Module automatically converts vocal communications into text, which is used by the Summarization Module to succinctly summarize crucial information that is then passed to the Analysis Module to derive and manage tasks. The Storage Module stores summaries and derived tasks, which are fed back to the Summarization Module and the Analysis Module for context. The User Interface displays a comprehensive real-time overview of summaries and derived tasks taken from the Storage Module. All this process is designed to take place in a real-time manner to enhance situational awareness and response efficiency, minimizing the cognitive load on responders by allowing them to concentrate on operational execution rather than communication management.

Speech-to-Text Module. This module uses the OpenAI's Whisper API[1] to transcribe real-time vocal communications, even in noisy emergency environments.

Summarization Module. The GPT-4[2] model is used to extract concise summaries based on the transcribed text from the Speech-to-Text Module and previous summaries. The summary emphasizes crucial information, which is achieved through iterative prompt design, including clear instructions and examples. The prompt embeds the previous summaries to enhance the situational understanding.

Analysis Module. The GPT-4 model is used to extract structured information, such as tasks, from all previous summaries. The prompt for the Analysis Module is designed iteratively to return the structured information, such as tasks, in the expected format. The prompt contains previous summaries and structured information to be aware of existing information, such as tasks. For instance, an unresolved task can be marked as resolved by the Analysis Module when the provided information indicates that the task is resolved.

Storage Module. This module provides storage for the CLEAR-Command system, allowing the Summarization and Analysis Module to use the existing information from the ongoing emergency. This storage prevents modifications to existing entries and is append-only. The Storage Module preserves the history, including the audio recording, the transcriptions, the summaries, and structured information, to ensure that everything is fully transparent, understandable, and accessible through the User Interface during and after the emergency.

User Interface. The User Interface is built in HTML, CSS, and JavaScript. The User Interface is composed of three panels, which are depicted in Figure 2. The Communication Overview panel shows the previous summaries and structured information from the Storage Module. The Tasks Overview panel shows the derived tasks and their status. The Radio panel allows users to interact with the system through direct communication.

Communication Overview panel. This panel provides the list of previous and current communications, each containing the sender, the time of creation, a transcription, a summarization, and a list of derived structural information, such as created tasks. Resolved tasks are marked as green, and unresolved tasks are marked as red when referenced in the overview. Ongoing communications are marked red, as seen in Figure 2 at the bottom of the Communication Overview Panel.

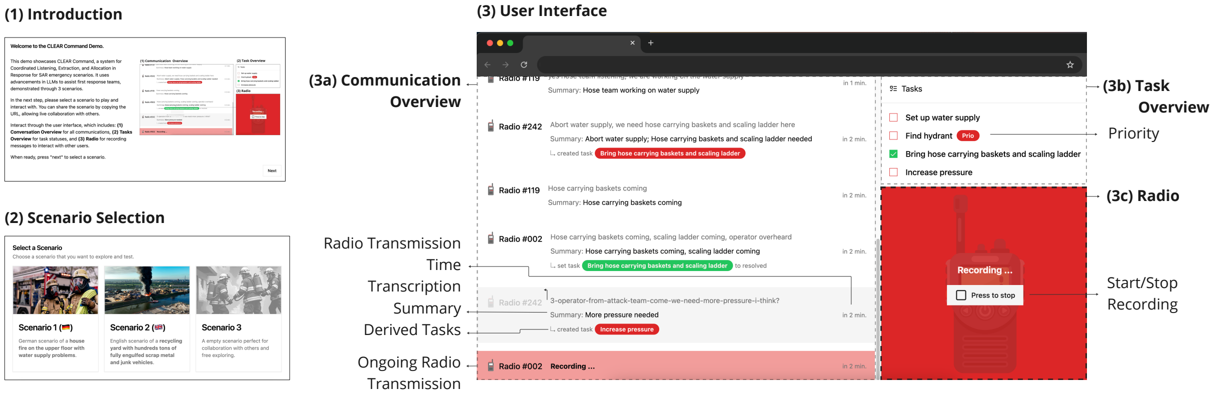
Tasks Overview panel. This panel shows a real-time view of resolved and unresolved tasks, along with an indication of their priority.

Radio panel. This panel contains a button for starting and stopping vocal recordings, which allows to interact with the system by adding new communications that will trigger the transcription, summarization, and analysis process.

Implementation Details. We implemented our system in Typescript running on NodeJS and Python, incorporating libraries from the Hugging Face ecosystem. Our Speech-to-Text and LLM models run on cloud inference endpoints, chosen for their capabilities to support the computational demands of the system's data processing and machine learning tasks and on-demand horizontal scaling to support multi-user access.

[1] Whisper: https://github.com/openai/whisper

[2] GPT-4: https://openai.com/index/gpt-4/



Welcome to the CLEAR Command Demo.

This demo showcases CLEAR Command, a system for Coordinated Listening, Extraction, and Allocation in Response for SAR emergency scenarios. It uses advancements in LLMs to assist first response teams, demonstrated through 3 scenarios.

In the next step, please select a scenario to play and interact with. You can share the scenario by copying the URL, allowing live collaboration with others.

Interact through the user interface, which includes: **(1) Conversation Overview** for all communications, **(2) Tasks Overview** for task statuses, and **(3) Radio** for recording messages to interact with other users.

When ready, press "next" to select a scenario.

Select a Scenario

Choose a scenario that you want to explore and test.

Scenario 1 (🇩🇪)
German scenario of a **house fire on the upper floor with water supply problems.**

Scenario 2 (🇬🇧)
English scenario of a **recycling yard with hundreds tons of fully engulfed scrap metal and junk vehicles.**

Scenario 3
A empty scenario perfect for collaboration with others and free exploring.

Radio #119 yes hose team listening, we are working on the water supply
Summary: Hose team working on water supply — in 1 min.

Radio #242 Abort water supply, we need hose carrying baskets and scaling ladder here
Summary: Abort water supply; Hose carrying baskets and scaling ladder needed — in 2 min.
↳ created task Bring hose carrying baskets and scaling ladder

Radio #119 Hose carrying baskets coming
Summary: Hose carrying baskets coming — in 2 min.

Radio #002 Hose carrying baskets coming, scaling ladder coming, operator overheard
Summary: Hose carrying baskets coming, scaling ladder coming — in 2 min.
↳ set task Bring hose carrying baskets and scaling ladder to resolved

Radio #242 3-operator-from-attack-team-come-we-need-more-pressure-i-think?
Summary: More pressure needed — in 2 min.
↳ created task Increase pressure

Radio #002 Recording ... — in 2 min.

Tasks
- □ Set up water supply
- □ Find hydrant Prio
- ☑ Bring hose carrying baskets and scaling ladder
- □ Increase pressure

Recording ...
□ Press to stop

Figure 2: **CLEAR-Command System Demo.** Users are (1) introduced to the CLEAR-Command System, can (2) select a scenario, and use the (3) user interface to get an (3a) overview of the communication, view the (3b) current tasks, and interact with the system using the (3c) radio during the scenario.

4 Evaluation

We evaluate CLEAR-Command through two main studies: an expert pre-study to assess the quality of summarization and instruction generation across different models and a user study to evaluate usability, trust, and workload in simulated emergency scenarios. The detailed text of the scenarios used in our evaluation, as well as further information about the experts and questions can be found in our Gitlab page https://gitlab.com/achref.d/clear-command.

4.1 Expert Pre-Study

In this study, 4 professionals and researchers with backgrounds in emergency response (firefighter), machine learning, user experience, and semantics assessed the quality of summarization and instruction generation using different models. We used three models for summarization and instruction generation: GPT4, LLama3-70B, LLama3-8B, and LLama3-8B-instruct[3].

During the study, the experts were provided a transcribed vocal interaction between a team of firefighters composed of a captain, a lieutenant, a research and rescue team, a perimeter control team, and a medical team. Furthermore, the outputs provided by the different models structured in (1) conversation summary and (2) extracted tasks were provided for the experts to assess the quality of their outputs. After reading each model output, the experts answered 5 questions structured as a 5-point Likert scale that address (1) the accuracy of task identification and prioritization, (2) the semantic accuracy and completeness, (3) the handling of domain-specific jargon and noisy data, (4) the usability of the summarized content, and (5) the overall effectiveness, respectively.

For the analysis of the results of the pre-study, depicted in Figure3, we use the Shapiro-Wilk test to assess normality and the Wilcoxon signed-rank test for paired samples to evaluate the significance of the results.

GPT-4. "Accurate Task Identification" showed normality with a Wilcoxon test statistic of 0.0, a p-value of 0.12, and a high mean score of 5.00. "Summarization Accuracy" also respected normality, with similar test results and a mean score of 4.00. "Domain-Specific Jargon Handling" did not respect normality, with a test statistic of 0.0, a p-value of 0.25, and a moderate mean score of 3.25. "User-Friendly Organization" showed a high mean score of 4.25, and 'Coherent and Concise Summary' had a mean score of 4.00, both respecting normality.

LLama3-8B. "Accurate Task Identification" respected normality with a test statistic of 0.0, a p-value of 0.12, and a mean score of 4.00. 'Summarization Accuracy' and 'Domain-Specific Jargon Handling' had moderate mean scores of 3.25 and 3.50, respectively. 'User-Friendly Organization' had a mean score of 3.75, and 'Coherent and Concise Summary' respected normality with a mean score of 4.00.

Llama-3-8B-Instruct. "Accurate Task Identification" respected normality with a test statistic of 0.0, a p-value of 0.12, and a mean score of 3.25. 'Summarization Accuracy' did not respect normality, with a low mean score of 2.25. 'Domain-Specific

[3] Llama models: https://huggingface.co/meta-llama

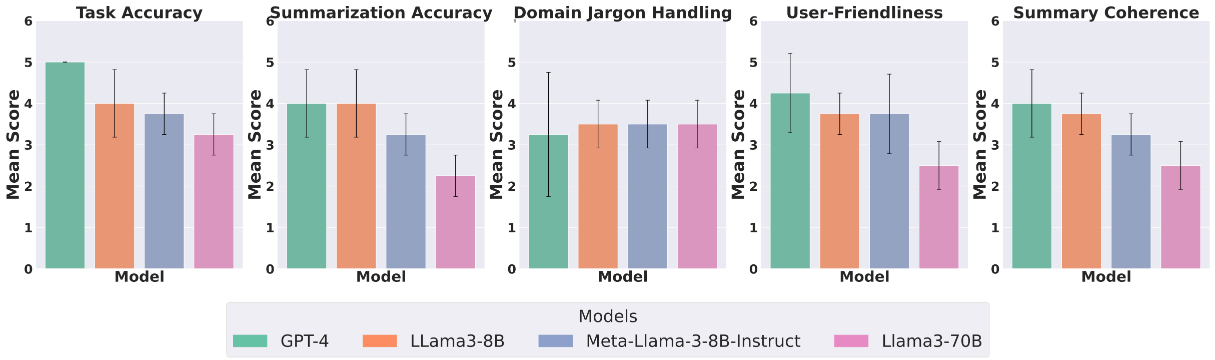


Figure 3: Expert pre-study results for different models (GPT-4, LLama3-8B, Meta-Llama-3-8B-Instruct, and Llama3-70B). Each subplot shows the mean and standard deviation of the expert ratings on a 5-point Likert scale for 5 evaluation metrics: Accurate Task Identification, Summarization Accuracy, Domain-Specific Jargon Handling, User-Friendly Organization, and Coherent and Concise Summary.

Jargon Handling' and 'User-Friendly Organization' had mean scores of 3.50 and 2.50, respectively. 'Coherent and Concise Summary' did not respect normality with a mean score of 2.50.

Llama3-70B. "Accurate Task Identification" respected normality with a test statistic of 0.0, a p-value of 0.1250, and a mean score of 4.00. 'Summarization Accuracy' and 'User-Friendly Organization' both had mean scores of 4.00 and 3.75, respectively. 'Domain-Specific Jargon Handling' had a mean score of 3.50, and 'Coherent and Concise Summary' respected normality with a mean score of 4.00.

4.2 User Study

We conducted a user study to compare CLEAR-Command to a baseline approach based on radio communication, which represents the current approach used for communication and task assignment in emergency interventions for firefighters. For the study, we recruited 10 participants aged between 18 and 34 from diverse backgrounds relevant to emergency response to evaluate the usability and trustworthiness of the system. The study adopts a within-subject design, meaning that participants interacted with our system and the baseline. During the study, participants heard a real interaction between firefighters over the radio, and their task was to write down the tasks that were assigned during the conversation. To assess the quality of CLEAR-Command in comparison to the plain radio communication baseline, we adopt the following metrics: Task Correctness (TC), Ease of Use, Perceived Information Clarity, and Confidence in Accuracy. The statistical analysis of the user study comparing the CLEAR-Command system to the baseline system reveals significant differences across all evaluated metrics. The results are depicted in Figure 4.

Task Correctness. This metric assesses the accuracy of task identification by participants using the CLEAR Command system compared to a baseline radio communication system. Participants' responses are evaluated on a scale from 0 to 10 based on a GPT4 assessment, reflecting how closely their identified tasks align with the correct set of tasks for a given emergency scenario. Higher scores indicate greater precision in task identification. The Mann-Whitney U Test yielded a statistic of 33.5 with a p-value of 0.0176, indicating a significant difference in correctness scores between the baseline and CLEAR-Command systems. Specifically, the CLEAR-Command system demonstrated a higher correctness score with an effect size of 1.034. This substantial effect size suggests that the CLEAR-Command system significantly outperforms the baseline system in accurately extracting tasks from the communication data.

Ease of Use. We measure the usability of the proposed systems in the study, using a 5-point Likert scale addressing the following question: "How easy was it to identify and understand the tasks using this system?". This metric yielded a Wilcoxon statistic of 0.0 with a p-value of 0.0019. This result, combined with an effect size of 2.138, indicates a statistically significant and substantial difference favoring the CLEAR-Command system over the baseline. Participants found it significantly easier to identify and understand tasks using the CLEAR-Command system.

Perceived Information Clarity. We evaluate the

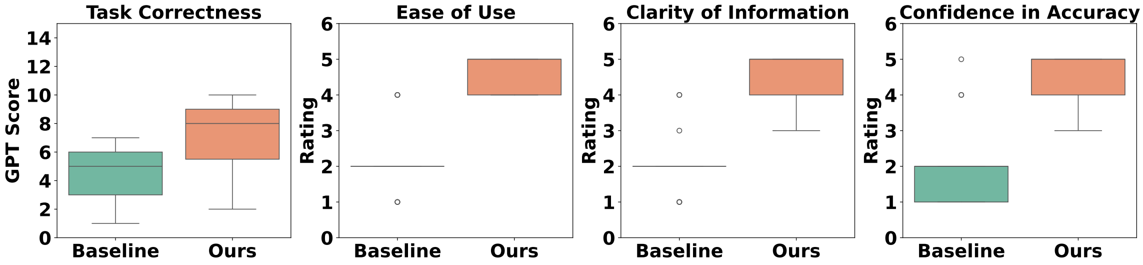


Figure 4: User study results comparing the baseline radio communication system (baseline) and the CLEAR-Command system (ours). Each subplot shows the mean and standard deviation of the user ratings on a 5-point Likert scale for four evaluation metrics: Task Correctness, Ease of Use, Clarity of Information, and Confidence in Accuracy. The results indicate significant improvements in all metrics when using the CLEAR-Command system.

perceived clarity of the information provided by each system using a 5-point Likert scale addressing the following question: "How clear was the information provided by the system in helping you identify the necessary tasks?". In this case, the Wilcoxon statistic was 0.0 with a p-value of 0.0018. The effect size of 1.968 also suggests a significant and notable improvement in the clarity of information provided by the CLEAR-Command system. This demonstrates that the CLEAR-Command system significantly enhances users' ability to comprehend the information needed to identify necessary tasks.

Confidence in Accuracy. We assessed the confidence of participants that they extracted the right tasks from the audio conversation using a 5-point Likert scale addressing the following question: "How confident did you feel about the completeness and accuracy of the tasks you wrote down?". The analysis of this metric showed a Wilcoxon statistic of 1.50 with a p-value of 0.0029. The effect size here was 1.637, indicating a significant and meaningful increase in user confidence regarding task completeness and accuracy when using the CLEAR-Command system.

5 Discussion

CLEAR-Command demonstrates significant improvements in emergency response through effective communication summarization and task allocation. The integration of speech recognition models for transcription and LLMs for summarization and instruction generation showcases the potential of state-of-the-art LLMs in critical situations. However, despite these advancements, certain limitations must be acknowledged.

While CLEAR-Command is effective in enhancing emergency response through communication summarization and task analysis, it has several limitations. Our reliance on specific models for transcription (OpenAI Whisper) and summarization and instruction generation (GPT-4) introduces dependencies on their performance, which may be affected by linguistic diversity and noisy environments. The substantial computational resources required by these models may also limit their deployability in resource-constrained settings.

The validity of our evaluation should be considered in the light of the following constraints. The expert study, involving 4 experts, and the user study, with 13 participants, provide preliminary insights that could be deepened through larger participant samples from different emergency sectors. Due to computational constraints, our evaluation considered solely a zero-shot setup to adapt the models to our specific use case. While achieving the reported results without training demonstrates the capabilities of state-of-the-art LLMs and their potential in emergency use cases, we believe that fine-tuning models can enhance the performance further.

6 Conclusion

In this paper, we introduced CLEAR-Command, a system designed to enhance emergency response efficiency through the transcription, summarization, and task allocation of emergency communications in SAR missions. By integrating OpenAI Whisper for voice-to-text transcription and Chat-GPT4 for summarization and task extraction, CLEAR-Command manages and organizes critical radio communication data in real-time. Our evaluation, comprising an expert pre-study and a user study, shows that CLEAR-Command outperforms traditional radio communication methods in terms of clarity, trust, and correctness.

Broader Impact

The deployment of CLEAR Command promises substantial improvements in emergency response efficacy by streamlining communication and task management. Automating the transcription and summarization of emergency communications enables responders to receive clear, concise, and actionable information swiftly, reducing cognitive load and minimizing errors in high-pressure environments. This system not only enhances immediate operational effectiveness but also contributes to longer-term improvements in response strategies and training programs. By refining communication processes, CLEAR Command can lead to more coordinated and timely interventions, ultimately saving lives and resources. Furthermore, the technologies and methodologies developed for this system can be adapted to other critical fields such as disaster management, military operations, and large-scale event coordination, demonstrating its potential to broadly enhance situational awareness and operational efficiency across various high-stakes domains. This broader applicability underscores CLEAR Command's role in fostering a safer and more resilient society.

Ethics Statement

The primary objective of this research is to enhance emergency response capabilities through the use of Natural Language Processing technologies. By leveraging Large Language Models, our system, CLEAR Command, aims to improve communication clarity, task identification accuracy, and operational efficiency in critical emergency scenarios. We are committed to the ethical use of LLMs and recognize the importance of balancing technological advancements with societal benefits. Our work focuses on using these models to support first responders, thereby potentially saving lives and reducing harm during emergencies. All participants in our studies provided informed consent, and their data was handled with strict confidentiality to ensure privacy and anonymity. Both studies were approaved by the ethical commity of our university. While we acknowledge the significant computational resources required for LLMs, which contribute to environmental impacts, our approach employs zero-shot to minimize energy consumption. Our intention is to harness the power of LLMs responsibly, aiming to create positive, real-world impacts in emergency management and response.

References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

Chenran Cai, Qianlong Wang, Bin Liang, Bing Qin, Min Yang, Kam-Fai Wong, and Ruifeng Xu. 2023. In-context learning for few-shot multimodal named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2969–2979.

Małgorzata Chałupnik and Sarah Atkins. 2020. "everyone happy with what their role is?": A pragmalinguistic evaluation of leadership practices in emergency medicine training. *Journal of Pragmatics*, 160:80–96.

Samer Chehade, Nada Matta, Jean-baptiste Pothin, and Rémi Cogranne. 2020. Handling effective communication to support awareness in rescue operations. *Journal of Contingencies and Crisis Management*, 28(3):307–323.

Laura Chiticariu, Yunyao Li, and Frederick Reiss. 2013. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 827–832.

Changwoo Chun, Songeun Lee, Jaehyung Seo, and Heui-Seok Lim. 2023. Cretihc: Designing causal reasoning tasks about temporal interventions and hallucinated confoundings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10334–10343.

Xiaodong Cui and Yifan Gong. 2007. A study of variable-parameter gaussian mixture hidden markov modeling for noisy speech recognition. *IEEE transactions on audio, speech, and language processing*, 15(4):1366–1376.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5884–5888. IEEE.

Dimitrios Doumanas, Andreas Soularidis, Konstantinos Kotis, and George Vouros. 2024. Integrating llms in the engineering of a sar ontology. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 360–374. Springer.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1602–1613.

Elliot Gestrin, Marco Kuhlmann, and Jendrik Seipp. 2024. Nl2plan: Robust llm-driven planning from minimal text descriptions. *arXiv preprint arXiv:2405.04215*.

Fábio Bif Goularte, Silvia Modesto Nassar, Renato Fileto, and Horacio Saggion. 2019. A text summarization method based on fuzzy rules and applicable to automated assessment. *Expert Systems with Applications*, 115:264–275.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Jiuzhou Han, Wray Buntine, and Ehsan Shareghi. 2024. Towards uncertainty-aware language agent. *arXiv preprint arXiv:2401.14016*.

Takaaki Hori, Jaejin Cho, and Shinji Watanabe. 2018. End-to-end speech recognition with word-based rnn language models. In *2018 IEEE spoken language technology workshop (SLT)*, pages 389–396. IEEE.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Sehoon Kim, Amir Gholami, Albert Shaw, Nicholas Lee, Karttikeya Mangalam, Jitendra Malik, Michael W Mahoney, and Kurt Keutzer. 2022. Squeezeformer: An efficient transformer for automatic speech recognition. *Advances in Neural Information Processing Systems*, 35:9361–9373.

James R Kirk, Robert E Wray, Peter Lindes, and John E Laird. 2024. Improving knowledge extraction from llms for task learning through agent analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18390–18398.

Yunyao Li, Frederick Reiss, and Laura Chiticariu. 2011. Systemt: A declarative information extraction system. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 109–114.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1641–1651.

Jessica Menold, Lydia Weizler, Yan Liu, Sven G Bilén, and Scarlett Miller. 2015. Identifying end-user requirements for communication systems in disadvantaged environments. In *2015 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 284–291. IEEE.

Nishanth Nakshatri, Siyi Liu, Sihao Chen, Dan Roth, Dan Goldwasser, and Daniel Hopkins. 2023. Using llm for improving key event discovery: Temporal-guided news stream clustering with event summaries. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4162–4173.

Hakan T Otal and M Abdullah Canbaz. 2024. Llm-assisted crisis management: Building advanced llm platforms for effective emergency response and public collaboration. *arXiv preprint arXiv:2402.10908*.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Christopher Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9308–9319.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. 2023. Conformal language modeling. *arXiv preprint arXiv:2306.10193*.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *The Eleventh International Conference on Learning Representations*.

Lin Ren, Yongbin Liu, Yixin Cao, and Chunping Ouyang. 2023. Covariance-based causal debiasing for entity and relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2627–2640.

George Saon, Zoltán Tüske, Daniel Bolanos, and Brian Kingsbury. 2021. Advancing rnn transducer technology for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5654–5658. IEEE.

Amina Saoutal, Jean-Pierre Cahier, and Nada Matta. 2014. Modelling the communication between emergency actors in crisis management. In *2014 International Conference on Collaboration Technologies and Systems (CTS)*, pages 545–552. IEEE.

Jianbin Shen, Junyu Xuan, and Christy Liang. 2023. Mitigating intrinsic named entity-related hallucinations of abstractive text summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15807–15824.

Ekundayo Shittu, Geoffrey Parker, and Nancy Mock. 2018. Improving communication resilience for effective disaster relief operations. *Environment Systems and Decisions*, 38:379–397.

Hagen Soltau, Hank Liao, and Hasim Sak. 2016. Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition. *arXiv preprint arXiv:1610.09975*.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009.

Jizhi Tang, Yansong Feng, and Dongyan Zhao. 2020. Understanding procedural text using interactive entity networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7281–7290.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, Bonan Min, and Thien Nguyen. 2022. Document-level event argument extraction via optimal transport. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1648–1658.

Ye Wang, Xiaojun Wan, and Zhiping Cai. 2022. Guiding abstractive dialogue summarization with content planning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3408–3413.

Christian Willms, Constantin Houy, Jana-Rebecca Rehse, Peter Fettke, and Ivana Kruijff-Korbayová. 2019. Team communication processing and process analytics for supporting robot-assisted emergency response. In *2019 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 216–221. IEEE.

Yiwen Zhang and Xuanmin Lu. 2018. A speech recognition acoustic model based on lstm-ctc. In *2018 IEEE 18th International Conference on Communication Technology (ICCT)*, pages 1052–1055. IEEE.