Cross-Modal Inconsistency in Multimodal LLMs during Crisis Communication

Anuridhi Gupta
Humanitarian Informatics Lab
George Mason University
Fairfax, VA, USA
agupta29@gmu.edu

Hemant Purohit
Humanitarian Informatics Lab
George Mason University
Fairfax, VA, USA
hpurohit@gmu.edu

Abstract

Effective crisis communication relies on language technologies that can serve everyone. Although multimodal large language models (MM-LLMs) accept both speech and text, it remains unclear whether equivalent inputs produce equivalent crisis guidance across modalities, particularly for people with access and functional needs. In this paper, we evaluate three state-of-the-art (SOTA) MM-LLMs on consistency of their outputs across text and audio modalities using real emergency alerts. Building on prior semantic-similarity and factual-overlap evaluation, we introduce a LIWC-aligned analysis of modality dependent differences in linguistic structure, interpersonal framing, risk, and action-oriented language. Our findings indicate that despite being unified models, capable of taking in input from different modalities, no single model was able to achieve reliable cross-modal consistency. Persona-level results further indicate that cross-modal disparities vary across user groups. We release text prompts, audio prompts, model outputs and evaluation materials to support the crisis language technology community in building equitable solutions for crisis management.

1 Introduction

Crisis informatics has emerged as a critical area of research at the intersection of natural language processing and humanitarian action. The emergence of multimodal large language models (MM-LLMs) capable of processing and generating content across text, audio, image, and video within a single unified architecture (Yin et al., 2024) represents the next avenue for crisis language technology. Conversational interfaces allow users to submit crisis queries by voice or text and receive actionable written guidance in return (Urbanelli et al., 2024).

However, an important question is whether these systems give the same information regardless of whether the user speaks or types. Disasters disproportionately harm those who are least equipped to respond, including the 43% of the U.S. population with access and functional needs (Federal Emergency Management Agency, 2024). Among them are approximately 48 million Americans with hearing loss (National Institute on Deafness and Other Communication Disorders, 2024) and 8.2 million with vision disabilities (USAFacts, 2024) who may be unable to use standard alerts. For these individuals, omitted safety-critical details can hinder informed decisions and compromise safety.

Human interaction is inherently multimodal as we combine sight, hearing, touch, and other senses to perceive the world (Turk, 2014). Yet, many people around the world are limited in these possibilities due to hearing, vision, speech, and cognitive disabilities (Karpov and Ronzhin, 2014), so emergency information must be accessible and adapted to their needs and preferences. High-stakes events such as disasters further worsen their experience, making them vulnerable to more physical and mental injuries (Tannenbaum-Baruchi et al., 2024; Stough and Kelman, 2017). Failure to account for their access and functional needs, such as when communicating information about the nearest shelters, can result in serious consequences.

Although artificial intelligence (AI) systems have shown promising applications for crisis communication, for AI systems to understand the situation, they need to be able to interpret and reason about these diverse interactions that individuals with disabilities require (Baltrušaitis et al., 2018). As these MM-LLMs move toward deployment, their evaluation plays a crucial role in development and deployment. Despite the claim of a unified MM system, these systems often demonstrate misalignment in understanding and response generation across input modalities (Zhao et al., 2025). They are prone to hallucinations, are fragile to adversarial input, and fail to produce appropriate responses in high-stakes domains, such as neglect-

Proceedings of the First Workshop on Language Technologies for Crisis Preparedness and Response, pages 115–125
November 9, 2026 ©2026 Association for Computational Linguistics

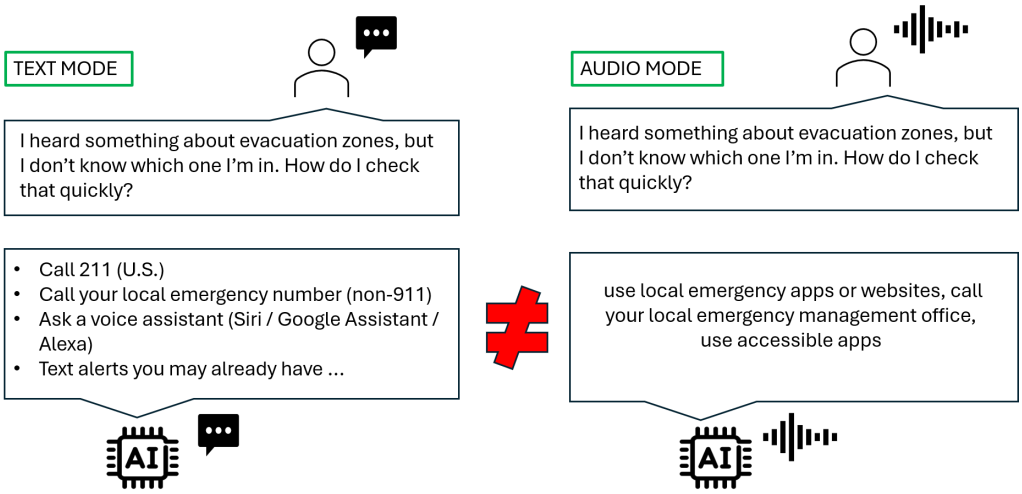


Figure 1: Illustrative GPT-4 interaction motivating our work on exploring the difference in responses across same emergency query given through text and voice input respectively.

ing the urgency of the risk and accessibility constraints in disaster risk communication. Previous work has suggested that some of these issues correlate with LLM inconsistency, generally defined as their tendency to generate low-confidence responses or conflicting responses when the same prompt is resampled (Wu et al., 2025). Accurate estimation of MM-LLM consistency is important during critical applications and supports user trust in the model (Liang et al., 2024). Building on the semantic-similarity and factual-overlap evaluation of Gupta et al. (2026), this study adds linguistic analysis using LIWC2015 categories and an adapted Language Style Matching measure.

Figure 1 demonstrates the problem through a GPT-4 interaction involving text and spoken responses. We test whether this pattern holds more systematically in three open-weight audio-language models: AudioFlamingo, SALMONN, and Qwen. Each model receives the same emergency question as typed text and as audio generated using text-to-speech (TTS), and we compare the two textual responses. Unlike Figure 1, all responses in our evaluation are text.

Problem Statement. MM-LLMs are increasingly capable of accepting semantically equivalent requests through speech and text, but multimodal capability does not ensure that the resulting guidance is consistent across input modalities. In crisis communication, modality-dependent omissions or changes can reduce the usefulness of guidance and create unequal service for vulnerable users. Existing multimodal benchmarks do not adequately measure this form of cross-modal informational consistency in safety-critical, crisis queries describing different user needs.

Contributions. We (1) extend an existing evaluation framework using audio and text prompts derived from Federal Emergency Management Agency (FEMA)'s Integrated Public Alert and Warning System (IPAWS) alert scenarios and instantiated across four access and functional needs personas; (2) provide a multifaceted evaluation combining semantic similarity, lexical overlap, preservation of safety-critical information, and LIWC-aligned linguistic analysis; and (3) release text prompts, audio prompts, model outputs, and evaluation materials for reproducible research on equitable crisis language technology. The text prompts, audio prompts, model responses are available in our public repository.[1]

Section 2 explores related work, Section 3 presents the framework, Section 4 reports results, Sections 5 discusses implications and Section 6 concludes with limitations and future work.

2 Related Work

2.1 Language Technologies in Crisis

Crisis informatics is a growing field that connects natural language processing with disaster preparedness, response, and recovery. Early work demonstrated the value of social media monitoring and text classification for situational awareness during

[1] https://github.com/samara-m1/multimodal-disaster-eval

active disasters (Purohit et al., 2014), establishing a foundation for NLP-driven crisis response that includes information extraction, rumor detection, and needs identification from crisis-generated text (Alam et al., 2021). The International Telecommunication Union has identified telecommunications and early warning systems as foundational infrastructure for all phases of disaster risk reduction, with particular emphasis on the importance of clear and usable guidance (International Telecommunication Union, 2025). With the emergence of LLMs, new possibilities have opened for interactive, personalized crisis communication (Urbanelli et al., 2024). Early work explored ChatGPT's potential for disaster prevention information dissemination and emergency response support because its natural language reasoning provided an advantage over static alert systems (Xue et al., 2023). More recently, LLMs have been used to classify crisis information from social media streams, monitor infrastructure during active disasters, and generate structured warning messages based on official guidelines (Linardos et al., 2025; Chen et al., 2024).

2.2 Accessibility in Crisis

People with access and functional needs are systematically underserved by existing crisis communication infrastructure. Traditional broadcast-based emergency alert systems such as sirens, television crawls, and Wireless Emergency Alerts have well-documented limitations in reaching individuals with sensory, cognitive, or linguistic barriers (Villarreal et al., 2025). Research has documented communication failures for hard of hearing individuals during Hurricane Rita, where announcements were exclusively accessible through specific television stations, translators were unavailable at shelters, and information was not conveyed in sign language or any other accessible manner (Tannenbaum-Baruchi et al., 2024). Adverse pregnancy outcomes have been documented during Hurricane Katrina (Callaghan et al., 2007), and significant cognitive deterioration has been observed in elderly individuals with dementia following hurricane and earthquake events (Fahmy and Alme, 2026; Güney and Çiçek Doğan, 2026). A systematic review found that emergency communication is not tailored to specific at-risk populations and adaptations change dissemination channels instead of message content (Lambert et al., 2025). This establishes the necessity of evaluating AI-based crisis communication systems across modalities for users with access and functional needs.

2.3 Cross-Modal Consistency in Speech and Language Models

Audio-capable language models such as SALMONN (Tang et al., 2024) and AudioFlamingo (Goel et al., 2025) have gained traction as they enable audio captioning, speech comprehension, and audio-based reasoning. Qwen3-Omni (Xu et al., 2025) can process text, audio, image, and video within a single unified training pipeline. AudioBench (Wang et al., 2025) provides a universal benchmark for audio LLMs across a range of listening tasks, establishing evaluation norms for the speech community that this work builds upon. A known challenge in these systems is cross-modality interference, where training on one modality degrades performance or consistency in another. Omni-modal LLMs attend to dominant modalities (text) while producing degraded or inconsistent outputs for audio inputs (Jiang et al., 2025).

Existing benchmarks address adjacent but different questions. AudioBench measures audio language understanding across listening tasks (Wang et al., 2025), while Holistic Evaluation of Multimodal Models (HEMM) evaluates models across tasks and datasets (Liang et al., 2024). Wu et al. (2025) show that BERTScore, BLEU, ROUGE, and USE capture different aspects of perceived consistency. Cross-modal consistency evaluates whether semantically equivalent speech and text inputs yield equivalent open-ended guidance. Prior work has not adequately evaluated this property. We address this gap by measuring whether meaning, safety-relevant information, and linguistic framing are preserved across speech and text conditions.

3 Cross-Modal Response Analysis Framework

Our framework consists of four components. The first component is *User Groups and Communication Needs*, which is based on FEMA and Centers for Disease Control and Prevention (CDC) guidelines for individuals with access and functional needs. The next component is *Persona Creation and Prompt Design*, which is grounded in real emergency alerts obtained from FEMA's alerts. Then we employ *Multimodal LLMs* to take inputs in the form of text and audio prompts which are then evaluated using the *Consistency Evaluation*. This framework

is designed to be open, reproducible, and extensible to additional disaster types and languages.

3.1 User Groups and Communication Needs

We grounded our selection of vulnerable populations in the U.S. Centers for Disease Control and Prevention (CDC) formal category of individuals with "access and functional needs" (Centers for Disease Control and Prevention, 2021), which refers to individuals who may need additional support during emergencies because of a temporary or permanent condition, without requiring a clinical diagnosis or medical evaluation. This is a deliberate design choice to keep the framework broadly applicable across many contexts. The CDC's full list spans a wide range of groups, including children, pregnant women, older adults, people with physical, sensory, intellectual, developmental, cognitive, or mental disabilities, people with chronic health conditions, those with limited English proficiency, and those facing financial, transportation, or legal barriers to preparedness. We narrowed this list to four groups for the current study. This includes *pregnant women, mothers with toddlers, hard of hearing individuals, and older adults with dementia*. These subgroups were selected because prior literature documents each group's vulnerability during disasters and each faces acute risk specifically during the communication phase of disaster response, which is this paper's focus.

Disasters have been linked to adverse outcomes for pregnant women, including increased rates of preterm birth, pregnancy complications, and maternal mortality (Harville et al., 2021; Callaghan et al., 2007). Mothers with toddlers face increased risk of acute illness, developmental disabilities, and maternal anxiety in post-disaster contexts (Yamazaki and Nakai, 2023; Nishihara et al., 2018). Hard of hearing individuals represent one of the most directly impacted groups as standard alerting systems including sirens, broadcast audio, and Wireless Emergency Alerts with sound are either inaccessible or partially inaccessible to them (Tannenbaum-Baruchi et al., 2024). Finally, research shows the adverse impacts of disasters on older adults with dementia, with documented declines in working memory, daily functioning, and general cognitive abilities after crisis events (Fahmy and Alme, 2026; Güney and Çiçek Doğan, 2026).

3.2 Persona Creation and Prompt Design

Each of the four stakeholder groups forms a persona used to generate realistic post-alert queries. We drew these scenarios from FEMA's Integrated Public Alert & Warning System (IPAWS)[2], the infrastructure that distributes verified alerts to the public over Wireless Emergency Alerts, the Emergency Alert System, and NOAA Weather Radio. Using real alerts, we built 40 prompts spanning alert types such as flash floods, hurricanes, tornadoes, dust advisories, and excessive heat warnings.

The prompt template follows a consistent structure intended to model naturalistic post-alert user behavior where a person receives an official emergency alert and asks an AI assistant for guidance relevant to their specific situation. This interaction is motivated by the Protective Action Decision Model, which describes warning reception, threat assessment, protective-action assessment, and information seeking as components of public responses to hazards (Lindell and Perry, 2012). Prior research emphasizes that people facing imminent threats need clear, specific, and actionable protective guidance and that warning responses are influenced by individual and situational characteristics (Sutton and Kuligowski, 2019; Kim et al., 2019). With that in mind, we designed the prompts. For example, a *Flash Flood Warning* prompt is *"I am a X, and I just received a flash flood warning. What should I do during this time to stay safe?"* where X represents one of the four personas. Appendix A provides representative queries for all four personas. For each persona–scenario combination, the same prompt was presented to the model as text and as audio. Audio inputs were synthesized from the query text with the Coqui TTS Python package, using the pretrained Tacotron 2 model (Shen et al., 2018), and saved as WAV files. Personas specify the user needs described in each query. Each persona is evaluated under both input conditions to test whether the model responds consistently to the same stated need. This design does not assume that members of a persona group prefer audio input. The synthesized voice does not simulate the speech characteristics of that group. The study does not evaluate variation across human speakers or challenging acoustic conditions such as background noise and reverberation.

[2] https://www.fema.gov/emergency-managers/practitioners/integrated-public-alert-warning-system

3.3 Multimodal LLMs

We evaluate three state-of-the-art open-weight MM-LLMs, selected for their distinct architectural approaches to speech and audio understanding, their public availability, and their relevance to the speech and language technology community. The generation settings are specified in our inference scripts on GitHub. AudioFlamingo 3 and Qwen3-Omni were configured with a maximum of 512 newly generated tokens, while SALMONN was invoked without explicit decoding overrides. Remaining decoding settings were inherited from the respective model configurations and implementations.

AudioFlamingo 3. Based on a 7B language model and the LLaVA architecture, AudioFlamingo 3 is trained on a unified AF-Whisper audio encoder that handles speech recognition and processing different audio signals, including sound, music, and speech (Goel et al., 2025). Its Whisper-based encoder makes it directly relevant to the task.

Qwen3-Omni. This model is capable of processing text, images, audio, and video while delivering real-time streaming responses in both text and natural speech (Xu et al., 2025). Its multilingual capabilities are relevant to our future work in low-resource and cross-cultural crisis communication.

SALMONN. Speech, Audio, Language, and Music Open Neural Network integrates pretrained audio encoders with an LLM through alignment layers, enabling audio captioning, speech comprehension, and audio-based reasoning (Tang et al., 2024). SALMONN is able to generalize across heterogeneous audio domains, making it suitable for applications requiring contextual understanding of real-world speech signals in crisis contexts.

3.4 Consistency Evaluation

In the quantitative evaluation, an *audio response* denotes the textual response generated when an audio input is processed through the models' audio branch, whereas a *text response* denotes the textual response generated from an input text through the models' text branch.

To quantify cross-modal consistency, we employ the following evaluation approaches across all 160 text and audio response pairs (40 prompts across 39 disaster alert types $\times$ 4 personas). The semantic-similarity and factual-overlap components follow the evaluation framework introduced by Gupta et al. (2026), while the psycholinguistic similarity extends that framework to examine differences in linguistic style and framing.

Semantic Similarity. We employ four different metrics. BERTScore (SBERT) (Zhang et al., 2020) captures deep semantic similarity using contextual embeddings. BLEU (SBLEU) (Papineni et al., 2002) and ROUGE (SROUGE) (Lin, 2004) capture lexical overlap at the n-gram level. Universal Sentence Encoder (SUSE) (Cer et al., 2018) captures sentence-level semantic similarity. Together, they provide a nuanced view of consistency.

Psycholinguistic Similarity. To complement the semantic metrics with a linguistic characterization of cross-modal consistency, we analyzed each response using LIWC2015 (Pennebaker et al., 2015). We then computed a category-level Language Style Matching (LSM) score (Niederhoffer and Pennebaker, 2002; Ireland and Pennebaker, 2010) between the responses produced by each pair of matched audio and text prompts. For LIWC category c, the score is

$$\mathrm{LSM}_c = 1 - \frac{|A_c - T_c|}{A_c + T_c + 0.0001}, \qquad (1)$$

where A_c and T_c are the percentages of words assigned to category c in the audio-prompt and text-prompt responses, respectively. The resulting score ranges from 0, indicating maximal divergence, to 1, indicating identical proportional usage. When both category values are zero, the score equals 1 and represents shared absence of that category.

Classical LSM focuses primarily on function-word coordination between human interlocutors. We extend its normalized similarity formulation to broader LIWC2015 categories to characterize modality-dependent differences in linguistic framing. Category-level scores were averaged across the 160 matched response pairs for each model. Where categories were grouped into broader dimensions, we then calculated the unweighted mean of their category-level scores. Word count was reported separately as a structural measure and excluded from the composite LSM score because LSM is intended for proportional language-category values rather than raw counts. The analyzed dimensions include function words and pronouns, cognitive and affective processes, social orientation, motivation, risk, perception, motion, and temporal focus. The purpose of this analysis is to determine whether responses that appear semantically similar differ in linguistic framing such as in their expression of urgency, certainty, risk, direct user address, or action orientation.

Table 1: Semantic similarity scores across the three MM-LLMs. (AF = AudioFlamingo)

Model	SROUGE	SBLEU	SUSE	SBERT
AF	0.223	0.065	0.619	0.495
SALMONN	0.249	0.132	0.696	0.625
Qwen	0.341	0.343	**0.896**	**0.896**

Factual Overlap. Following Gupta et al. (2026), we use a safety-critical information-overlap proxy that combines named entities, safety keywords, and ROUGE-L, treating the text response as the reference and the audio response as the comparison. It estimates whether instructions such as "stay indoors," "move to higher ground," or "avoid downed power lines" are preserved across modes. We report factual overlap scores both overall and broken down by persona in Table 2, enabling analysis of whether consistency gaps are uniform across user groups or concentrated among specific access and functional need populations. The response to text input serves as a comparison reference. Lower agreement does not, by itself, establish that either response is less correct, complete, actionable, or safe.

4 Results and Discussion

We present our findings through semantic similarity, factual overlap, linguistic analysis, and qualitative error examination, each revealing a distinct layer of the cross-modal inconsistency problem. Together, the findings build a cumulative case that inconsistency in MM-LLMs for disaster risk communication is not a surface-level quality gap.

4.1 Cross-Model Consistency Analysis

Table 1 reports semantic similarity scores for the three models. Qwen achieved the strongest performance across all four metrics, with an SBERT score of 0.896 indicating that the underlying meaning of responses was preserved more effectively across modalities than in the other two models. SALMONN demonstrated moderate performance, and AudioFlamingo (AF) performed worst across all metrics, indicating the weakest alignment in both meaning and wording across modalities.

4.2 Cross-Modal Consistency Analysis

A critical pattern that emerges across all models is that SUSE and SBERT are consistently higher than SBLEU and SROUGE, indicating that while models often preserve the general meaning of a response, they vary substantially in wording and structure across modalities. This distinction has direct implications for crisis communication. Qwen's close scores between SUSE (0.896) and SBERT (0.896) indicate relatively robust semantic grounding. However, its substantially lower SBLEU and SROUGE scores indicate that consistency is achieved through paraphrasing rather than faithful content reproduction. In safety-critical scenarios, reformulation of key guidance can significantly alter the actionability of a response. Semantic similarity alone is therefore insufficient to evaluate cross-modal consistency in crisis communication contexts. We address this in the next subsection.

Table 2: Factual overlap scores by persona across all three MM-LLMs. (AF = AudioFlamingo)

Persona	AF	SALMONN	Qwen
Pregnant woman	0.335	0.296	**0.452**
Mother w/ toddler	0.265	0.315	**0.449**
Elderly w/ dementia	0.271	**0.335**	0.426
Hard of hearing	0.237	0.306	**0.397**
Overall	0.277	0.313	**0.431**

4.3 Cross-Persona Consistency Analysis

Table 2 presents factual overlap scores broken down by persona for all three models, capturing whether safety-critical facts present in text responses are preserved in corresponding audio responses. Within AudioFlamingo and Qwen, the hard-of-hearing persona has the lowest factual-overlap score among the four personas evaluated. This pattern does not hold for SALMONN, where the pregnant-woman persona scores lowest, while the elderly with dementia persona scores highest (0.335). The rankings therefore differ across models. These scores indicate differences in response agreement and lack of stability across models. Persona rankings differ across models and a model can appear broadly consistent while systematically under serving specific user groups.

4.4 Language Style Matching Analysis

Table 3 reports mean LSM scores across the three models for the LIWC2015 dimensions most relevant to crisis communication. These include direct address (You), authoritative language (Power), risk language (Risk), uncertainty (Tentative), present-focused action framing (Focuspresent), social and motion language. Qwen has the highest LSM score in each reported linguistic dimension and

Table 3: Language Style Matching (LSM) scores between text and audio responses across crisis communication dimensions, computed from LIWC2015. Audio and Text indicate the mean category score per 100 words (raw word count for the Word Count row). LSM ranges from 0 (no correspondence) to 1 (identical usage).

Dimension	Model	Audio	Text	LSM
You	AudioFlamingo	0.44	3.58	**0.209**
	SALMONN	4.61	5.01	0.698
	Qwen	5.76	5.57	0.917
Power	AudioFlamingo	1.28	4.20	**0.268**
	SALMONN	2.57	3.15	0.678
	Qwen	2.84	2.75	0.881
Tentative	AudioFlamingo	2.00	5.07	0.315
	SALMONN	5.20	5.99	0.755
	Qwen	6.33	5.95	0.913
Risk	AudioFlamingo	7.25	3.85	0.558
	SALMONN	4.89	3.46	0.727
	Qwen	3.54	3.69	0.896
Social	AudioFlamingo	9.27	7.16	0.606
	SALMONN	9.35	8.66	0.789
	Qwen	11.96	12.14	0.934
Motion	AudioFlamingo	8.73	6.81	0.716
	SALMONN	5.44	5.83	0.752
	Qwen	4.46	4.55	0.895
Word Count	AudioFlamingo	28.07	45.48	0.701
	SALMONN	117.24	122.62	0.860
	Qwen	411.23	449.61	0.933
Focuspresent	AudioFlamingo	10.42	13.74	0.772
	SALMONN	12.62	14.43	0.859
	Qwen	13.04	12.85	0.938
Mean LSM	AudioFlamingo	0.492		
	SALMONN	0.751		
	Qwen	0.911		

the highest mean LSM score (0.911), followed by SALMONN (0.751), with AudioFlamingo showing the lowest consistency by a wide margin (0.492). AudioFlamingo's weakest dimensions are You (0.209) and Power (0.268), indicating that its audio responses diverge most from text in how directly they address the user and how authoritatively they instruct. Responses to audio input were shorter on average than responses to text input for all three models.

Persona-level LSM scores, computed identically to the model-level scores but broken out by persona (pooled across all three models), show that consistency loss is not uniform across the four groups and does not have a single ranking across dimensions. The hard-of-hearing persona has the lowest overall composite LSM (0.699) of the four personas, driven specifically by the You (0.580) and Motion (0.748) dimensions, indicating that responses for the hard of hearing persona exhibit particularly large modality dependent differences in direct address and action-oriented language. The mother with toddler persona shows its weakest consistency on Tentative (0.616) and Social (0.745) language, a difference in certainty and social framing rather than in direct address. The elderly with dementia persona is weakest specifically on Power (0.588) which is the authoritative dimension. The pregnant woman persona shows the highest composite consistency (0.741) and is the least divergent persona on four of the seven dimensions (You, Power, Tentative, Focuspresent). The observed linguistic differences vary across personas.

4.5 Qualitative Response Analysis

The quantitative and linguistic differences identified above also appear in the response examples shown in Table 4. In both cases, text responses provide detailed, actionable guidance tailored to the specific hazard, for example, avoiding dust exposure, staying hydrated, limiting physical exertion whereas corresponding audio responses are generic, shorter, and in some cases fail to address the scenario meaningfully.

In the dust advisory case, the audio response omits critical protective measures such as filtration, indoor sheltering, and mask usage and replaces them with vague instructions to "follow safety protocols." In a humanitarian context, where affected individuals may lack access to follow-on resources or reliable connectivity, such vagueness is not merely a quality issue but a safety risk. Elderly individuals with dementia are particularly vulnerable here as they are unlikely to seek clarification or consult secondary sources under crisis conditions (Bell et al., 2023), meaning the first response they receive is often the only one they act on. For the second case the text response provides specific, actionable guidance such as staying indoors, drinking plenty of water and using fans while the audio response instructs the user to "follow safety protocols" which is entirely unrelated to the specific disaster and the user's condition. Additional Qwen examples in Appendix B further demonstrate these differences.

5 Implications for Crisis Language Technology

Our findings indicate that cross-modal inconsistency in MM-LLMs is a systematic problem that ap-

Table 4: Examples of poor cross-modal consistency in AudioFlamingo. Blue: situation; Green: text-only guidance.

Input	Text Response	Audio Response
Elderly w/ dementia; dust advisory	Stay indoors, avoid dust, close windows, use a HEPA filter. Wear a dust mask if going outside.	Follow safety protocols, seek healthcare guidance, stay informed.
Mother w/ toddler; heat warning	Stay indoors, avoid sun, drink plenty of water, use fans or AC.	Follow safety protocols, seek shelter, stay informed.

pears across all three models and all four personas. The inconsistency is evident in semantic similarity (SBERT and SUSE), lexical overlap (SBLEU and SROUGE), preservation of safety-critical information (factual overlap), and linguistic style and framing (LIWC-aligned categories). Additionally, response agreement varies across personas and models, without a uniform ranking across groups.

For the crisis language technology community, these findings motivate three concrete directions. **First, cross-modal consistency must be treated as a top priority evaluation criterion** alongside accuracy and fluency, with training pipelines incorporating explicit consistency objectives (Tang et al., 2024; Xu et al., 2025). **Second, evaluation frameworks for crisis-oriented speech and language systems must include persona-level analysis** to reveal the inequity that emerges at the level of individual user groups. **Third, the linguistic dimensions most affected by cross-modal inconsistency are direct address (You), authoritative language (Power), and uncertainty (Tentative), and they provide an actionable target set for future fine-tuning efforts on disaster risk communication corpora**. A model unable to preserve directive, personally addressed language in audio input is not fit for deployment in humanitarian crisis contexts, despite its semantic similarity scores.

6 Conclusion

This paper extends prior evaluation of cross-modal consistency in disaster assistance with an analysis of linguistic style and framing. This is a critical property for crisis language technology systems serving users with access and functional needs. Using real emergency alerts from the publicly available FEMA IPAWS archive and persona-based scenarios grounded in CDC guidelines, we evaluated AudioFlamingo, SALMONN, and Qwen across text and audio modalities using semantic similarity metrics, factual overlap scoring, linguistic analysis, and qualitative error examination.

Our findings reveal that no model achieves reliable consistency across text and audio modalities. Audio responses were generally shorter and preserved fewer safety-relevant details than their text-prompt counterparts. The LIWC analysis further revealed modality-dependent differences in direct address, authoritative framing, and uncertainty.

Limitations and Future Work. This study has several limitations that point directly to future directions. Audio prompts were synthesized using a single TTS configuration and do not capture variation across human speakers. Future work will incorporate diverse speaker profiles and personas. The evaluation is English-only and uses U.S.-based alert scenarios, limiting generalization to multilingual and cross-cultural disaster contexts and extending this framework to low-resource language settings is an immediate priority aligned with the LT4CPR community's mission. Preliminary repeated testing on same modality prompt showed variation across generations, but the variation was not in terms of specificity or actionable detail. We plan to conduct this testing more systematically across a set of metrics. We also plan on independently assessing each model's speech recognition accuracy. Finally, this evaluation is purely computational, meaning that people with access and functional needs were not involved as participants. We are actively conducting focus groups and interviews with this population to validate our persona designs with lived experiences to inform the next phase of this research.

7 Acknowledgment

This research was partially supported by the grant # 2531369 from the National Science Foundation. During the preparation of this work, the authors used Claude and OpenAI Codex for code assistance, literature review support and proofreading. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the manuscript.

References

F. Alam, H. Sajjad, M. Imran, and P. Nakov. 2021. 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.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Sue Anne Bell, Marie Lynn Miranda, Julie PW Bynum, and Matthew A Davis. 2023. Mortality after exposure to a hurricane among older adults living with dementia. *JAMA network open*, 6(3):e232043.

William M Callaghan, Sonja A Rasmussen, Denise J Jamieson, Stephanie J Ventura, Sherry L Farr, Paul D Sutton, Thomas J Mathews, Brady E Hamilton, Katherine R Shealy, Dabo Brantley, et al. 2007. Health concerns of women and infants in times of natural disasters: lessons learned from hurricane katrina. *Maternal and child health journal*, 11(4):307–311.

Centers for Disease Control and Prevention. 2021. Access and functional needs toolkit: Integrating a community partner network to inform risk communication strategies. Technical report, CDC Center for Preparedness and Response, Atlanta, GA.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174.

Minze Chen, Zhenxiang Tao, Weitong Tang, Tingxin Qin, Rui Yang, and Chunli Zhu. 2024. Enhancing emergency decision-making with knowledge graphs and large language models. *International Journal of Disaster Risk Reduction*, 113:104804.

Christina Fahmy and Maria Nordheim Alme. 2026. Effects of natural disasters on the cognitive state of older adults with dementia: a scoping review. *European Geriatric Medicine*, pages 1–11.

Federal Emergency Management Agency. 2024. Access and functional needs support. Fact sheet, FEMA Mass Care and Emergency Assistance, Washington, DC.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*.

Seda Güney and Özlem Çiçek Doğan. 2026. Experiences of people with dementia and their family caregivers after earthquakes: A qualitative study. *Dementia*, 25(2):315–331.

Anuridhi Gupta, Samara Mansoor, and Hemant Purohit. 2026. Evaluating multimodal llms across text and audio modalities for accessible disaster assistance. *arXiv preprint arXiv:2608.14651*.

Emily W Harville, Leslie Beitsch, Christopher K Uejio, Samendra Sherchan, and Maureen Y Lichtveld. 2021. Assessing the effects of disasters and their aftermath on pregnancy and infant outcomes: A conceptual model. *International Journal of Disaster Risk Reduction*, 62:102415.

International Telecommunication Union. 2025. The use of telecommunications/ICT for disaster risk reduction and management. Output Report on ITU-D Question 3/1 D-STG-SG01.03.1-2025, ITU-D Study Group 1, Geneva, Switzerland. Study period 2022–2025.

Molly E Ireland and James W Pennebaker. 2010. Language style matching in writing: synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, 99(3):549.

Shixin Jiang, Jiafeng Liang, Jiyuan Wang, Xuan Dong, Heng Chang, Weijiang Yu, Jinhua Du, Ming Liu, and Bing Qin. 2025. From specific-mllms to omni-mllms: a survey on mllms aligned with multi-modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8617–8652.

Alexey Karpov and Andrey Ronzhin. 2014. A universal assistive technology with multimodal input and multimedia output interfaces. In *International Conference on Universal Access in Human-Computer Interaction*, pages 369–378. Springer.

Grace Kim, Andrea Martel, David Eisenman, Michael Prelip, Armen Arevian, Kerri L Johnson, and Deborah Glik. 2019. Wireless emergency alert messages: Influences on protective action behaviour. *Journal of Contingencies and Crisis Management*, 27(4):374–386.

Hélène Lambert, Dominique Doumont, Nina Reviers, Mieke Vandenbroucke, and Isabelle Aujoulat. 2025. Enhancing accessibility of crisis communication to people in vulnerable circumstances. *Journal of Public Health*, pages 1–13.

Paul Pu Liang, Akshay Goindani, Talha Chafekar, Leena Mathur, Haofei Yu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2024. Hemm: Holistic evaluation of multimodal foundation models. *Advances in Neural Information Processing Systems*, 37:42899–42940.

C.-Y. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Vasileios Linardos, Maria Drakaki, and Panagiotis Tzionas. 2025. Utilizing llms and ml algorithms in disaster-related social media content. *GeoHazards*, 6(3):33.

Michael K Lindell and Ronald W Perry. 2012. The protective action decision model: Theoretical modifications and additional evidence. *Risk Analysis: An International Journal*, 32(4):616–632.

National Institute on Deafness and Other Communication Disorders. 2024. Quick statistics about hearing.

Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of language and social psychology*, 21(4):337–360.

Mika Nishihara, Yasuhide Nakamura, Toru Fuchimukai, and Mayumi Ohnishi. 2018. Factors associated with social support in child-rearing among mothers in post-disaster communities. *Environmental health and preventive medicine*, 23(1):58.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report, University of Texas at Austin.

Hemant Purohit, Carlos Castillo, Fernando Diaz, Amit Sheth, and Patrick Meier. 2014. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday*.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Laura M Stough and Ilan Kelman. 2017. People with disabilities and disasters. In *Handbook of disaster research*, pages 225–242. Springer.

Jeannette Sutton and Erica D Kuligowski. 2019. Alerts and warnings on short messaging channels: Guidance from an expert panel process. *Natural Hazards Review*, 20(2):04019002.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pages 16607–16629.

Carolina Tannenbaum-Baruchi, Isaac Ashkenazi, and Carmit Rapaport. 2024. Risk inclusion of vulnerable people during a climate-related disaster: a case study of people with hearing loss facing wildfires. *International Journal of Disaster Risk Reduction*, 103:104335.

Matthew Turk. 2014. Multimodal interaction: A review. *Pattern recognition letters*, 36:189–195.

Angelica Urbanelli, Antonella Frisiello, Luca Bruno, and Claudio Rossi. 2024. The ermes chatbot: A conversational communication tool for improved emergency management and disaster risk reduction. *International Journal of Disaster Risk Reduction*, 112:104792.

USAFacts. 2024. How many people have a disability in the US?

Melissa Villarreal, Carson MacPherson-Krutsky, and Mary Angelica Painter. 2025. Barriers and best practices for inclusive emergency alerts and warnings. *International Journal of Disaster Risk Reduction*, 125:105581.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. 2025. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the NAACL-2025*, pages 4297–4316.

Xiaoyuan Wu, Weiran Lin, Omer Akgul, and Lujo Bauer. 2025. Estimating llm consistency: A user baseline vs surrogate metrics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30518–30532.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Zhiwen Xue, Chong Xu, and Xiwei Xu. 2023. Application of chatgpt in natural disaster prevention and reduction. *Natural Hazards Research*, 3(3):556–562.

Chisato Yamazaki and Hisao Nakai. 2023. Understanding mothers' worries about the effects of disaster evacuation on their children: a cross-sectional study. *International journal of environmental research and public health*, 20(3):1850.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. 2025. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*.

A Representative Persona Prompts

Table 5 presents representative queries for the four personas.

Table 5: Representative persona prompts used to obtain textual responses, derived from real Wireless Emergency Alerts.

Persona Name	**Alert Type**	**Persona Query**
Pregnant woman	Flood warning	*"I am a pregnant woman, and I just received a flood warning. What should I do during this time to stay safe?"*
Mother with toddler	Flood warning	*"I am a mother with a toddler, and I just received a flood warning. What should I do during this time to stay safe?"*
Hard-of-hearing individual	Flood warning	*"I am a hard-of-hearing individual, and I just received a flood warning. What should I do during this time to stay safe?"*
Elderly individual with dementia	Flood warning	*"I am an elderly individual with dementia, and I just received a flood warning. What should I do during this time to stay safe?"*

B Additional Qwen Response Examples

Table 6 presents selected Qwen response pairs illustrating differences in the specific details included under text and audio input conditions.

Table 6: Selected Qwen response excerpts for the same query under text and TTS audio input conditions. Both responses are textual. Blue highlights the hazard; green highlights details present in the text-input response but absent from the corresponding full audio-input response. Excerpts preserve wording, with Markdown formatting removed.

Input	**Text Response**	**Audio Response**
I am a hard of hearing individual, and I just received a flash flood warning. What should I do during this time to stay safe?	Monitor local news and weather updates via TV with closed captioning, apps with visual alerts, or text-based emergency notifications.	Follow local news on TV or radio if accessible.
I am an elderly individual with dementia, and I just received a excessive heat warning. What should I do during this time to stay safe?	Use fans if you don't have air conditioning, but know that fans alone may not prevent heat illness in extreme heat.	Use fans or air conditioning
I am a mother with a toddler, and I just received a excessive heat warning. What should I do during this time to stay safe?	Use a wide-brimmed hat and sunscreen (SPF 30+), and reapply every 2 hours or after sweating or wiping.	Use a wide-brimmed hat and sunglasses if outside.