A Dynamic Fusion Model for Consistent Crisis Response

Xiaoying Song[1] Anirban Saha Anik[1] Eduardo Blanco[2]
Vanessa Frias-Martinez[3] Lingzi Hong[1]
[1] University of North Texas
[2] University of Arizona [3] University of Maryland
{xiaoyingsong, anirbansahaanik}@my.unt.edu
eduardoblanco@arizona.edu, vfrias@umd.edu, lingzi.hong@unt.edu

Abstract

In response to the urgent need for effective communication with crisis-affected populations, automated responses driven by language models have been proposed to assist in crisis communications. A critical yet often overlooked factor is the consistency of response style, which could affect the trust of affected individuals in responders. Despite its importance, few studies have explored methods for maintaining stylistic consistency across generated responses. To address this gap, we propose a novel metric for evaluating style consistency and introduce a fusion-based generation approach grounded in this metric. Our method employs a two-stage process: it first assesses the style of candidate responses and then optimizes and integrates them at the instance level through a fusion process. This enables the generation of high-quality responses while significantly reducing stylistic variation between instances. Experimental results across multiple datasets demonstrate that our approach consistently outperforms baselines in both response quality and stylistic uniformity.

1 Introduction

People in crisis often turn to social networks for information, support, and assistance, especially when other sources cannot be relied upon (Bukar et al., 2022). Although some responses in social media from the general public offer valuable information and emotional support, others may be inaccurate and even misleading to those in crisis (Jafar et al., 2023). For example, during Hurricane Irma, users on Twitter (now X) shared conflicting information about whether shelters required identity checks, which affected whether some immigrants decided to evacuate (Hunt et al., 2022).

Direct communication from relevant government agencies or NGOs that carry out disaster relief efforts is critical to providing accurate information and verifying misleading information. However, authorities and NGOs often do not have enough resources to respond promptly to all affected individuals. At the same time, people's needs are so different that a one-size-fits-all response is rarely effective (Paulus et al., 2024; Lenz and Eckhard, 2023). This challenge can be mitigated using LLM-based chat engines to understand natural conversations and generate informed responses (Song et al., 2025a). Leveraging AI to improve the efficiency, scalability, and accuracy of crisis communication has become a critical research focus (Ziberi et al., 2024).

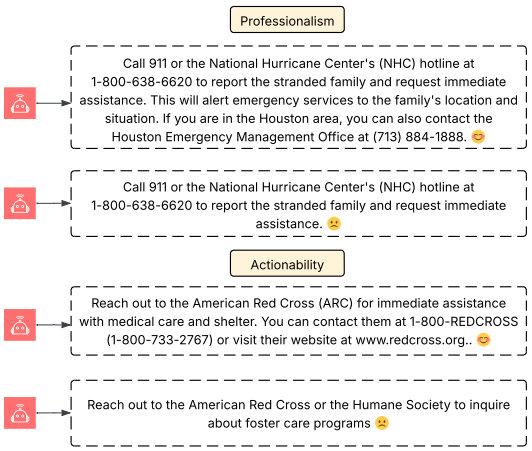

Figure 1: Examples of responses with high and low professionalism and actionability. Professional responses include explanations backing recommendations, demonstrating authority. Actionable responses offer specific guidance (e.g., phone numbers, website links) that users can follow to seek help. In this paper, we focus on generating *consistent* responses, i.e., ensuring that professionalism, actionability, and relevance are roughly the same across all responses.

Recent studies have explored the potential role of LLMs in supporting crisis communication (Hong et al., 2025; Xiao and Yu, 2025; Otal et al., 2024; Grigorev et al., 2024). These systems aim to provide actionable, real-time guidance to affected individuals, focusing on user satisfaction, responsive

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2753–2768
November 4-9, 2025

interaction, and efficient use of resources (Lei et al., 2025). However, an important issue remains overlooked: the consistency of automatically generated responses.

Authorities and NGOs have shown bias in their responses to people in crisis, which leads to inequitable access to aid and distrust (van Voorst et al., 2022; Huang and Su, 2009). We define consistency as the uniformity of the style in which information is conveyed across all responses. In particular, the core information conveyed should maintain the same level of quality regardless of the audience, crisis scenario, or communication platform. Consistency signals organizational reliability. When messages remain aligned, audiences are more likely to trust the source (Correia, 2024). In contrast, inconsistent responses can be confusing and diminish trust (Chatratichart et al., 2024). For example, if some responses offer clear guidance while others are vague or off-topic, users may be uncertain about what to believe or do. Figure 1 shows examples of replies with different degrees of professionalism and actionability. When responses vary in quality across users, those receiving lower quality replies may perceive the interaction as inattentive or dismissive, resulting in dissatisfaction.

Previous studies have explored the generation of consistent responses in general-purpose dialogue systems, with particular attention to persona consistency (Lee et al., 2024), semantic consistency (Fan et al., 2025), and factual consistency (Mesgar et al., 2021). Few studies have addressed style consistency in crisis communications (Huang and Su, 2009). Additionally, these studies typically employ fine-tuned generative models to increase consistency (Lee et al., 2024; Mesgar et al., 2021).

There are no established metrics to evaluate the consistency of responses in crisis communication. Effective crisis communication requires adherence to critical communicative functions (Sellnow and Seeger, 2021; Coombs, 2007). These responses should be professional (Steimle et al., 2024; Coombs, 2007), actionable (Coche et al., 2021; Bono, 2024), and relevant to user needs. Response consistency, therefore, entails delivering messages with stable characteristics across these dimensions, regardless of user query or scenario. We propose a task-oriented definition for crisis communication: consistency refers to the degree to which all responses have similar characteristics across the three dimensions: professionalism, actionability, and relevance, while exhibiting minimal variation across responses.

In addition, we propose a fusion framework to generate crisis responses with improved consistency. The approach integrates the strengths of the responses generated by multiple methods, taking advantage of their complementary advantages to produce highly effective outputs in all evaluation dimensions, resulting in reduced variations. Our approach employs state-of-the-art generation methods and explores various fusion methods. We evaluate the generation approaches in the three critical dimensions (professionalism, actionability, and relevance) as well as consistency across these dimensions. Experiments show the fusion framework enables the generation of responses with higher overall quality and consistency. Specifically, we propose a novel fusion method grounded on assigning tailored weights to each dimension. We experiment with Llama and Mistral and demonstrate that our fusion method results in superior performance compared to alternatives.

The contributions of this study include:

- We introduce a novel crisis response evaluation metric, **Consistency**, designed to ensure uniformity across key evaluation dimensions while addressing diverse information-need queries across crisis events.
- We propose a **Fusion Framework** that generates responses by integrating the strengths of outputs from different models, achieving strong performance on key evaluation metrics while ensuring consistency.
- We conduct detailed analyses demonstrating the fusion mechanisms obtains strong performance across LLMs, crisis scenario, and other realistic scenarios.

2 Related Work

Information Needs and Responses in Crisis Individuals frequently use social media platforms to seek assistance in times of crisis. Previous studies have proposed methods for detecting and classifying user needs. Several datasets offer granular categorizations of needs (Alam et al., 2021a,b). Recent studies have proposed using LLMs to facilitate timely responses (Hong et al., 2025; Otal et al., 2024; Yin et al., 2024; Chowdhury et al., 2024). For example, Goecks and Waytowich (2023) and Otal et al. (2024) leveraged LLMs to generate actionable plans or guidance to crisis-affected individuals. Grigorev et al. (2024) developed IncidentRespon-

seGPT, which leverages LLMs to automatically generate traffic incident response plans by synthesizing guidelines and processing real-time accident reports to inform authorities. Rawat (2024) introduced DisasterQA, which is designed to evaluate LLMs in disaster response scenarios. They experimented with several prompting methods to answer crisis questions.

These prior studies investigate approaches to generating responses for crisis communication. We are the first, however, to investigate the consistency of responses, with a focus on maintaining a uniform style across varying scenarios.

Consistent Response Generation Consistent responses are essential for ensuring trust. In particular, it is important to avoid contradictions when addressing different audiences at different times, maintain a consistent tone, and ensure the conveyed information remains aligned (Lee et al., 2024).

Previous studies have explored various aspects of consistent response generation, including persona consistency, semantic consistency, and factual consistency. Persona consistency refers to the alignment between generated responses and the established persona in dialogue systems (Lee et al., 2024; Kim et al., 2023; Mesgar et al., 2021). Semantic consistency ensures the generated responses logically follow the context without introducing irrelevant (Fan et al., 2025; Song et al., 2025b). Factual consistency refers to the accuracy and correctness of generated content (Mesgar et al., 2021). While these forms of consistency are crucial in general-purpose dialogue, they do not address consistency in balancing the critical communication dimensions required for crisis response, including professionalism, actionability, and relevance. To our knowledge, no prior work has systematically defined or evaluated consistency in the context of crisis communication, highlighting a gap that our work aims to address.

3 Consistency in Crisis Communication

Consistency in crisis communication is crucial for maintaining trust and clarity. Our consistency involves producing professional, actionable, and relevant responses, as defined below. Maintaining consistency across these dimensions is essential because variation can lead to confusion, reduced trust, and even harmful outcomes.

- *Professionalism* Professional responses ensure accurate, reliable, and credible assistance by leveraging knowledge and expertise to address crisis challenges effectively (Steimle et al., 2024; Broekema et al., 2018).
- *Actionability* Actionable responses deliver clear, practical, and relevant steps or guidance to address the concern or needs. In crisis response, solutions need to be straightforward and easy to implement (Coche et al., 2021).
- *Relevance* It evaluates how closely connected or appropriate generated responses are to the requests or queries showing needs.

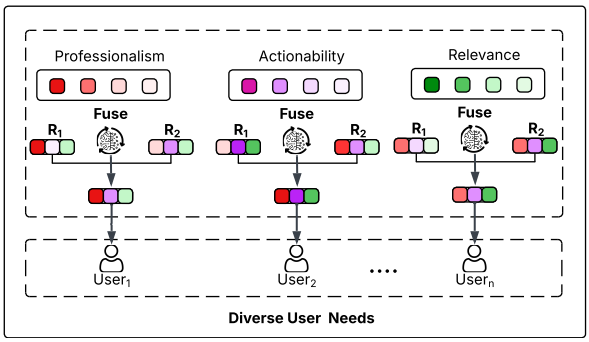

Figure 2: Overview of our fusion framework. Initial responses vary in professionalism (red), actionability (purple), and relevance (green); darker indicates higher. The fusion mechanism results in *consistent* responses that address individual needs and combine the strengths of the initial responses: all users receive responses with high professionalism, actionability, and relevance.

Given a set of responses, the degree of variation is measured as the variance of scores in the three dimensions across all responses.

$$\text{Variation} = \frac{1}{3}\left(\text{Var}_{\text{prof}} + \text{Var}_{\text{act}} + \text{Var}_{\text{rel}}\right) \quad (1)$$

where Var_{prof}, Var_{act}, and Var_{rel} represent the variances in *professionalism*, *actionability*, and *relevance*, respectively.

The consistency score is defined as:

$$\text{Consistency Score} = 1 - \text{Variation} \quad (2)$$

Higher scores indicate better consistency, which refers to minimized fluctuation in standards that reliably address user needs across diverse queries and requests (Kovač et al., 2024). It also supports scalability by ensuring that all users receive uniformly relevant, actionable, and professional guidance regardless of context or input variation.

4 A Fusion Framework for Consistent Generation

We propose a fusion framework to achieve consistent response generation in crisis communica-

tion. The framework is designed to integrate the strengths of conventional controllable response generation methods, balancing the key dimensions in crisis communication to achieve maximum consistency. Figure 2 illustrates the fusion framework.

The framework leverages a fusion-based generation strategy that integrates generations from state-of-the-art approaches. Rather than selecting one output, we introduce a prompt-driven fusion mechanism that evaluates outputs by different models across critical communicative dimensions and synthesizes a new, improved response that draws on the strengths of both. We represent the process using the following formulation with the example of using Instructional Prompt (IP) and Retrieval-Augmented Generation (RAG):

$$\mathrm{CC}(N, D) = \mathcal{L}\Big(\mathrm{Fuse}\big(M_{\mathrm{IP}}(N),\ M_{\mathrm{RAG}}(N),\ \mathbf{s}_{\mathrm{IP}},\ \mathbf{s}_{\mathrm{RAG}}\big)\Big)$$

$\mathrm{CC}(N, D)$ represents the response generation process for a given crisis needs N within a crisis-specific context D. The model generates two candidate responses: $M_{\mathrm{IP}}(N)$ via Instructional Prompt, and $M_{\mathrm{RAG}}(N)$ via Retrieval-Augmented Generation. $\mathbf{s}_{\mathrm{IP}}$ and $\mathbf{s}_{\mathrm{RAG}}$ represent the score vectors of the Instructional Prompt and RAG outputs respectively, evaluated along three communicative dimensions: professionalism, actionability, and relevance. $\mathrm{Fuse}(\cdot)$ compares and balances the strengths of $M_{\mathrm{IP}}(N)$ and $M_{\mathrm{RAG}}(N)$ in these dimensions and generates a new response optimized across all aspects. The process is further detailed in three steps.

Candidate Response Generation We employ state-of-the-art inference strategies to generate candidate responses, including the Instructional Prompt and RAG. These two methods are selected as they represent complementary approaches to response generation: Instructional Prompting leverages the reasoning and generalization capabilities of LLMs through carefully designed prompts, while RAG incorporates external evidence retrieved from a knowledge corpus to ground responses in factual content. This combination enables both flexibility and factuality, which are crucial for high-quality response generation. While other advanced methods exist, such as fine-tuned generation models or knowledge editing, we focus on Instructional Prompting and RAG due to their strong empirical performance, modularity, and ease of integration in diverse downstream tasks.

Instructional Prompt leverages zero-shot learning to generate crisis responses. As detailed in Appendix C, the prompt is crafted to define both the structure and intent of the response. The primary objective is to elicit outputs that consistently demonstrate high levels of professionalism, actionability, and relevance. We experiment with variations of prompts and choose the one with the best performance in three evaluation dimensions for the following experiments (See Appendix B).

Another method to generate candidate replies is RAG, which integrates external knowledge to provide factual information. We refer to the authoritative resources from the Federal Emergency Management Agency (FEMA)[1] to build our knowledge base, for example the *Individual Assistance Program and Policy Guide,* which provides accessible programs and policies designed to support individuals during disaster.[2] FEMA's publications are grounded in government-endorsed emergency management protocols, ensuring their reliability as sources of factual information. They are tailored to various crisis scenarios, including hurricanes, wildfires, floods, and pandemics, offering relevant information for crisis responses.

After collecting the knowledge, we construct a knowledge base for retrieval. Given the resources $S = \{D_1, D_2, \ldots, D_N\}$ from FEMA, we split the content into individual documents to form the knowledge base $K = \{d_1, d_2, \ldots, d_N\}$ for downstream retrieval. To enhance retrieval effectiveness, we adopt a hybrid approach that combines keyword-based and semantic retrieval methods, which has been shown to outperform single-method retrieval (Anik et al., 2025; Sawarkar et al., 2024). The hybrid retriever (R_h) integrates the strengths of keyword-based (R_k) and semantic retrieval (R_s) via union: $R_h = R_k \cup R_s$. When retrieving the top-N documents ($R_h = \{d_1, d_2, \ldots, d_N\}$), these documents are concatenated into a single context: $C = \mathrm{concat}(d_1, d_2, \ldots, d_N)$. The concatenated context C is then paired with the input query q to construct the prompt for the LLM to generate responses r. We acknowledge we haven't incorporated real-time information, which could enhance adaptability in crisis communication, but this also incurs higher computational costs. We plan to explore the integration of real-time data in future work to further improve crisis communications.

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

[2] https://www.fema.gov/sites/default/files/documents/fema_iappg-1.1.pdf

Multi-dimensional Evaluation After obtaining candidate responses, evaluations are conducted to provide criteria for fusion. For professionalism and actionability, the evaluation measures how users in crises would perceive these qualities. Given the lack of established automatic metrics for these dimensions and the high cost of recruiting real users, we utilize LLMs (GPT-4o mini[3]) as evaluators to assist with the evaluations (Coche et al., 2021). The detailed instructions and generation are fed to LLMs to obtain the professionalism and actionability score. For relevance, we refer to previous studies to assess the similarity between generated responses and crisis needs using BERTscore (Zhang et al.; Zhou et al., 2024; Liusie et al., 2024). Additionally, we implement human evaluations to validate the assessment of LLMs. The details of evaluations are presented in Appendix B.

Fusion-based Generation The output of a single model may be unstable. To address this, we aggregate the outputs of multiple models, leveraging the strengths of each model. This fusion-based approach enables us to generate more balanced results across various critical dimensions, demonstrating higher overall quality and exhibiting consistency.

We design various in-context learning-based fusion methods. First, we experiment with *Fusion with Evaluation Scores* (Fusion w/ Eval). This method provides the LLM with numeric scores (e.g., professionalism, actionability, and relevance) associated with each candidate response. The model uses these scores as implicit guidance to identify and integrate the stronger elements of each response. However, without further instructions, the model may not consistently interpret or act upon the scores effectively. Second, we design *Fusion with Evaluation Scores and Structured Instructions* (Fusion w/ Eval & Instruct). Building upon the first method, this approach augments the score information with a prompt template that explicitly instructs the model to reason over the scores. The template directs the LLM to compare the candidate responses, retain the strengths from one, integrate key elements from the other, and synthesize them into a well-rounded output. This ensures more deliberate, interpretable fusion behavior and mitigates ambiguity in how the model uses the evaluation scores. Third, we define *Fusion with Weighted Evaluation Guidance* (Fusion w/ Eval & Weight Instruct). Recognizing that optimizing all quality dimensions simultaneously may not always be feasible, we introduce weighted scores that reflect the relative importance of each dimension (e.g., 40% professionalism, 40% actionability, 20% relevance). These weights guide the model to prioritize more critical dimensions during synthesis. This approach supports targeted optimization and helps enhance the overall response quality, especially in settings where trade-offs between dimensions are necessary.

[3]Available at: https://platform.openai.com

5 Experiments and Results

5.1 Dataset

We use a Twitter (now X) dataset containing 1,013,313 geotagged posts from U.S. states affected by hurricanes Harvey, Irma, and Maria between August 15 and October 12, 2017. Geotagged tweets are used to ensure posts are from crisis-affected individuals.

Detect Information Needs Related Posts We train three RoBERTa models to predict whether a tweet expresses information needs (Alam et al., 2021b). Our classifiers are trained with three crisis datasets annotated with "needs or request" and other categories (Alam et al., 2021a,b). A tweet is labeled as "needs-related" if all three classifiers predict it as such. We opt for three smaller models rather than directly relying on LLMs for detection because they are more accurate and efficient. We then conduct human validation to verify the predictions (Song et al., 2025c) (See details in Appendix A.1). Two research assistants are employed to annotate crisis needs. 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. We finally obtain 540 information needs related posts for experiments.

5.2 Experiment Setup

We experiment with several open-sourced LLMs, including Llama-3.1-8B-Instruct [4] and Ministral-8B-Instruct-2410 [5] , which are good at conversational communications (Taori et al., 2023; Zheng et al., 2024; Li et al., 2024).

[4]Available at https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[5]Available at https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

Model	Category	Method	Professionalism	Actionability	Relevance	Overall Quality	Consistency
Llama	Baseline	Instructional Prompt	0.74 (0.33)	0.52 (0.36)	0.80 (0.02)	0.66	0.76
		RAG	0.96 (0.14)	0.63 (0.33)	0.80 (0.02)	0.80	0.84
		RAG-PE	0.94 (0.19)	0.50 (0.14)	0.80 (0.02)	0.74	0.88
		Prompt and Select	0.50 (0.50)	0.98 (0.14)	0.79 (0.02)	0.75	0.78
		Fusion w/o Eval	0.55 (0.27)	0.97 (0.16)	0.79 (0.02)	0.77	0.85
	Fusion	Fusion w/ Eval	0.98 (0.10)	0.77 (0.27)	0.79 (0.02)	0.86	0.87
		Fusion w/ Eval & Instruct	0.92 (0.19)	0.99 (0.07)	0.79 (0.02)	0.92	0.91
		Fusion w/ Eval & Weight Instruct	**0.99 (0.07)**	**0.99 (0.09)**	**0.79 (0.02)**	**0.95**	**0.94**
Mistral	Baseline	Instructional Prompt	0.87 (0.34)	0.98 (0.15)	0.79 (0.02)	0.90	0.83
		RAG	0.87 (0.22)	0.97 (0.11)	0.81 (0.03)	0.90	0.88
		RAG-PE	0.76 (0.26)	0.96 (0.15)	0.80 (0.02)	0.85	0.86
		Prompt and Select	0.75 (0.39)	0.81 (0.39)	0.80 (0.03)	0.78	0.73
		Fusion w/o Eval	0.93 (0.25)	1.00 (0.04)	0.80 (0.02)	0.93	0.90
	Fusion	Fusion w/ Eval	0.92 (0.28)	1.00 (0.08)	0.80 (0.02)	0.93	0.87
		Fusion w/ Eval & Instruct	0.96 (0.13)	1.00 (0.05)	0.80 (0.02)	0.94	**0.93**
		Fusion w/ Eval & Weight Instruct	**0.97 (0.13)**	**1.00 (0.08)**	**0.80 (0.02)**	**0.95**	0.92

Table 1: Results (mean and standard deviation) using Llama and Mistral for response generation. Overall quality is the weighted average of professionalism, actionability, and relevance. While relevance remains roughly the same across all methods, our fusion approach generates the most consistent responses across the board while increasing both professionalism and actionability with Llama, and professionalism with Mistral.

5.2.1 Baselines

Instructional Prompt We use the prompt detailed in Appendix C as a baseline model and for generating candidate responses for fusion. We further experiment with various temperature settings and find out TEMPERATURE 0.6 performs better in our task (Table 2).

RAG As mentioned in Section 4, we collect resources from FEMA to construct the knowledge base and use a hybrid search method incorporating two retrieval methods: keyword-based retrieval and semantic retrieval, using all-mpnet-base-v2[6] as the embedding model. In the generation process, we select the top-5 retrieved documents and concatenate them into a single context, providing additional knowledge for LLMs. The combined context and the full prompt are fed into the LLMs to generate responses.

RAG with Prompt Engineering (RAG-PE) To examine whether the consistency and overall quality will be improved by prompt engineering and prove the necessity of the fusion work, we experiment with RAG-PE, where the prompt is iteratively refined based on RAG's performance to generate effective responses across three dimensions. This method combines the strengths of RAG and Instruction Prompt with refined guidance. However, as RAG-PE relies on a single model, we hypothesize that RAG-PE may not achieve the same level of consistency as fusion models.

[6]Available at https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Prompt and Select Following prior work on response generation (Hong et al., 2024; Zhu and Bhat, 2021), we implement this approach, where LLMs are prompted to generate multiple candidates and the better response is chosen based on the evaluation scores. This method allows us to investigate whether selecting the most suitable response without fusion can improve consistency in the generated outputs. The fusion approaches allow for further optimization of candidate responses, presumably enabling the generation of outputs with better quality and reduced variances.

Fusion without Evaluation Score (Fusion w/o Eval) Given that all our fusion methods incorporate evaluation scores as guidance, we design an experiment to examine whether LLMs can independently recognize the strengths without such kind of instructions. Therefore, we conduct an experiment where candidate responses are fused without referencing evaluation scores.

5.2.2 Validation of Evaluators

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 (See details in Appendix A.2). We randomly sample 100 tweets and their responses 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 the LLM evaluator. The agreement

Setup	Pro	Act	Rel	Consist
TEMPERATURE 0.4	0.41	0.40	0.02	**0.72**
TEMPERATURE 0.5	0.48	0.24	0.02	**0.75**
TEMPERATURE 0.6	0.33	0.36	0.02	**0.76**
TEMPERATURE 0.7	0.30	0.39	0.02	**0.76**
TEMPERATURE 0.8	0.38	0.43	0.02	**0.72**

Table 2: Professionalism, actionability, relevance and consistency using different temperatures experimenting on Instructional Prompt using Llama-3.1-8B-Instruct.

rate and Cohen's Kappa ($\kappa \geq 0.72$) between human evaluation and LLM evaluation demonstrate substantial agreement.

5.2.3 Model Settings

We set all parameters the same for LLMs in the experiment. We set `max_new_tokens=256` for detailed yet concise responses. Sampling is enabled (`do_sample=True`) with a temperature (`temperature=0.6`) as it generates the best results. The `top_p=0.9` setting allows for some diversity while filtering unlikely tokens. Fusion prompts are detailed in Appendix C.

5.3 Results

Table 1 presents the results generated by the baseline and the fusion models.

Moderate temperatures yield the highest consistency in baseline generation. We first examine the effect of the temperature parameter on the consistency of generated responses. As shown in Table 2, setting the temperature to 0.6 or 0.7 produces the highest consistency scores (0.76). This suggests that moderate levels of randomness strike an effective balance between diversity and stability in generation. In contrast, lower temperatures (e.g., 0.4) constrain variation but slightly reduce consistency, while higher settings (e.g., 0.8) increase variability at the cost of stable response patterns. Overall, our findings indicate that a mid-range temperature optimizes consistency.

Fusion methods outperform all baselines in overall quality across models. Fusion models retain similar relevance scores compared to baseline models; however, they can achieve much higher scores in professionalism and actionability, leading to high overall quality and low variance. In both Llama and Mistral, Fusion w/ Eval & Weight Instruct achieves the best overall quality score of 0.95. This indicates that integrating the strengths of candidate responses produces higher-quality results than relying solely on a single model.

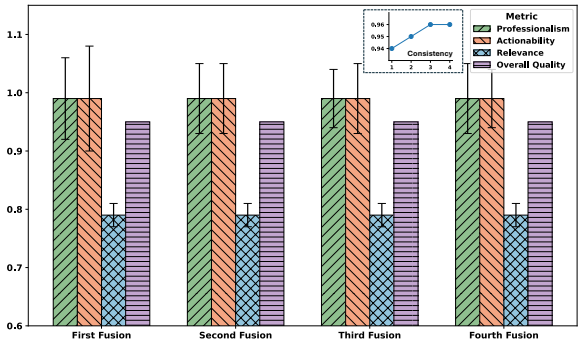

Figure 3: Results after one and more iterations of fusion with Eval & Weight Instruct and Llama-3.1-8B-Instruct. Consistency scores are visualized in a mini line chart. Average professionalism, actionability, and relevance remain high from the first iteration. On the other hand, consistency plateaus after three iterations.

Evaluation guidance is essential. Comparing fusion without evaluation guidance (Fusion w/o Eval), fusion with guidance (Fusion w/ Eval, Fusion w/ Eval & Instruct, and Fusion w/ Eval & Weight Instruct) achieves higher overall quality and consistency. The experiment confirms that fusion with evaluation guidance is more effective.

Consistency improves under structured fusion methods. For both the Llama and Mistral model, Fusion w/ Eval & Instruct and Fusion w/ Eval & Weight Instruct demonstrate better consistency compared to all five baseline methods. This indicates that LLMs with guided instructions are better at aggregating the strengths of individual responses, resulting in better consistency.

More fusion iterations do not lead to further improved performance. We further fuse the fused responses with responses generated by Instructional Prompt and RAG iteratively, using the Fusion with Eval & Weight Instruct configuration as a representative example. As illustrated in Figure 3, performance in key dimensions, professionalism, actionability, and relevance, remains consistently stable in multiple iterations of the fusion, while consistency improves slightly and reaches an optimal after three iterations.

5.4 Inconsistency Cause Analysis

To further investigate the cause of inconsistency, we have conducted a finer-grained analysis by grouping the crisis requests into need categories defined by previous studies (Zguir et al., 2025; Yang et al., 2024), and evaluating the variance of responses generated by Instructional Prompt using Llama-3.1-8B-Instruct on each category. Addition-

Needs Category	Professionalism	Actionability	Relevance	Overall Quality	Consistency
Evacuation	1.00 (0.00)	1.00 (0.00)	0.80 (0.02)	0.96	0.99
Food	1.00 (0.00)	1.00 (0.00)	0.80 (0.01)	0.96	1.00
Others	1.00 (0.00)	0.97 (0.18)	0.81 (0.02)	0.95	0.93
Rescue	0.98 (0.14)	0.98 (0.14)	0.80 (0.02)	0.94	0.90
Shelter	1.00 (0.00)	1.00 (0.00)	0.80 (0.02)	0.96	0.99
Average	0.99 (0.10)	0.98 (0.14)	0.80 (0.02)	0.95	0.91

Table 3: Few-shot learning performance across various dimensions.

ally, we conduct additional analyses exploring how user query characteristics, such as detailedness, sentiment, and formality, affect the consistency of LLM-generated crisis responses.

Specifically, we categorize our crisis requests by need type and annotate each query for its level of detail (vague, medium, and specific), sentiment (neutral and emotional), and formality (casual and formal). We then calculate professionalism, actionability, relevance, and consistency scores for responses within each group. The results, shown in Appendix D Table 7, reveal several important trends:

Response consistency is sensitive to linguistic variation within the same need type. For the Evacuation need, specific, neutral, and formal queries (Consistency: 0.90) outperform specific, emotional, and formal queries (Consistency: 0.74). This suggests that neutral sentiment in crisis scenarios may prompt more stable LLM behavior, potentially because emotional language introduces interpretive ambiguity or distracts from actionable content (Gandhi and Gandhi, 2025; Wang et al., 2025).

The type of crisis need influences response variance. For instance, Shelter queries that are specific and either neutral or formal achieve some of the highest consistency scores (0.82), while categories such as Rescue exhibit more moderate consistency and overall quality.

The role of sentiment is context-dependent. The sentiment dimension does not show a uniform impact across categories. In Rescue, both emotional and neutral sentiments yield comparable consistency (0.76 vs. 0.74), whereas in Food, emotional sentiment results in higher consistency (0.82) than neutral (0.77). This suggests that certain topics (like Food) benefit from emotional language, while others (like Evacuation) perform better with neutral expressions.

Furthermore, previous researchers found that few-shot learning reduces variability in responses to the same sample despite prompt variations (Zhuo et al., 2024). We have further conducted few-shot learning in our crisis response generation to investigate whether this method may improve the response consistency in crisis scenarios. We have drafted several response examples designed for diverse crisis needs and applied them in the few-shot learning experiment.

The results are shown in Table 3. The average performance of the few-shot learning approach (Overall Quality: 0.95; Consistency: 0.91) remains slightly lower than our best-performing model (the Fusion w/ Eval & Weight Instruct using Llama-3.1-8B-Instruct), which achieved an Overall Quality of 0.95 and a higher Consistency score of 0.94. Importantly, the fusion approach maintains robust performance and stability across a diverse range of user queries.

While few-shot learning effectively narrows the performance gap, especially when high-quality and targeted exemplars are available, our dynamic fusion model offers a more scalable and generalizable solution. It does not rely much on handcrafted prompts tailored to specific scenarios, making it more adaptable to real-world applications.

Moreover, our fusion method seamlessly integrates responses from RAG. Under this setting, we incorporate authoritative crisis-related knowledge from trusted sources such as FEMA, ensuring that the information provided is both accurate and contextually relevant. The inclusion of RAG also helps reduce hallucinations commonly produced by large language models, thereby further improving the factual reliability of responses.

6 Cross Crisis Generalization

To investigate the robustness of our fusion framework, we carry out experiments to generate responses to other crises such as earthquakes and typhoons. We employ the CrisisBench dataset (Alam et al., 2021b), which comprises a diverse set of crisis events.

We use the best-performing model, Llama-3.1-8B-Instruct (Consistency: 0.94, Overall Quality: 0.95), for the experiment. Table 4 reports the performance of baseline and fusion methods. Among the baseline methods, Prompt and Select performs better in consistency (0.91) and overall quality (0.93). Notably, fusion-based methods outperform the baseline methods. Especially, Fusion w/ Eval & Weight Instruct achieves the best consistency (0.96) and overall quality (0.95). These findings indicate

Method	Professionalism	Actionability	Relevance	Overall Quality	Consistency
Baseline Methods					
Instructional Prompt	0.93 (0.24)	0.94 (0.23)	0.79 (0.02)	0.91	0.84
RAG	0.94 (0.23)	0.97 (0.12)	0.77 (0.02)	0.92	0.88
RAG-PE	0.76 (0.39)	0.72 (0.40)	0.77 (0.02)	0.75	0.73
Prompt and Select	0.97 (0.12)	0.98 (0.12)	0.77 (0.02)	0.93	0.91
Fusion w/o Eval	0.96 (0.21)	0.97 (0.13)	0.78 (0.02)	0.93	0.88
Fusion-Based Methods					
Fusion w/ Eval	0.98 (0.10)	0.98 (0.11)	0.78 (0.02)	0.94	0.92
Fusion w/ Eval & Instruct	0.96 (0.15)	0.97 (0.15)	0.78 (0.02)	0.93	0.89
Fusion w/ Eval & Weight Instruct	**1.00 (0.00)**	**0.99 (0.11)**	0.78 (0.02)	**0.95**	**0.96**

Table 4: Cross-crisis generalization results (earthquake and typhoon) with Llama-3.1-8B-Instruct. While relevance decreases compared to the same-crisis scenario (Table 1), professionalism, actionability, and overall quality remain very high.

Metric	IP	RAG	Fusion
Agreement Metrics			
Agreement Rate	0.86	0.72	0.78
Cohen's Kappa	0.76	0.60	0.62
Evaluation Results			
User Preference	0.48	0.47	**0.86**
Consistency	0.83	0.77	**0.86**

Table 5: Human agreement and evaluation results across three strategies: IP = Instructional Prompt, RAG = Retrieval-Augmented Generation, Fusion = Fusion w Eval & Weight Instruct.

that our fusion framework not only performs well in hurricane-related contexts but also generalizes effectively to other crisis scenarios, confirming its applicability and robustness. We also repeat the experiments multiple times and present the results in Appendix E, Figure 4. The results show performance remains consistent across multiple rounds.

7 Qualitative Analysis

To investigate how humans perceive the generated crisis responses, we recruited two PhD students with a background in crisis computing to evaluate our responses. We select 50 responses generated by Instructional Prompt, RAG, and Fusion w/ Eval & Weight Instruct using Llama-3.1-8B-Instruct. (See evaluation guidance in Appendix A.3) We report the results in Table 5, which indicate a higher preference for the fused responses, with an average rating of 0.86 and a consistency score of 0.86.

Through human evaluations, we observe distinct characteristics across the different strategies. **Instructional Prompt:** Some responses offer clear and detailed instructions, while others are general and less actionable (e.g., *"Stay safe and indoors, away from floodwaters and fallen power lines"*). In some cases, the model incorrectly refuses to generate a response, citing concerns about facilitating a scam, although the original crisis need was legitimate. **RAG:** Some responses lack informativeness or appear evasive, using phrases such as *"I don't know."* Although a few responses provide detailed action steps, but some are vague and lack actionable clarity (e.g., *"Reach out to the American Red Cross or the Humane Society to inquire about foster care programs"*). **Fusion w/ Eval & Weighted Instruct:** Most responses follow a consistent structure that includes both guidance and concise explanation. These responses provide concrete instructions with reliable references (e.g., *"Reach out to the Harris County Emergency Management Office at (713) 755-5000 or the City of Houston's Emergency Management Office at (713) 837-0311 ...*). Compared to other methods, the fusion approach generates responses with high quality consistently.

8 Conclusion

We introduce the evaluation of consistency for crisis communication, which requires that responses are uniformly professional, actionable, and relevant for all contexts. To achieve the generation of consistent responses, we propose a fusion framework and conduct experiments with various open-sourced LLMs. Results show that our fusion framework can achieve better consistency and higher overall quality across professionalism, actionability, and relevance. In particular, the evacuation scores are beneficial and enhance the fusion process. Cross-crisis experiments have been conducted to show the robustness of our framework across diverse crisis contexts. Human evaluation proves that our fusion-based generation obtains more preference.

Limitations

Limited Candidate Generation. Even though we select the current state-of-the-art generation method to produce responses, there are still many other potential generation methods that could be used to further enhance the quality of fused responses. We will explore more diverse models and leverage their strengths to facilitate candidate response generation.

Limited Resource for RAG response. We collect information from FEMA, which is well-suited for our task. However, it is not sufficient to fully support crisis response generation due to the dynamic nature of real-world crises. In the future, we will collect more factual information from diverse sources and incorporate real-time information to assist crisis response generation.

Ethics Statement

This study makes use of publicly available data collected from Twitter (now X). All data were accessed in accordance with Twitter's Terms of Service and applicable platform policies. We ensured that the dataset does not contain personally identifying information beyond what is publicly visible, and we took steps to minimize potential risks to individual users. Specifically, any user identifiers were anonymized or removed, and only aggregated results are reported. We acknowledge that Twitter data may contain offensive, biased, or otherwise harmful content. Such instances were carefully considered during data processing, and filtering strategies were applied where appropriate to reduce the propagation of harmful material. The use of this dataset is strictly for research purposes, and no attempts were made to deanonymize users or to use the data outside of its original research context.

References

Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. 2021a. Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks. In *Proceedings of the International AAAI Conference on Web and social media*, volume 15, pages 933–942.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021b. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the International AAAI conference on web and social media*, volume 15, pages 923–932.

Anirban Saha Anik, Xiaoying Song, Elliott Wang, Bryan Wang, Bengisu Yarimbas, and Lingzi Hong. 2025. Multi-agent retrieval-augmented framework for evidence-based counterspeech against health misinformation. *arXiv preprint arXiv:2507.07307*.

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

Wout Broekema, Carola van Eijk, and René Torenvlied. 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.

Umar Ali Bukar, Marzanah A Jabar, Fatimah Sidi, RNH Binti Nor, Salfarina Abdullah, and Iskandar Ishak. 2022. How social media crisis response and social interaction is helping people recover from covid-19: an empirical investigation. *Journal of computational social science*, pages 1–29.

Waraporn Chatratichart, Yaninee Petcharanan, and Phansasiri Kularb. 2024. Inconsistency and obscurity of government-led communication during the pandemic. *Risk Communication and COVID-19: Governmental Communication and Management of Pandemic*.

Md Towhidul Absar Chowdhury, Soumyajit Datta, Naveen Sharma, and Ashiqur R KhudaBukhsh. 2024. Infrastructure ombudsman: Mining future failure concerns from structural disaster response. In *Proceedings of the ACM on Web Conference 2024*, pages 4664–4673.

Julien Coche, Jess Kropczynski, Aurélie Montarnal, Andrea Tapia, and Frederick Benaben. 2021. Actionability in a situation awareness world: Implications for social media processing system design. In *ISCRAM 2021-18th International conference on Information Systems for Crisis Response and Management*, 2391, pages p–994.

W Timothy Coombs. 2007. Protecting organization reputations during a crisis: The development and application of situational crisis communication theory. *Corporate reputation review*, 10:163–176.

Tiago Correia. 2024. Trust building in public health approaches: The importance of a "people-centered" concept in crisis response. *Risk Management and Healthcare Policy*, pages 1903–1908.

Wenlu Fan, Yuqi Zhu, Chenyang Wang, Bin Wang, and Wentao Xu. 2025. Consistency of responses and continuations generated by large language models on social media. *arXiv preprint arXiv:2501.08102*.

Vishal Gandhi and Sagar Gandhi. 2025. Prompt sentiment: The catalyst for llm change. *arXiv preprint arXiv:2503.13510*.

Vinicius G Goecks and Nicholas R Waytowich. 2023. Disasterresponsegpt: Large language models for accelerated plan of action development in disaster response scenarios. *arXiv preprint arXiv:2306.17271*.

Artur Grigorev, Adriana-Simona Mihaita Khaled Saleh, and Yuming Ou. 2024. Incidentresponsegpt: Generating traffic incident response plans with generative artificial intelligence. *arXiv preprint arXiv:2404.18550*.

Lingzi Hong, Pengcheng Luo, Eduardo Blanco, and Xiaoying Song. 2024. Outcome-constrained large language models for countering hate speech. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4523–4536.

Lingzi Hong, Xiaoying Song, Anirban Saha Anik, and Vanessa Frias-Martinez. 2025. Dynamic fusion of large language models for crisis communication. In *Proceedings of the International ISCRAM Conference*.

Yi-Hui Huang and Shih-Hsin Su. 2009. Determinants of consistent, timely, and active responses in corporate crises. *Public Relations Review*, 35(1):7–17.

Kyle Hunt, Puneet Agarwal, and Jun Zhuang. 2022. Monitoring misinformation on twitter during crisis events: a machine learning approach. *Risk analysis*, 42(8):1728–1748.

Zain Jafar, Jonathan D Quick, Heidi J Larson, Verner Venegas-Vera, Philip Napoli, Godfrey Musuka, Tafadzwa Dzinamarira, Kolar Sridara Meena, T Raju Kanmani, and Eszter Rimányi. 2023. Social media for public health: Reaping the benefits, mitigating the harms. *Health promotion perspectives*, 13(2):105.

Donghyun Kim, Youbin Ahn, Wongyu Kim, Chanhee Lee, Kyungchan Lee, Kyong-Ho Lee, Jeonguk Kim, Donghoon Shin, and Yeonsoo Lee. 2023. Persona expansion with commonsense knowledge for diverse and consistent response generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1139–1149.

Grgur Kovač, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2024. Stick to your role! stability of personal values expressed in large language models. *Plos one*, 19(8):e0309114.

Kyungchan Lee, Chanhee Lee, Donghyun Kim, and Kyong-Ho Lee. 2024. Dialogue act-based partner persona extraction for consistent personalized response generation. *Expert Systems with Applications*, 254:124380.

Zhenyu Lei, Yushun Dong, Weiyu Li, Rong Ding, Qi Wang, and Jundong Li. 2025. Harnessing large language models for disaster management: A survey. *arXiv preprint arXiv:2501.06932*.

Alexa Lenz and Steffen Eckhard. 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.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, et al. 2024. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *arXiv preprint arXiv:2402.13064*.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151.

Mohsen Mesgar, Edwin D Simpson, and Iryna Gurevych. 2021. Improving factual consistency between a response and persona facts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 549–562.

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

David Paulus, Ramian Fathi, Frank Fiedrich, Bartel Van de Walle, and Tina Comes. 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.

Sharon Lisseth Perez, Xiaoying Song, and Lingzi Hong. 2025. Analyzing the language of rejection: a study of user flagging responses to hate speech on reddit. *Information Research an international electronic journal*, 30(iConf):815–823.

Rajat Rawat. 2024. Disasterqa: A benchmark for assessing the performance of llms in disaster response. *arXiv preprint arXiv:2410.20707*.

Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. 2024. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. *arXiv preprint arXiv:2404.07220*.

Timothy L Sellnow and Matthew W Seeger. 2021. *Theorizing crisis communication*. John Wiley & Sons.

Xiaoying Song, Anirban Saha Anik, Dibakar Barua, Pengcheng Luo, Junhua Ding, and Lingzi Hong. 2025a. Speaking at the right level: Literacy-controlled counterspeech generation with rag-rl. *arXiv preprint arXiv:2509.01058*.

Xiaoying Song, Sujana Mamidisetty, Eduardo Blanco, and Lingzi Hong. 2025b. Assessing the human likeness of ai-generated counterspeech. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3547–3559.

Xiaoying Song, Sharon Lisseth Perez, Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2025c. Echoes of discord: Forecasting hater reactions to counterspeech. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4892–4905.

Larissa Steimle, Sebastian von Peter, and Fabian Frank. 2024. Professional relationships during crisis interventions: A scoping review. *Plos one*, 19(2):e0298726.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm/. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Stijn van Voorst, Sandra L Resodihardjo, and Andrea Schneiker. 2022. Humanitarian aid ngos' accountability towards large donors: the case of the european union's dg echo. *Journal of International Humanitarian Action*, 7(1):20.

Yifei Wang, Ashkan Eshghi, Yi Ding, and Ram Gopal. 2025. Echoes of authenticity: Reclaiming human sentiment in the large language model era. *PNAS nexus*, 4(2):pgaf034.

Yi Xiao and Shubin Yu. 2025. Can chatgpt replace humans in crisis communication? the effects of ai-mediated crisis communication on stakeholder satisfaction and responsibility attribution. *International Journal of Information Management*, 80:102835.

Pingjing Yang, Ly Dinh, Alex Stratton, and Jana Diesner. 2024. Detection and categorization of needs during crises based on twitter data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1713–1726.

Kai Yin, Chengkai Liu, Ali Mostafavi, and Xia Hu. 2024. Crisissense-llm: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics. *arXiv preprint arXiv:2406.15477*.

Ahmed El Fekih Zguir, Ferda Ofli, and Muhammad Imran. 2025. Detecting actionable requests and offers on social media during crises using llms. In *Proceedings of the International ISCRAM Conference*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Chen Zheng, Ke Sun, Hang Wu, Chenguang Xi, and Xun Zhou. 2024. Balancing enhancement, harmlessness, and general capabilities: Enhancing conversational llms with direct rlhf. *arXiv preprint arXiv:2403.02513*.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149. Association for Computational Linguistics (ACL).

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. Prosa: Assessing and understanding the prompt sensitivity of llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976.

Linda Ziberi, Lara Lengel, Artan Limani, and Victoria A Newsom. 2024. Affect, credibility, and solidarity: strategic narratives of ngos' relief and advocacy efforts for gaza. *Online Media and Global Communication*, 3(1):27–54.

A Human Evaluation

A.1 Crisis Needs Detection Guidance

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.

A.2 Validation of Evaluators

We engage two PhD students with a background in crisis computing to serve as human annotators. Each is provided with crisis needs paired with corresponding responses. We define the evaluation criteria in Table 6.

A.3 Qualitative Analysis

We provide the following evaluation guidance: *Assuming you are a user experiencing a crisis. Below is a crisis-related need and a generated response. Please rate the response on a scale from 1 to 5*

Metric	Definition	Criteria (Scoring Scale)
Professionalism	The extent to which the response conveys authority, credibility, and a well-substantiated foundation.	**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.
Actionability	The degree to which the response offers clear, practical, and relevant steps or guidance to address the concern or need expressed in the tweet.	**Score 0 (Non-Actionable):** The response fails to provide any practical guidance or relevant steps. It may be vague, off-topic, or merely acknowledge the problem without offering a solution. **Score 1 (Partially Actionable):** The response provides some guidance but lacks clarity and specificity. It may contain useful information but is incomplete, unclear, or too general to be effectively acted upon. **Score 2 (Fully Actionable):** The response clearly and specifically provides detailed guidance or steps that the user can take immediately. It includes direct actions, useful resources, or concrete advice that fully addresses the concern.

Table 6: Definitions and scoring criteria for response evaluation metrics.

based on your personal preference, considering the response's professionalism, actionability, and relevance to the given need. They independently evaluate the responses. The agreement rate and Cohen's Kappa score were both above 0.60, indicating moderate inter-rater reliability.

To ensure a thorough understanding of the evaluation criteria, the annotators undergo training using example samples. Each annotator independently reviews and labels the data. Upon completion, a discussion is conducted to resolve disagreements. If consensus cannot be reached, an expert reviewer provides the final adjudicated label.

B Evaluation Details

We prompt LLM to evaluate the professionalism and actionability of generated responses. We design detailed guidelines for both professionalism and actionability, as outlined in the following prompts. Additionally, 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 demonstrate substantial agreement.

1. Professionalism Evaluation:

```
You are an expert evaluator tasked with
    assessing the professionalism of
    responses. Your goal is to determine
     how well each response adheres to
    professional communication standards
    , providing reliable and well-
    structured information.
Criteria:
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.
Assign a score (0, 1, or 2) according to
     the criteria and reply with only
    the numerical score.
```

2. Actionability Evaluation:

```
 You are an expert evaluator tasked with
     assessing the actionability of
     responses. Your goal is to
     determine how well each response
     provides clear, practical steps to
     address the concern or need
     expressed in the original tweet.
Criteria:
Score 0 (Non-Actionable): The response
    fails to provide any practical
    guidance or relevant steps. It may
    be vague, off-topic, or merely
    acknowledge the problem without
    offering a solution.
Score 1 (Partially Actionable): The
    response provides some guidance but
    lacks clarity and specificity. It
    may contain useful information but
    is incomplete, unclear, or too
    general to be effectively acted upon
    .
Score 2 (Fully Actionable): The response
     clearly and specifically provides
    detailed guidance or steps that the
    user can take immediately. It
    includes direct actions, useful
    resources, or concrete advice that
    fully addresses the concern.\\

Assign a score (0, 1, or 2) and provide
    a brief justification for the
    assigned score.
```

C Generation Prompts

Candidate Response Generation Prompts

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

Need Category	Detailedness	Sentiment	Formality	Evaluation Metrics (mean (sd))			
				Professionalism	Actionability	Relevance	Consistency
Rescue	medium	emotional	formal	0.89 (0.22)	0.39 (0.33)	0.80 (0.02)	0.81
	specific	emotional	casual	0.77 (0.34)	0.50 (0.45)	0.80 (0.02)	0.73
	specific	emotional	formal	0.76 (0.35)	0.53 (0.36)	0.80 (0.02)	0.76
	specific	neutral	formal	0.55 (0.44)	0.65 (0.34)	0.79 (0.02)	0.74
Shelter	specific	emotional	formal	0.83 (0.32)	0.72 (0.33)	0.80 (0.02)	0.78
	specific	neutral	formal	0.67 (0.26)	0.75 (0.27)	0.79 (0.02)	0.82
Evacuation	specific	emotional	formal	0.68 (0.34)	0.55 (0.42)	0.81 (0.02)	0.74
	specific	neutral	formal	0.50 (0.00)	0.67 (0.29)	0.79 (0.02)	0.90
Food	specific	emotional	formal	0.68 (0.28)	0.42 (0.26)	0.80 (0.02)	0.82
	specific	neutral	formal	0.62 (0.23)	0.44 (0.42)	0.79 (0.03)	0.77
Others	medium	emotional	formal	0.62 (0.31)	0.33 (0.33)	0.80 (0.02)	0.78
	specific	emotional	casual	0.75 (0.42)	0.35 (0.24)	0.81 (0.03)	0.77
	specific	emotional	formal	0.78 (0.31)	0.52 (0.37)	0.80 (0.02)	0.77
	specific	neutral	formal	0.71 (0.26)	0.50 (0.39)	0.81 (0.02)	0.78
	vague	emotional	formal	1.00 (0.00)	1.00 (0.00)	0.80 (0.00)	1.00
$ComConne^{\dagger}$	specific	emotional	formal	0.67 (0.41)	0.58 (0.38)	0.80 (0.04)	0.73
$EmoPsycho^{\dagger}$	specific	emotional	formal	1.00 (0.00)	0.50 (0.00)	0.80 (0.01)	1.00
$MisTrap^{\dagger}$	specific	emotional	formal	0.50 (0.71)	0.00 (0.00)	0.80 (0.01)	0.76
Medical Help	specific	emotional	formal	0.57 (0.35)	0.21 (0.27)	0.79 (0.02)	0.79

Table 7: The variance of response across the same crisis needs with diverse linguistic features. $ComConne^{\dagger}$ indicates Communication or Connectivity Issues. $EmoPsycho^{\dagger}$ means Emotional or Psychological Support. $MisTrap^{\dagger}$ refers to Missing or Trapped Persons.

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, and relevant.

RAG-PE

You are an AI assistant designed to provide practical, actionable, and relevant advice for individuals seeking help related to crisis on social media. Use the provided documents to address the needs expressed in the tweet. If you don't know the answer, clearly state, "I don't know."
Guidelines:
1. Prioritize Immediate Actions: Break down advice into clear, numbered steps labeled as Step 1, Step 2, etc.
2. Explain Each Action: For every step, include a brief follow-up sentence explaining its importance or how to implement it.
3. Provide Resources: Include links, organizations, or contact information where relevant to help the user take action.
4. Stay Concise: Keep responses clear and to the point, avoiding unnecessary details.

Prompt and Select

You are an AI assistant designed to provide professional, actionable and relevant advice for someone seeking help during crises on social media. Two responses are provided, each with scores in three categories: Professionalism, Actionability, and Relevance.
Response 1: {response1}
Scores: {scores1}
Response 2: {response2}
Scores: {scores2}

Your task: Compare the two responses based on their scores. Return only the response that has the better overall performance.

Fusion w/o Eval

You are an AI assistant tasked with synthesizing two responses into one that optimally balances three key qualities: Professionalism, Actionability, and Relevance. Two responses are provided.
Response 1: {response1}
Response 2: {response2}
Your task is to merge these two responses into a single, cohesive answer. In doing so, you should maintain high levels of Professionalism, Actionability, and Relevance. Integrate the strongest elements from both responses and present the final response clearly. Only provide the final response.

Fusion w Eval

You are an AI assistant tasked with synthesizing two responses into one that optimally balances three key qualities: Professionalism, Actionability, and Relevance. Two responses are provided, each with scores in three categories: Professionalism, Actionability, and Relevance.
Response 1: {response1}
Scores: {scores1}
Response 2: {response2}
Scores: {scores2}
Your tasks are:
1. Internally analyze and compare the two responses based on their provided scores, identifying the strengths and essential elements of each.
2. Merge the strong qualities of Response 1 with the essential elements of Response 2 into a single, cohesive response that effectively balances Professionalism, Actionability, and Relevance. Only provide the final response.

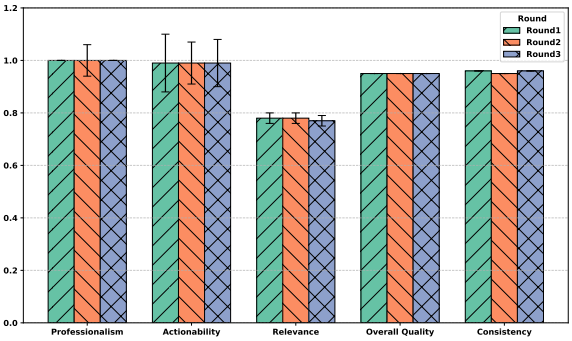

Figure 4: Multiple rounds of fusion w Eval & Weight Instruct in generalization experiments. The results demonstrate that the method produces stable performance regardless of the number of fusion rounds.

Fusion w/ Eval & Instruct

You are an AI assistant tasked with synthesizing two responses into one that optimally balances three key qualities: Professionalism, Actionability, and Relevance. Two responses are provided, each with scores in three categories: Professionalism, Actionability, and Relevance.
Response 1: {response1}
Scores: {scores1}
Response 2: {response2}
Scores: {scores2}
Your task: 1. Compare the two responses based on their scores.
2. Retain the {} and {} qualities from Response 1.
3. Incorporate the {} and {} elements from Response 2.
4. Merge these aspects into a single, well-rounded response that balances Professionalism, Actionability, and Relevance.
5. Provide only the final merged response.

Fusion w/ Eval & Weight Instruct

You are an AI assistant evaluating and fusing two

```
responses. Each response is accompanied by scores in four categories: Professionalism, Actionability, and Relevance.
Response 1:
response1
Scores: scores1
Response 2:
response2
Scores: scores2
Your task is:
1. Compare the two responses based on their scores in each category.
2. Synthesize the strengths of both responses to create a new, improved response that excels in all three areas.
3. The final quality of the improved response is determined by:
- Professionalism: 40%
- Actionability: 40%
- Relevance: 20%
4. Clearly list steps and explanations, resources, and provide contact information for the user to access help, the format:
- Step 1: Explanation, resources, and contact information
- Step 2: Explanation, resources, and contact information
Your objective is to produce a response that integrates the best elements of both responses, thereby achieving a higher overall quality.
```

D Inconsistency Cause Analysis

In Section 5.4, we have conducted a finer-grained analysis by grouping the crisis requests into need categories and evaluating the variance of needs and responses. Specifically, we categorized our crisis requests by need type and annotated each query for its level of detail (vague, medium, and specific), sentiment (neutral and emotional), and formality (casual and formal), referring to the linguistic analysis of Perez et al. (2025). We then calculated professionalism, actionability, relevance, and consistency scores for responses within each group. The results are shown in Table 7.

E Multiple Rounds Fusion

We run the fusion experiments multiple times to investigate whether the performance is stable. The results in Figure 4 suggest that our fusion method, incorporating evaluation scores and weighted instructions, is robust and maintains stable performance across multiple rounds of application. This indicates that increasing the number of fusion rounds does not significantly degrade or improve performance; it remains consistently strong across key quality dimensions.

F Computing Resources

The computational resources applied in this research include a high-performance server equipped with an Intel Xeon Gold 6226R processor, 128 GB memory, and 3 Nvidia RTX 8000 GPUs.

G Use of AI Assistants

We acknowledge the use of AI tools to assist with code writing and expression refinement. The authors developed all core ideas, methods, analyses, and conclusions. The final content reflects the authors' independent scholarly contributions.