Dynamic Fusion of Large Language Models for Crisis Communication

Xiaoying Song
University of North Texas
xiaoyingsong@my.unt.edu

Anirban Saha Anik
University of North Texas
anirbansahaanik@my.unt.edu

Vanessa Frías-Martínez
University of Maryland
vfrias@umd.edu

Lingzi Hong*
University of North Texas
lingzi.hong@unt.edu

ABSTRACT

People affected by crisis increasingly rely on social media platforms for real-time information and assistance. This underscores the need for a robust and reliable approach to provide accurate and timely information to affected individuals. Large Language Models (LLMs), which can understand user queries and generate responses, have the potential to act as assistants in crisis response. We explore various approaches, including instruction prompts, retrieval-augmented generation, and dynamic fusion of LLMs, to generate responses that address the information needs of affected individuals on social media platforms. Experiments demonstrate that the dynamic fusion approach produces better crisis responses across key evaluation dimensions, including professionalism, actionability, empathy, and relevance.

Keywords

crisis communication, large language models, social networks, retrieval-augmented generation, dynamic fusion

INTRODUCTION

Social networks have become a popular space for people in crisis to seek support and assistance, as they allow widespread outreach and connection with a larger audience (Bukar et al., 2022; Hong et al., 2018; X. Li et al., 2021). Although some responses provide valuable information and emotional support, others may be inaccurate or outdated to people in crisis (Jafar et al., 2023). Therefore, direct communication of credible information from NGOs and local governments that perform crisis response support is needed (Ziberi et al., 2024). However, providing accurate crisis responses to address individuals' needs presents significant challenges (Paulus et al., 2024). Due to the large volume of posts, authorities often face resource constraints and rapidly evolving situations that hinder their ability to provide accurate responses in time (Lenz & Eckhard, 2023).

Advanced AI has been widely used in crisis management, such as crisis prediction (Fatima et al., 2024), needs detection (Hong et al., 2018; Yin et al., 2024), and damage assessment (Saravi et al., 2019). New AI techniques, such as LLMs, can assist in crisis communication, such as combating misinformation (Saha & Srihari, 2024; Yue et al., 2024) and hostility (Hong et al., 2024). However, few studies investigate the use of LLMs to assist people in need.

In this study, we explore different approaches, including the instructional prompting LLMs (Instruction Prompt), Retrieval-Augmented Generation with LLMs (RAG), and the dynamic fusions of LLMs (Dynamic Fusion), for generating responses to address the information needs posted on social networks. We further evaluate the strengths and weaknesses of responses generated by these methods from four critical aspects: professionalism, actionability, empathy, and relevance.

Instruction Prompt is a common method used to generate responses, where tasks-specific guidance and constraints are provided in the prompt for generating responses with desired characteristics (Sahoo et al., 2024). However, the

*corresponding author

performance of models is usually sensitive to the framing of prompts (Mishra et al., 2022). In addition, LLMs are prone to hallucinations when no explicit knowledge is provided for the generation (Taveekitworachai et al., 2024). RAG retrieves external knowledge for generations, which can mitigate LLM hallucinations (Song et al., 2024). The reliance on complementary knowledge ensures more reliable and domain-specific responses and minimizes the need for extensive re-training of LLM (Mao et al., 2024). However, the RAG system may not generate desirable responses as the model focuses more on the retrieved documents and deviates from user instructions (Dai et al., 2024). In addition to the state-of-the-art methods, we propose a dynamic fusion method to generate responses aiming to combine the strengths from both Instruction Prompt and RAG. This approach introduces a dynamic fusion agent that combines the generations by Instruction Prompt and RAG for optimized responses.

We experiment with these three approaches to generate responses for users' information needs posted on Twitter (now X) during Hurricane Irma. The responses generated are evaluated on four important dimensions: professionalism, actionability, empathy, and relevance. A composite score aggregating the four dimensions assesses overall quality. We conduct the experiments with three open-source LLMs, including Llama, Mistral, and Qwen. The results show that the RAG generations consistently show higher empathy scores, while the dynamic fusion approach performs better in professionalism, actionability, and overall quality.

LITERATURE REVIEW

Previous studies have explored the detection of informative social media posts (Xie et al., 2021) and the identification of posts expressing the needs of affected individuals (Yu et al., 2022) in crisis. Multiple datasets have been developed, for example, HumanAID (Alam, Qazi, et al., 2021) and Crsisbench (Alam, Sajjad, et al., 2021). These datasets provide granular categorizations of humanitarian aspects, including the categorization of "requests or urgent needs." Crisisbench consolidates multiple crisis-related datasets and identifies both needs and responses in tweets. These efforts provide the foundation for identifying need-related posts for effective responses.

Additionally, researchers have explored the summarization task in crisis management involving condensing large volumes of information from various sources into concise and coherent summaries (Pereira et al., 2022). These summaries aim to provide an overview of the situation, highlight key developments, and support decision-making processes. For instance, Vitiugin and Castillo (2022) proposes a method for retrieving and summarizing crisis-relevant information from multilingual social media postings to aid emergency management. While summarization provides a macro-level understanding of crisis events by distilling essential information from vast data streams, crisis response generation operates at a micro-level, delivering tailored support to individuals.

Recent studies have explored the application of LLMs in crisis management, using these models to identify social media emergencies, improve real-time information extraction, and enhance public collaboration (Otal et al., 2024; Yin et al., 2024). Goecks and Waytowich (2023) leverage LLMs to generate actionable plans for users during emergencies. There is a notable gap in research on applying LLMs to generate personalized responses to the information needs of affected people in crisis.

Crisis response generation can be seen as a specialized application of open-domain question answering (QA), which requires additional considerations such as professionalism, actionability, and consistency across responses to ensure effective communication. Open-domain QA generation has emerged as a critical area of research (Kasai et al., 2024). Social networks enable users to post informal questions, which differ from the structured QA in Wikipedia or news outlets (Ritter et al., 2011). TweetsQA is the first large-scale social media QA dataset that includes user-generated questions and answers in tweets (Xiong et al., 2019). Unlike typical social media QA generation, crisis-related questions often express urgent information needs in crisis. Responding to such inquiries requires professional knowledge, and the response should be actionable to address their concern efficiently (Lamsal et al., 2024).

The fusion approach is a technique to integrate the strengths of multiple models or methods (Aniol et al., 2019; Sagi & Rokach, 2018). Jiang et al. (2023) design a fuser to integrate top-ranked candidates from multi LLMs, enabling the fuser to generate enhanced responses. Similarly, Pitis et al. (2023) utilize a set of prompts to generate answers and employ majority voting to select the best answers. Several studies have also explored approaches that first predict the best expert model and then choose its output as the final result (Shnitzer et al., 2023; Wang et al., 2023). Barabucci et al. (2024) has applied the fusion idea to medical domains, using scores of each unique diagnosis and aggregating across all LLMs. Existing studies only consider the static fusion of multiple responses, and few have investigated dynamic fusion strategies, which adjust the fusion method for each data point. We propose a dynamic fusion approach that can adaptively fuse the strengths of different generations to generate optimized responses.

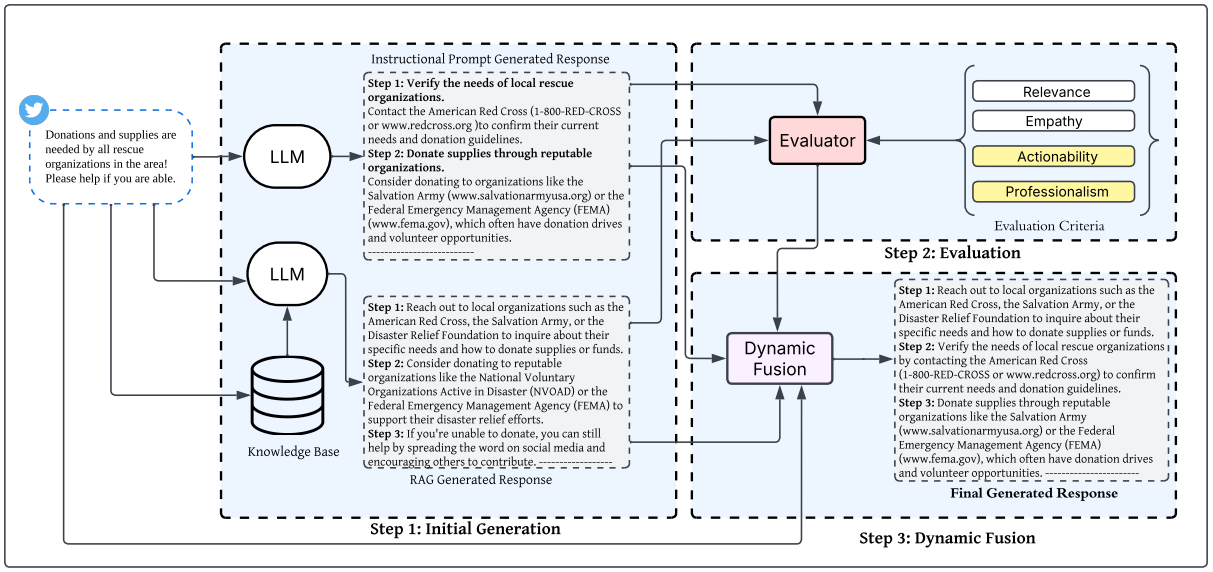

Figure 1. The overview of Dynamic Fusion.

METHODS

We first present the evaluation metrics for the AI-generated responses, as these metrics are considered in designing generation methods. Then, we introduce the baseline methods, including Instruction Prompt and RAG, and the proposed dynamic fusion method.

Evaluation

Four dimensions are used to assess the responses: professionalism, actionability, empathy, and relevance. Professionalism ensures that the response is credible, reliable and adheres to ethical standards, which is essential for maintaining public trust during a crisis. Actionability assesses whether the response offers practical guidance that individuals can follow to protect themselves or mitigate crisis impact. Empathy addresses the emotional needs of affected individuals, fostering a sense of support and understanding, which is crucial in crisis responses. Lastly, relevance ensures that the response directly addresses the specific context and unique challenges of the crisis. Together, these dimensions provide a comprehensive framework for evaluating the effectiveness of crisis communication.

Professionalism Professional responses ensure accurate, reliable, and credible assistance by leveraging knowledge and expertise to address information needs (Broekema et al., 2018; Steimle et al., 2024). We employ LLMs as evaluators (H. Li et al., 2024) to assess the professionalism of the responses generated. We design a 3-scale measurement for professionalism. Score 0 (Not Professional): The response is vague, lacks details, and does not mention specific organizations or actionable information. Score 1 (Moderately Professional): The response provides some professional elements but lacks specificity, such as mentioning general organizations without details on what they offer or how to contact them. Score 2 (Highly Professional): The response is well-structured, references specific organizations and programs, explains their relevance, and includes real contact information such as links, phone numbers, or emails. The evaluations are validated with human annotations to ensure reliability.

Actionability Actionable responses deliver clear, practical, and relevant steps or guidance to address the concern or need, which is important in crisis response (Coche et al., 2021). Specifically, there are three levels of actionability. Score 0 (Non-Actionable): The response fails to provide any practical guidance or relevant steps. Score 1 (Partially Actionable): The response provides some guidance but lacks clarity and specificity. Score 2 (Fully Actionable): The response clearly and specifically provides detailed guidance or steps that users can take immediately. This dimension is also evaluated by an LLM and validated by human annotations.

Empathy Crisis response aims to support the people suffering hard times, making it essential to convey empathy during such interactions (Schoofs et al., 2022). Empathetic responses help build trust, provide emotional support, and ensure the audience feels understood and respected in difficult situations (Bono, 2024). We utilize an empathetic dialogue dataset that categorizes empathy into three scales to build a classifier for assessing the empathy of responses (Sharma et al., 2020).

Relevance Relevance is evaluated using the BERTScore between the response and the query (Zhang et al., n.d.).

In addition, we design a metric integrating the four dimensions to evaluate the overall quality responses. Compared to relevance and empathy, professionalism and actionability are assigned with higher weights, as they ensure that responses are reliable and easy to follow (Whims, 2024). We detail the score formula as:

$$Q = 0.4 \times S(\text{Pro}) + 0.4 \times S(\text{Act}) + 0.1 \times S(\text{Emp}) + 0.1 \times S(\text{Rel})$$

This approach prioritizes professionalism and actionability while still maintaining strong relevance and empathy.

Instruction Prompt

An instructional prompt contains clear guidelines and constraints to direct an LLM in task completion. We design prompts that require the model to consider professionalism, actionability, empathy, and relevance in the responses to information need posts. By defining the scope of the task, the expected attitude, and specific contextual information, instructional prompts guide the model in producing appropriate responses to address the information needs of affected individuals. This method serves as the baseline.

RAG

RAG method includes three key components: knowledge base selection, document retrieval, and response generation. (1) Knowledge Base Construction. We refer to the authoritative resources from FEMA[1] to build our knowledge base. Given the resources $S = \{D_1, D_2, \ldots, D_N\}$ from FEMA, the resources are split into documents to create the knowledge base $K = \{d_1, d_2, \ldots, d_N\}$ for subsequent retrieval. FEMA provides official guidelines, risk mitigation strategies, crisis response protocols, and post-disaster recovery plans, ensuring accurate and up-to-date knowledge, which is well-suited for crisis response systems(Miller, 2024).

(2) Document Retrieval. The hybrid RAG integrates multiple retrieval methods and outperforms any single retrieval approach (Sawarkar et al., 2024). Our retrieval module combines keyword-based (R_k) and semantic retrieval (R_s) methods. The hybrid retriever (R_h) integrates the strengths of two methods: $R_h = R_k \cup R_s$. Specifically, R_k refers to traditional keyword-based retrieval, which employs methods like BM25 or TF-IDF to rank documents based on exact keyword matches between the query and the documents. In contrast, R_s uses semantic retrieval by generating dense vector embeddings for both the query and the documents, and then ranks them using similarity measures such as cosine similarity. The hybrid retriever R_h combines these two approaches to leverage the precision of keyword matching and the contextual understanding of semantic retrieval, resulting in improved overall document retrieval performance.

(3) Response Generation. The retrieved top-N documents ($R_h = \{d_1, d_2, \ldots, d_N\}$) are concatenated into a single context: $C = \text{concat}(d_1, d_2, \ldots, d_N)$.The concatenated context C is then paired with the input query q to construct the prompt for an LLM to generate a response r.

Dynamic Fusion

The dynamic fusion model adaptively integrates the strengths of multiple models to ensure high-quality responses across diverse user needs and contexts. The fusion process adjusts the integration based on the strengths of each model in every case and combines the best aspects of professionalism, actionability, empathy, and relevance from candidate responses for optimized results.

The dynamic fusion process includes three steps (See Figure 1): (1) *Initial Generation*: responses are generated by state-of-the-art crisis response generation methods, including the instructional prompts and RAG. (2) *Evaluation*: Responses by different methods are evaluated in four dimensions: relevance, empathy, actionability, and professionalism. (3) *Dynamic Fusion*: an LLM agent is used to integrate these responses by selectively combining the best aspects from candidate responses, for example, integrating the professionalism and empathy from one response and the actionability and relevance from another. If one response is superior in all dimensions, this response is used as the final output without further need of fusion. The process is represented in the following.

Given two initial responses, r_1 and r_2, along with their evaluation scores across four dimensions, our goal is to generate a final response r_{final}.

[1] https://www.fema.gov/

1. **Selection**: Selecting the better response if one outperforms the other across all dimensions.

$$r_{\text{final}} = \begin{cases} r_1, & \text{if } S(r_1, e) \geq S(r_2, e),\ \forall e \in \mathcal{E} \\ r_2, & \text{if } S(r_2, e) \geq S(r_1, e),\ \forall e \in \mathcal{E} \end{cases}$$

   $\mathcal{E} = \{\text{Rel}, \text{Act}, \text{Pro}, \text{Emp}\}$ represent the set of evaluation dimensions: relevance (Rel), actionability (Act), professionalism (Pro), and empathy (Emp). The performance score of response r_i on dimension $e \in \mathcal{E}$ as $S(r_i, e)$, where $i \in \{1, 2\}$.

2. **Fusion**: If there is no single response that excels in all dimensions, generate a new, fused response based on their evaluation strengths.

$$r_{\text{final}} = \text{fuse}(r_1, r_2, \{e \mid \arg\max_i S(r_i, e)\})$$

The function selects the best-performing response in each dimension by identifying which response has the highest score for that criterion. The result is a set of dimensions e where the corresponding best response contributes to the final response. The fusion process then merges the selected aspects into a new, optimized output that maintains the strongest qualities from both candidates, ensuring a well-rounded and effective final response.

EXPERIMENTS

Dataset

The dataset used in the study comprises over 1 million geotagged tweets collected from six southern U.S. states (Florida, Georgia, South Carolina, North Carolina, Tennessee, Alabama) during Hurricane Irma (August 15 – October 12, 2017) from Twitter (now X). After cleaning and filtering, 1,013,313 tweets from 127,181 users were retained, with 6,726 users meeting the criteria for evacuation flow analysis based on their tweet frequency across different hurricane phases. Among them, 5,307 were identified as evacuees, with 85.4% staying within their state. Florida had the largest sample, with 2,486 Twitter users and 2,020 evacuees, 62% of whom remained in their original counties. We use posts on information needs for experiments.

The needs-related posts are buried in many irrelevant ones. We train three RoBERTa models (Liu, 2019) to predict whether a tweet expresses needs. The RoBERTa model outperforms in identifying humanitarian tweets (Alam, Sajjad, et al., 2021). Our classifiers are trained with three crisis datasets annotated with "needs or request" and other categories (Alam, Qazi, et al., 2021; Alam, Sajjad, et al., 2021) [2]. A tweet is labeled as "needs-related" if all three classifiers predict it as such.

Additionally, we conduct human validation to verify the predictions. Two research assistants are employed to annotate crisis needs. We provide detailed guidelines in the following: Read the tweet and identify tweets where people seek help in crisis, such as food, medical supplies, and emotional support. Label the tweet as 1 if it demonstrates a need, and 0 if it does not. Examples are also provided to annotators for guidance. For instance, tweets like *"We need tents, water, food, lanterns, medicine. In Peguy Ville..." or "My dog is hurt, is there any help around?..."* would be labeled as 1. The agreement rate between two annotators is 94.5%, with a Cohen's Kappa of 0.87. The agreement rate between classifiers and humans is 95%, with a Kappa of 0.79, indicating the predictions are reliable.

LLM Setup

We experiment with several LLMs, including **Llama-3.1-8B-Instruct**, **Mistral-8B-Instruct-2410**, and **Qwen2.5-7B-Instruct**, which are good at conversational chatting. The *pipeline* function from the *transformers* library is used to load the models. The maximum token length (*max_new_tokens=256*) ensures the generated response remains concise while allowing enough space for meaningful content. The temperature (*temperature=0.6*) is set to balance creativity and coherence. The nucleus sampling parameter (*top_p=0.9*) filters out unlikely tokens while retaining natural variation in the generated responses. All the experiments run on a server with an Intel Xeon Gold 6226R processor, 128 GB memory, and 3 Nvidia RTX 8000 graphic cards.

[2] https://www.kaggle.com/datasets/ulkutuncerkucuktas/turkey-earthquake-relief-tweets-dataset

Crisis Response Generations

Instruction Prompt Needs-related tweets are considered as the input to feed into LLMs to obtain response predictions. The prompt for instructing LLMs to generate the responses is:

```
role: system
content: You are an AI assistant designed to provide professional, actionable, empathetic and
    relevant advice for someone seeking help related to a hurricane on social media.

role: user
content: Given the following tweet expressing needs during a hurricane, provide a detailed
    solution. If you don't know the answer, clearly state, 'I don't know'.

Guidelines:
- Prioritize immediate actions, clearly labeled as **Step 1**, **Step 2**, etc.
- For each action, provide a brief follow-up sentence to explain its importance or how to
    implement it.
- Include links, organizations, or contact information where relevant.
- Response should be professional, actionable, empathetic and relevant.
```

RAG We select two resources from U.S. Federal Emergency Management Agency (FEMA): one is *Individual Assistance Program and Policy Guide*, which provides accessible programs and policies designed to support individuals during disaster [3]. Another is *A Citizen's Guide to Disaster Assistance*[4], which documents comprehensive guidance and resources to assist citizens during crises.

```
role: system
content: You are an AI assistant designed to provide professional, actionable, empathetic and
    relevant advice for someone seeking help related to a hurricane on social media. Use the
    provided documents to address the needs expressed in the tweet.

role: user
content: Given the following tweet expressing needs during a hurricane, provide a detailed
    solution. If you don't know the answer, clearly state, 'I don't know'.

Guidelines:
- Prioritize immediate actions, clearly labeled as **Step 1**, **Step 2**, etc.
- For each action, provide a brief follow-up sentence to explain its importance or how to
    implement it.
- Include links, organizations, or contact information where relevant.
- Response should be professional, actionable, empathetic and relevant.
```

We use a hybrid search method incorporating keyword-based and semantic retrieval in RAG. The embedding model all-mpnet-base-v2 is used for semantic retrieval. We implement a custom hybrid retriever and set the mode parameter as "OR" to maximize the chances of retrieving relevant information. We select the top-5 retrieved documents and concatenate them into a single context, providing additional knowledge for LLMs. The top-5 selection ensures the most relevant context is included while minimizing redundant information. The combined context and the full prompt are fed into the LLMs to generate responses.

Dynamic Fusion This method aims to either select the better response or generate a fused response that integrates the strengths of candidate responses. If one response performs better across all dimensions, it will be selected directly as the final output. Otherwise, LLMs will create a fused response. The prompt we use is as follows.

```
You are an AI assistant designed to choose the better response or create an infused response
    by analyzing evaluation scores. Consider the strengths of each response based on
    professionalism, actionability, relevance, and empathy.
Response 1:
{response1}

Scores: {scores1}

Response 2:
{response2}

Scores: {scores2}

Generate the best response based on the provided details.
```

[3] https://www.fema.gov/sites/default/files/documents/fema_iappg-1.1.pdf
[4] https://training.fema.gov/emiweb/downloads/is7complete.pdf

RESULTS

We conduct evaluations using the aforementioned metrics for responses generated by different approaches. To validate the evaluations of professionalism and actionability by LLMs, we engage human annotators to view the response and manually annotate based on the 3-scale definitions. We randomly sample 100 tweets and response pairs for annotations. The agreement rates between two annotations are above 85% with Cohen's Kappa ($\kappa \geq 0.80$), indicating the human annotation is reliable. An expert assigns the final label for the human annotation, which will be used to compare with model evaluation. The agreement rate and Cohen's Kappa ($\kappa \geq 0.72$) between human evaluation and model evaluation demonstrates substantial agreement.

Figure 2 presents the comparison of generated responses by different approaches and LLMs. RAG responses are more actionable, professional, and empathic but less relevant than Instruction Prompt responses when Llama model is used. When using Qwen and Ministral models, RAG responses are more empathetic and relevant but less actionable and professional. Neither method outperforms the other across all dimensions. This justifies the use of the dynamic fusion approach.

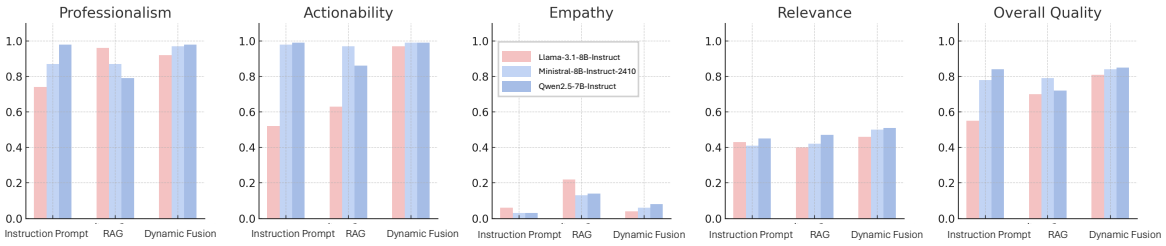

Figure 2. Performance comparison of different models and methods.

Notably, RAG responses are consistently more empathetic than Instruction Prompt responses. This is probably because they leverage the knowledge from the RAG knowledge base, which emphasizes humanitarian and empathetic communication. When different LLMs are used, the dynamic fusion approach can generate responses with higher professionalism and actionability. Considering the overall quality, the dynamic fusion approach is the best when using Llama and Ministral.

We further present the detailed evaluation results in Table 1 with the mean scores across all samples. Dynamic Fusion consistently outperforms baseline methods in overall response quality, achieving an overall score of 0.81 with Llama-3.1-8B-Instruct, 0.84 with Mistral-8B-Instruct-2410, and 0.85 with Qwen2.5-7B-Instruct. Specifically, Dynamic Fusion with Llama-3.1-8B-Instruct excels in actionability (0.97), relevance (0.46), and a comparably high score in professionalism (0.92), but it shows slightly lower performance in empathy. Dynamic Fusion with Mistral-8B-Instruct-2410 achieves higher scores in professionalism (0.97), actionability (0.99), and relevance (0.50), despite a lower empathy score (0.06). Nevertheless, it still attains a significantly higher overall quality score, demonstrating its effectiveness in generating crisis responses. Dynamic Fusion, when applied to Qwen2.5-7B-Instruct, generates the highest quality crisis response, achieving scores of 0.98, 0.99, and 0.51 in professionalism, actionability, and relevance. The experiments with different LLMs consistently demonstrate the superiority of the dynamic fusion method.

CONCLUSION

We explore different approaches using LLMs to generate responses to address the needs of affected individuals in crisis. An innovative dynamic fusion approach is proposed, which can integrate the strengths of Instruction Prompt and RAG for a better generation evaluated by professionalism, actionability, empathy, and relevance. Our initial generation reveals that neither Instruction Prompt nor RAG consistently outperforms the other across all dimensions. Fusing two responses is essential. The dynamic fusion approach significantly improves response quality compared to Instruction Prompt and RAG, especially in professionalism and actionability.

The current dynamic fusion method primarily relies on the automatic integration done by an LLM. More sophisticated fusion methods will be developed for better results. We experiment with the setting of one integration with results from Instruction Prompt and RAG. Future research will look into the iterative integration for optimized results.

We experiment with generation methods in one crisis scenario of Hurricane Irma. However, each crisis situation may present unique nuances that can affect the quality and tone of the response. In the future, we will experiment methods with datasets from different crisis scenarios to understand the generalizability of the proposed method. We

Table 1. Performance comparison of different models and methods on multiple evaluation criteria.

Model	Method	Professionalism	Actionability	Empathy	Relevance	Overall Quality
Llama	Instruction Prompt	0.74	0.52	0.06	0.43	0.55
	RAG	**0.96**	0.63	**0.22**	0.40	0.70
	Dynamic Fusion	0.92	**0.97**	0.04	**0.46**	**0.81**
Mistral	Instruction Prompt	0.87	0.98	0.03	0.41	0.78
	RAG	0.87	0.97	**0.13**	0.42	0.79
	Dynamic Fusion	**0.97**	**0.99**	0.06	**0.50**	**0.84**
Qwen	Instruction Prompt	0.98	0.99	0.03	0.45	0.84
	RAG	0.79	0.86	**0.14**	0.47	0.72
	Dynamic Fusion	**0.98**	**0.99**	0.08	**0.51**	**0.85**

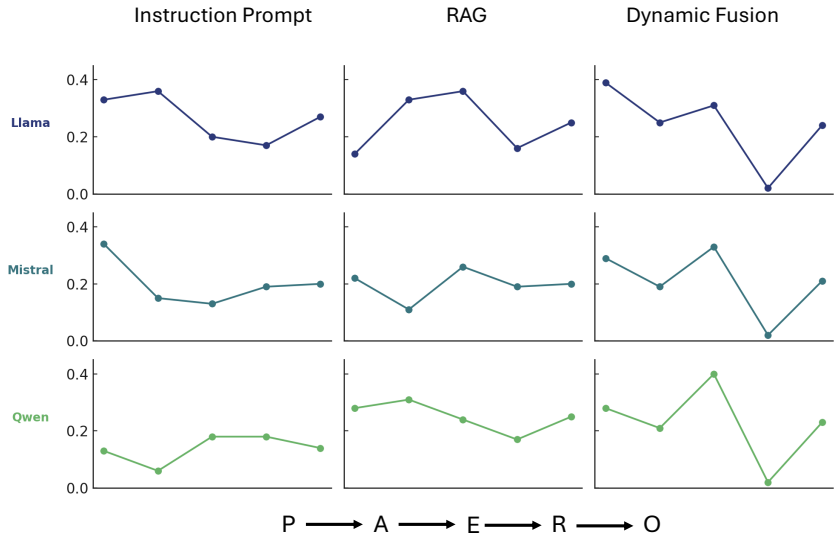

Figure 3. Performance variance comparison of different models and methods across multiple evaluation metrics. P: professionalism, A: actionability, E: empathy, R: relevance, O: overall quality

also find that balancing conflicting objectives, such as empathy versus actionability, is challenging. We will explore alternative fusion methods to explore their performance in the overall quality of responses across different crisis contexts, ultimately enhancing the robustness and adaptability of crisis response models.

ACKNOWLEDGMENTS

This work was supported by the Institute of Museum and Library Services (IMLS) National Leadership Grants under LG256661-OLS-24 and LG-256666-OLS-24.

REFERENCES

Alam, F., Qazi, U., Imran, M., & Ofli, F. (2021). Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks. *Proceedings of the International AAAI Conference on Web and social media*, *15*, 933–942.

Alam, F., Sajjad, H., Imran, M., & Ofli, F. (2021). Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. *Proceedings of the International AAAI conference on web and social media*, *15*, 923–932.

Aniol, A., Pietron, M., & Duda, J. (2019). Ensemble approach for natural language question answering problem. *2019 Seventh International Symposium on Computing and Networking Workshops (CANDARW)*, 180–183.

Barabucci, G., Shia, V., Chu, E., Harack, B., & Fu, N. (2024). Combining insights from multiple large language models improves diagnostic accuracy. *arXiv e-prints*, arXiv–2402.

Bono, O. (2024). Effectiveness of crisis communication strategies on public trust in chad. *American Journal of Public Relations*, *3*(1), 36–45.

Broekema, W., van Eijk, C., & Torenvlied, R. (2018). The role of external experts in crisis situations: A research synthesis of 114 post-crisis evaluation reports in the netherlands. *International journal of disaster risk reduction*, *31*, 20–29.

Bukar, U. A., Jabar, M. A., Sidi, F., Nor, R. B., Abdullah, S., & Ishak, I. (2022). How social media crisis response and social interaction is helping people recover from covid-19: An empirical investigation. *Journal of computational social science*, 1–29.

Coche, J., Kropczynski, J., Montarnal, A., Tapia, A., & Benaben, F. (2021). Actionability in a situation awareness world: Implications for social media processing system design. *ISCRAM 2021-18th International conference on Information Systems for Crisis Response and Management*, (2391), p–994.

Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., & Xu, J. (2024). Bias and unfairness in information retrieval systems: New challenges in the llm era. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6437–6447.

Fatima, K., Shareef, H., Costa, F. B., Bajwa, A. A., & Wong, L. A. (2024). Machine learning for power outage prediction during hurricanes: An extensive review. *Engineering Applications of Artificial Intelligence*, *133*, 108056.

Goecks, V. G., & Waytowich, N. R. (2023). Disasterresponsegpt: Large language models for accelerated plan of action development in disaster response scenarios. *arXiv preprint arXiv:2306.17271*.

Hong, L., Fu, C., Wu, J., & Frias-Martinez, V. (2018). Information needs and communication gaps between citizens and local governments online during natural disasters. *Information Systems Frontiers*, *20*, 1027–1039.

Hong, L., Luo, P., Blanco, E., & Song, X. (2024). Outcome-constrained large language models for countering hate speech. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4523–4536.

Jafar, Z., Quick, J. D., Larson, H. J., Venegas-Vera, V., Napoli, P., Musuka, G., Dzinamarira, T., Meena, K. S., Kanmani, T. R., & Rimányi, E. (2023). Social media for public health: Reaping the benefits, mitigating the harms. *Health promotion perspectives*, *13*(2), 105.

Jiang, D., Ren, X., & Lin, B. Y. (2023). Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14165–14178.

Kasai, J., Sakaguchi, K., Le Bras, R., Asai, A., Yu, X., Radev, D., Smith, N. A., Choi, Y., Inui, K., et al. (2024). Realtime qa: What's the answer right now? *Advances in Neural Information Processing Systems*, *36*.

Lamsal, R., Read, M., Karunasekera, S., & Imran, M. (2024). Crema: Crisis response through computational identification and matching of cross-lingual requests and offers shared on social media. *IEEE Transactions on Computational Social Systems*.

Lenz, A., & Eckhard, S. (2023). Conceptualizing and explaining flexibility in administrative crisis management: A cross-district analysis in germany. *Journal of Public Administration Research and Theory*, *33*(3), 485–497.

Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., & Liu, Y. (2024). Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Li, X., Bahursettiwar, A., & Kogan, M. (2021). Hello? is there anybody in there? analysis of factors promoting response from authoritative sources in crisis. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21.

Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, *364*.

Mao, K., Liu, Z., Qian, H., Mo, F., Deng, C., & Dou, Z. (2024). Rag-studio: Towards in-domain adaptation of retrieval augmented generation through self-alignment. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 725–735.

Miller, S. (2024). Emerging technology in emergency management: Using generative artificial intelligence (ai) to improve emergency messaging. *URL: https://doi. org/10.13140/RG*, *2*(21739.25124).

Mishra, S., Khashabi, D., Baral, C., & Hajishirzi, H. (2022). Cross-task generalization via natural language crowdsourcing instructions. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3470–3487.

Otal, H. T., Stern, E., & Canbaz, M. A. (2024). Llm-assisted crisis management: Building advanced llm platforms for effective emergency response and public collaboration. *2024 IEEE Conference on Artificial Intelligence (CAI)*, 851–859.

Paulus, D., Fathi, R., Fiedrich, F., de Walle, B. V., & Comes, T. (2024). On the interplay of data and cognitive bias in crisis information management: An exploratory study on epidemic response. *Information Systems Frontiers*, *26*(2), 391–415.

Pereira, J. A., do Nascimento Fidalgo, R., de Alencar Lotufo, R., & Nogueira, R. F. (2022). Using neural reranking and gpt-3 for social media disaster content summarization. *TREC*.

Pitis, S., Zhang, M. R., Wang, A., & Ba, J. (2023). Boosted prompt ensembles for large language models. *arXiv preprint arXiv:2304.05970*.

Ritter, A., Cherry, C., & Dolan, B. (2011). Data-driven response generation in social media. *Empirical Methods in Natural Language Processing (EMNLP)*.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *8*(4), e1249.

Saha, S., & Srihari, R. K. (2024). Integrating argumentation and hate-speech-based techniques for countering misinformation. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11109–11124.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Saravi, S., Kalawsky, R., Joannou, D., Rivas Casado, M., Fu, G., & Meng, F. (2019). Use of artificial intelligence to improve resilience and preparedness against adverse flood events. *Water*, *11*(5), 973.

Sawarkar, K., Mangal, A., & Solanki, S. R. (2024). Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. *arXiv preprint arXiv:2404.07220*.

Schoofs, L., Fannes, G., & Claeys, A.-S. (2022). Empathy as a main ingredient of impactful crisis communication: The perspectives of crisis communication practitioners. *Public Relations Review*, *48*(1), 102150.

Sharma, A., Miner, A., Atkins, D., & Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5263–5276.

Shnitzer, T., Ou, A., Silva, M., Soule, K., Sun, Y., Solomon, J., Thompson, N., & Yurochkin, M. (2023). Large language model routing with benchmark datasets. *Annual Conference on Neural Information Processing Systems*.

Song, J., Wang, X., Zhu, J., Wu, Y., Cheng, X., Zhong, R., & Niu, C. (2024). Rag-hat: A hallucination-aware tuning pipeline for llm in retrieval-augmented generation. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1548–1558.

Steimle, L., von Peter, S., & Frank, F. (2024). Professional relationships during crisis interventions: A scoping review. *Plos one*, *19*(2), e0298726.

Taveekitworachai, P., Abdullah, F., & Thawonmas, R. (2024). Null-shot prompting: Rethinking prompting large language models with hallucination. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13321–13361.

Vitiugin, F., & Castillo, C. (2022). Cross-lingual query-based summarization of crisis-related social media: An abstractive approach using transformers. *Proceedings of the 33rd ACM conference on hypertext and social media*, 21–31.

Wang, H., Polo, F. M., Sun, Y., Kundu, S., Xing, E., & Yurochkin, M. (2023). Fusing models with complementary expertise. *Annual Conference on Neural Information Processing Systems*.

Whims, T. (2024). Ai at the helm: Transforming crisis communication through theory and advancing technology.

Xie, Z., Jayanth, A., Yadav, K., Ye, G., & Hong, L. (2021). Multi-faceted classification for the identification of informative communications during crises: Case of covid-19. *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 924–933.

Xiong, W., Wu, J., Wang, H., Kulkarni, V., Yu, M., Chang, S., Guo, X., & Wang, W. Y. (2019). Tweetqa: A social media focused question answering dataset. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5020–5031.

Yin, K., Liu, C., Mostafavi, A., & Hu, X. (2024). Crisissense-llm: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics. *arXiv preprint arXiv:2406.15477*.

Yu, X., Xie, Z., Mashhadi, A., & Hong, L. (2022). Multi-task models for multi-faceted classification of pandemic information on social media. *Proceedings of the 14th ACM Web Science Conference 2022*, 327–335.

Yue, Z., Zeng, H., Lu, Y., Shang, L., Zhang, Y., & Wang, D. (2024). Evidence-driven retrieval augmented response generation for online misinformation. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5628–5643.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (n.d.). Bertscore: Evaluating text generation with bert. *International Conference on Learning Representations*.

Ziberi, L., Lengel, L., Limani, A., & Newsom, V. A. (2024). Affect, credibility, and solidarity: Strategic narratives of ngos' relief and advocacy efforts for gaza. *Online Media and Global Communication*, *3*(1), 27–54.