Preprints.org

Article

Optimizing AI-Driven Disaster Management through LLMs

Wei Chen * and Jiing Fang

Article

Optimizing AI-Driven Disaster Management through LLMs

Jiing Fang and Wei Chen *

Henan University of Technology

*   Correspondence: 1606081059@stu.sqxy.edu.cn

Abstract: Disaster issue consultation is crucial for effective emergency management, requiring timely and accurate information to support decision-making processes. This paper explores the integration of large language models (LLMs) with advanced prompt engineering to enhance the performance of AI systems in disaster management. We propose a novel multi-round prompt engineering approach that guides LLMs to iteratively refine their responses based on evolving disaster scenarios. Our method leverages diverse disaster-related datasets and employs a combination of accuracy metrics and a customized GPT-4 scoring system for evaluation. We conduct comprehensive experiments across multiple LLM platforms, including Qwen2, ChatGPT, Claude, and GPT-4, comparing our approach with baseline models and the Tree of Thoughts (ToT) method. The results demonstrate significant improvements in both accuracy and response quality, highlighting the effectiveness of our method in providing contextually relevant and actionable insights. Additional analysis confirms the robustness and generalizability of our approach across various disaster scenarios, establishing it as a valuable tool for enhancing disaster management practices.

Keywords: large language models; natural language processing

---

1. Introduction

Disaster issue consultation is a critical aspect of modern emergency management, encompassing the collection, analysis, and dissemination of information necessary for effective disaster response and recovery. The ability to provide timely and accurate advice during disasters can significantly impact the effectiveness of relief efforts, potentially saving lives and reducing economic losses. The integration of artificial intelligence (AI) in this domain has shown promise, particularly in processing large volumes of data and generating actionable insights [1–4]. However, there are significant challenges that need to be addressed to fully leverage AI's potential in disaster issue consultation.

One of the primary challenges in disaster management is the dynamic and unpredictable nature of disasters, which requires AI systems to adapt quickly to new information and evolving scenarios. Traditional AI models often struggle with integrating diverse data sources and providing contextually relevant responses. Additionally, the complexity of disaster scenarios demands a nuanced understanding that goes beyond simple data processing, necessitating advanced capabilities in natural language understanding and generation [5–8]. The motivation for our research is to enhance the effectiveness of AI in disaster issue consultation by harnessing the capabilities of large language models (LLMs) like GPT-4, combined with sophisticated prompt engineering techniques.

Our approach focuses on leveraging LLMs to address the aforementioned challenges by designing specialized prompts that guide these models to perform specific disaster management tasks efficiently. Prompts are crafted to extract critical information, generate concise summaries, and offer actionable recommendations. For instance, a prompt for assessing real-time damage might be: "Based on the following satellite images and sensor data, summarize the extent of building damage and suggest priority areas for rescue operations." This tailored prompting enables the LLM to provide more accurate and relevant responses, thereby improving decision-making processes during disasters [2].

To validate our approach, we collected a comprehensive dataset encompassing various disaster scenarios, including natural calamities, infrastructure damage reports, and resource allocation needs. We then applied our prompt engineering techniques to this dataset and evaluated the performance of

GPT-4 in generating useful outputs. The results of these evaluations demonstrated that our method significantly enhances the LLM's ability to provide context-aware and actionable insights in disaster scenarios [2,6].

- We propose a novel approach that leverages large language models combined with advanced prompt engineering to enhance disaster issue consultation.
- We develop and refine specialized prompts to guide LLMs in performing specific disaster management tasks, leading to more accurate and relevant responses.
- We validate our approach with a comprehensive dataset and demonstrate the effectiveness of our method in improving AI-supported disaster management through detailed evaluations of GPT-4's performance.

2. Related Work

In this section, we review the relevant literature in two primary areas: Large Language Models (LLMs) and Disaster Issue Consultation. We summarize the key contributions and highlight the advancements in each field that inform our research.

2.1. Large Language Models

Large Language Models (LLMs) have seen significant advancements in recent years, leading to breakthroughs in natural language processing and understanding. These models, such as GPT-3, GPT-4, and PaLM, are characterized by their extensive parameter sizes and training on vast amounts of data, which enable them to perform a wide range of language tasks with high proficiency.

Several surveys provide comprehensive overviews of LLMs, discussing their architectures, training methods, and applications. For instance, [9,10] outline the evolution of LLMs and their impact on AI research, noting how models like GPT-4 are pushing the boundaries of artificial general intelligence. The ability of LLMs to understand and generate human-like text has revolutionized fields such as NLP, information retrieval, and even computer vision through multimodal models [11–14].

Evaluation of LLMs encompasses various aspects, including their knowledge, reasoning, and ethical considerations. Studies like [15,16] categorize evaluation metrics and discuss the importance of transparency and safety in deploying LLMs. Additionally, [17] focuses on aligning LLM outputs with human values to mitigate risks such as bias and undesirable content.

2.2. Disaster Issue Consultation

Disaster issue consultation involves the use of technology to improve the management and response to disasters. This includes real-time data collection, analysis, and dissemination to support decision-making during emergencies. The integration of AI and other advanced technologies has shown promise in enhancing disaster management capabilities.

Blockchain technology has been proposed to improve the efficiency and resilience of disaster relief networks. Papers such as [4] discuss how blockchain can facilitate data propagation, peer-to-peer communication, and automated operations through smart contracts. The use of AI in disaster management includes deploying drones and communication networks to enhance situational awareness and response strategies.

Social media analytics is another area of significant interest. Reviews like [18] explore how social media data can provide real-time insights and improve coordination during disasters. Similarly, [19] discusses the role of IoT and cloud computing in disaster management, covering all phases from mitigation to recovery.

Ethical considerations are crucial in disaster management, especially when analyzing personal data shared on social media during crises. [20] emphasizes the need for privacy and the prevention of bias in automated systems. Furthermore, frameworks such as [21,22] provide guidelines for the responsible use of AI in sensitive areas, ensuring that technological interventions do not exacerbate vulnerabilities.

3. Method

In this section, we detail the methodology employed to leverage large language models (LLMs) for enhanced disaster issue consultation through advanced prompt engineering. Our approach revolves around the design and iterative refinement of multi-round prompts that guide the LLM to perform specific disaster management tasks effectively. The following subsections elaborate on the motivation behind multi-round prompting, the structure of prompts, their inputs and outputs, and the significance of this method.

3.1. Motivation for Multi-Round Prompting

The dynamic nature of disaster scenarios necessitates an AI system that can handle evolving information and provide contextually accurate responses. Single-round prompts often fall short in capturing the complexity and depth required for effective disaster management. Multi-round prompting addresses this by allowing the LLM to iteratively refine its responses based on additional context and feedback. This iterative process mimics human consultation, where responses are continuously refined and adjusted based on new information and evolving requirements.

3.2. Structure of Prompts

The structure of our prompts is designed to maximize the LLM's understanding and output quality. Each prompt is crafted to provide clear instructions, contextual information, and specific tasks. Prompts are divided into multiple rounds, each building upon the previous to refine the response. The initial prompt provides a broad overview and requests a preliminary analysis, while subsequent prompts introduce additional data, request further detail, and seek specific recommendations.

3.3. Prompt Inputs and Outputs

The inputs to our prompts include various forms of disaster-related data such as satellite images, sensor readings, textual reports, and social media posts. For example, an initial prompt might be:

> "Given the following satellite images and sensor data from the recent earthquake, provide an initial assessment of the affected areas, highlighting major damage zones and potential risks."

The output from this prompt would be an initial assessment report, summarizing the damage and risks. A subsequent prompt might refine this by asking:

> "Based on the initial assessment, prioritize the areas for rescue operations and recommend the optimal routes for emergency responders, considering the current infrastructure damage."

This round would produce a prioritized list of areas and suggested routes, providing actionable insights for emergency responders.

3.4. Significance of Multi-Round Prompting

The significance of multi-round prompting lies in its ability to generate more accurate and contextually relevant responses. By iteratively refining the LLM's output, we can ensure that the final recommendations are based on comprehensive and up-to-date information. This method also allows the model to handle ambiguities and incomplete data more effectively, as each round can address specific uncertainties and refine the initial findings.

3.5. Why This Approach is Effective

The effectiveness of our approach stems from its structured yet flexible nature. Multi-round prompting leverages the LLM's strengths in natural language understanding and generation, guiding it to produce detailed and accurate disaster management outputs. This method enables the model to handle complex and evolving scenarios, providing emergency responders with the critical information

they need to make informed decisions. Additionally, the iterative nature of the prompts allows for continuous improvement and adaptation, ensuring that the AI system remains relevant and effective in real-time disaster situations.

By combining the power of LLMs with advanced prompt engineering, our method represents a significant advancement in AI-supported disaster issue consultation. It enhances the ability of AI systems to provide timely, accurate, and actionable insights, ultimately improving the overall effectiveness of disaster management efforts.

4. Experiments

In this section, we present the experimental setup, dataset collection process, detailed evaluation metrics, and a comprehensive analysis of the results. We conducted a series of experiments to compare the performance of our proposed prompt engineering method with baseline LLMs and the Tree of Thoughts (ToT) method across multiple state-of-the-art language models.

4.1. Dataset Collection

To create a robust dataset for our experiments, we manually collected data from various online sources. This dataset includes:

- Satellite images and sensor data from recent disaster events such as earthquakes, floods, and hurricanes.
- Textual reports from governmental and non-governmental organizations detailing the impact and response measures.
- Social media posts that provide real-time information about ongoing disaster situations.
- Academic articles and case studies on disaster management practices.

The collected data was preprocessed to ensure consistency and relevance, and then annotated to facilitate the generation and evaluation of prompts.

4.2. Evaluation Metrics

To evaluate the performance of our method, we used two primary metrics:

- **Accuracy (ACC)**: This metric measures the correctness of the information extracted and the relevance of the recommendations provided by the LLM.
- **GPT-4 Scoring System**: This involves expert reviewers using GPT-4 to rate the quality of the responses on a scale from 1 to 10 based on criteria such as relevance, clarity, and actionability.

4.3. Experimental Setup

We implemented our method on four different LLM platforms: Qwen2, ChatGPT, Claude, and GPT-4. Each platform was tested with three different configurations:

- **Base LLM**: The language model without any prompt engineering.
- **ToT Method**: The Tree of Thoughts method, a known baseline for comparison.
- **Our Method**: The proposed multi-round prompt engineering approach.

4.4. Results

The results of our experiments are detailed in Tables 1 and 2. Our method consistently outperformed the base LLM and ToT methods across all platforms.

Table 1. Comparison of Methods Across Different LLM Platforms (Part 1).

Model	Method	Accuracy (ACC)	GPT-4 Score
Qwen2	Base LLM	72.3%	6.5
	ToT	75.1%	7.1
	Our Method	**81.4%**	**8.3**
ChatGPT	Base LLM	74.8%	6.8
	ToT	77.2%	7.4
	Our Method	**83.6%**	**8.7**

Table 2. Comparison of Methods Across Different LLM Platforms (Part 2).

Model	Method	Accuracy (ACC)	GPT-4 Score
Claude	Base LLM	70.9%	6.2
	ToT	73.5%	6.9
	Our Method	**79.8%**	**8.1**
GPT-4	Base LLM	78.2%	7.2
	ToT	80.4%	7.8
	Our Method	**85.9%**	**9.0**

4.5. Analysis and Validation

To further validate our method, we conducted a detailed analysis of the experimental results. The significant improvement in both accuracy and GPT-4 scores across all models highlights the efficacy of our prompt engineering approach. Specifically, the iterative refinement process embedded in our prompts allows the LLMs to better handle complex disaster scenarios by providing more accurate and contextually relevant information.

We also performed a qualitative analysis by examining specific examples where our method outperformed the baselines. For instance, in a scenario involving flood management, our method was able to provide a detailed analysis of the most affected areas, prioritize emergency responses, and suggest optimal routes for relief teams. This level of detail and accuracy was not observed in the outputs of the base LLM or the ToT method.

Moreover, we tested the robustness of our method by applying it to new, previously unseen disaster data. The results were consistent with our initial findings, demonstrating that our approach is not only effective but also generalizable to different types of disaster scenarios.

The additional experiments on unseen data, as shown in Table 3, reaffirm the robustness and adaptability of our method. These results indicate that our multi-round prompt engineering approach effectively improves the performance of LLMs in real-world disaster management scenarios, making it a valuable tool for emergency responders and decision-makers.

Table 3. Performance on Unseen Disaster Data.

Model	Method	Accuracy (ACC)	GPT-4 Score
Qwen2	Base LLM	70.1%	6.3
	ToT	73.0%	7.0
	Our Method	**80.5%**	**8.2**
ChatGPT	Base LLM	72.5%	6.7
	ToT	76.0%	7.3
	Our Method	**82.9%**	**8.6**
Claude	Base LLM	69.0%	6.1
	ToT	72.8%	6.8
	Our Method	**78.6%**	**8.0**
GPT-4	Base LLM	77.5%	7.1
	ToT	79.8%	7.7
	Our Method	**85.3%**	**8.9**

Overall, our experiments demonstrate that combining large language models with advanced prompt engineering significantly enhances their capability to provide timely, accurate, and actionable insights in disaster issue consultation. This approach not only improves the immediate response but also contributes to better preparedness and resilience in the face of future disasters.

5. Conclusion

In this study, we presented a novel approach to disaster issue consultation by leveraging large language models (LLMs) combined with advanced prompt engineering. Our multi-round prompt method was designed to iteratively refine the outputs of LLMs, thereby enhancing their ability to provide accurate, relevant, and actionable insights during disaster scenarios. We validated our approach through extensive experiments on diverse disaster-related datasets, using multiple LLM platforms including Qwen2, ChatGPT, Claude, and GPT-4. Our results showed a marked improvement in performance metrics such as accuracy and GPT-4 scores when compared to baseline models and the Tree of Thoughts (ToT) method. Further analysis underscored the robustness and adaptability of our method, demonstrating its effectiveness across different types of disasters. This research highlights the potential of integrating LLMs with prompt engineering to significantly improve the capabilities of AI systems in disaster management, offering valuable support for emergency responders and decision-makers. Future work will focus on expanding the dataset, refining the prompts, and exploring real-time applications to further enhance the practical utility of our approach.

References

1. Leveraging AI for Natural Disaster Management : Takeaways From The Moroccan Earthquake. *CoRR* **2023**, *abs/2311.08999*, [2311.08999]. doi:10.48550/ARXIV.2311.08999.
2. Sharma, S.; Devreaux, P.; Sree, S.; Scribner, D.; Grynovicki, J.; Grazaitis, P.J. Artificial intelligence agents for crowd simulation in an immersive environment for emergency response. The Engineering Reality of Virtual Reality 2019, Burlingame, CA, USA, 13-17 January 2019; Dolinsky, M.; McDowall, I.E., Eds. Society for Imaging Science and Technology, 2019. doi:10.2352/ISSN.2470-1173.2019.2.ERVR-176.
3. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. *arXiv preprint arXiv:2402.11574* **2024**.
4. Wang, Y.; Hu, Q.; Su, Z.; Zou, X.; Zhou, J. Blockchain-Envisioned Disaster Relief Networks: Architecture, Opportunities, and Open Issues. *CoRR* **2023**, *abs/2310.05180*, [2310.05180]. doi:10.48550/ARXIV.2310.05180.
5. Ortiz, B.; Kahn, L.H.; Bosch, M.; Bogden, P.; Pavon-Harr, V.; Savas, O.; McCulloh, I. Improving Community Resiliency and Emergency Response With Artificial Intelligence. 17th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2020, May 2020; Hughes, A.L.; McNeill, F.; Zobel, C.W., Eds. ISCRAM Digital Library, 2020, pp. 35–41.
6. Otal, H.T.; Canbaz, M.A. LLM-Assisted Crisis Management: Building Advanced LLM Platforms for Effective Emergency Response and Public Collaboration. *CoRR* **2024**, *abs/2402.10908*, [2402.10908]. doi:10.48550/ARXIV.2402.10908.
7. Zhou, Y.; Shen, T.; Geng, X.; Long, G.; Jiang, D. ClarET: Pre-training a Correlation-Aware Context-To-Event Transformer for Event-Centric Generation and Classification. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2559–2575.
8. Zhou, Y.; Geng, X.; Shen, T.; Long, G.; Jiang, D. Eventbert: A pre-trained model for event correlation reasoning. Proceedings of the ACM Web Conference 2022, 2022, pp. 850–859.
9. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *CoRR* **2023**, *abs/2307.06435*, [2307.06435]. doi:10.48550/ARXIV.2307.06435.
10. Wornow, M.; Xu, Y.; Thapa, R.; Patel, B.S.; Steinberg, E.; Fleming, S.L.; Pfeffer, M.A.; Fries, J.A.; Shah, N.H. The Shaky Foundations of Clinical Foundation Models: A Survey of Large Language Models and Foundation Models for EMRs. *CoRR* **2023**, *abs/2303.12961*, [2303.12961]. doi:10.48550/ARXIV.2303.12961.
11. Wornow, M.; Xu, Y.; Thapa, R.; Patel, B.S.; Steinberg, E.; Fleming, S.L.; Pfeffer, M.A.; Fries, J.A.; Shah, N.H. The Shaky Foundations of Clinical Foundation Models: A Survey of Large Language Models and Foundation Models for EMRs. *CoRR* **2023**, *abs/2303.12961*, [2303.12961]. doi:10.48550/ARXIV.2303.12961.

12. Zhou, Y.; Long, G. Style-aware contrastive learning for multi-style image captioning. *arXiv preprint arXiv:2301.11367* **2023**.

13. Zhou, Y.; Geng, X.; Shen, T.; Zhang, W.; Jiang, D. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 5822–5834.

14. Zhou, Y.; Geng, X.; Shen, T.; Pei, J.; Zhang, W.; Jiang, D. Modeling event-pair relations in external knowledge graphs for script reasoning. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* **2021**.

15. Guo, Z.; Jin, R.; Liu, C.; Huang, Y.; Shi, D.; Supryadi.; Yu, L.; Liu, Y.; Li, J.; Xiong, B.; Xiong, D. Evaluating Large Language Models: A Comprehensive Survey. *CoRR* **2023**, *abs/2310.19736*, [2310.19736]. doi:10.48550/ARXIV.2310.19736.

16. Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; Du, M. Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 20:1–20:38. doi:10.1145/3639372.

17. Xu, Y.; Hu, L.; Zhao, J.; Qiu, Z.; Ye, Y.; Gu, H. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. *CoRR* **2024**, *abs/2404.00929*, [2404.00929]. doi:10.48550/ARXIV.2404.00929.

18. Karimiziarani, M. Social Media Analytics in Disaster Response: A Comprehensive Review. *CoRR* **2023**, *abs/2307.04046*, [2307.04046]. doi:10.48550/ARXIV.2307.04046.

19. Gaire, R.; Sriharsha, C.; Puthal, D.; Wijaya, H.; Kim, J.; Keshari, P.; Ranjan, R.; Buyya, R.; Ghosh, R.K.; Shyamasundar, R.K.; Nepal, S. Internet of Things (IoT) and Cloud Computing Enabled Disaster Management. *CoRR* **2018**, *abs/1806.07530*, [1806.07530].

20. Kraft, A.; Usbeck, R. The Ethical Risks of Analyzing Crisis Events on Social Media with Machine Learning. Proceedings of the International Workshop on Data-driven Resilience Research 2022 co-located with Data Week Leipzig 2022 (DATAWEEK 2022), Leipzig, Germany, July 6, 2022; Arndt, N.; Gründer-Fahrer, S.; Holze, J.; Martin, M.; Tramp, S., Eds. CEUR-WS.org, 2022, Vol. 3376, *CEUR Workshop Proceedings*.

21. Wang, Q.; Hu, H.; Zhou, Y. Memorymamba: Memory-augmented state space model for defect recognition. *arXiv preprint arXiv:2405.03673* **2024**.

22. Mörch, C.; Gupta, A.; Mishara, B.L. Canada Protocol: an ethical checklist for the use of Artificial Intelligence in Suicide Prevention and Mental Health. *CoRR* **2019**, *abs/1907.07493*, [1907.07493].