DAISY: A Comparative Evaluation of Fine-Tuned GPT-3.5 and Zephyr-RAG for Natural Disaster Information Dissemination

Amanda He[1,2], Dominick Pelaia[3], Nathan He[4], Sadie Jones[3] & Flora Zhu[5]

Received February 12, 2026
Accepted May 11, 2026
Electronic access June 15, 2026

Natural disasters pose a great threat to communities worldwide, creating a pressing need for tools that rapidly disseminate accurate information. We present DAISY (Disaster Awareness and Information Serving You), a conversational AI chatbot that enhances public access to real-time disaster information. DAISY integrates Large Language Models (LLMs), GPT-3.5 Turbo and Zephyr-7B Alpha, with Retrieval-Augmented Generation (RAG)—a method that allows the model to retrieve relevant information from external data sources in real time rather than relying only on pre-trained knowledge—to deliver accurate, context-specific responses on environmental disasters such as hurricanes, volcanic eruptions, wildfires, tornadoes, earthquakes, subsidence, and floods. Real-time data is fetched from the Global Disaster Alert and Coordination System (GDACS). We evaluated the two models (GPT-3.5 Turbo and Zephyr-7B Alpha with RAG) on three metrics: answer length (word count), answer depth (human evaluation), and answer accuracy (verified using ChatGPT-4o with human supervision), with a score of 2 out of 3 metrics constituting a win. The Zephyr-RAG system outperformed the fine-tuned GPT-3.5 Turbo model in six of seven disaster categories, providing longer, more detailed, and more accurate responses. By leveraging RAG, the Zephyr-based chatbot was able to retrieve up-to-date external information and provide direct source citations in its responses. Utilizing relevant real-world information improved the accuracy and relevance of the responses, as reflected in higher accuracy and depth scores. Using earlier, simpler models allows for clearer comparison of system behavior in the presence of RAG. Nonetheless, future work includes replicating this analysis using newer-generation large language models to assess whether the advantages of RAG persist as base model capabilities continue to evolve.

Keywords: Chatbot, Machine Learning, Information Dissemination, Natural Disasters, Disaster Communication Systems, Retrieval Augmented Generation

Introduction

Natural disasters such as earthquakes, floods, hurricanes, wildfires, volcanic eruptions, tornadoes, and subsidence have posed significant threats to life and property. Recent natural disasters across the globe, such as the flooding in Brazil since April 2024 and wildfires in New Mexico, United States, have caused severe damage to the local populations and economies[1,2]. There is a pressing need for communication tools that can rapidly and reliably disseminate disaster information. This paper describes the design and implementation of a chatbot called DAISY (Disaster Awareness and Information Serving You) that utilizes fresh data to inform users about natural disasters across the globe.

Traditional methods of information dissemination, such as news broadcasts and public announcements, may not reach everyone promptly and cannot accommodate individual needs. DAISY leverages the LLMs of GPT and Zephyr to provide instant, real-time updates and personalized responses to user queries, ensuring that critical information is accessible to all users.

In real-world disaster scenarios, access to timely and accurate information can directly impact safety outcomes. For instance, during the 2023 Maui wildfires, many residents experienced challenges accessing timely evacuation information and sought information through various means to determine safe escape routes[3]. Similarly, following the 2023 Turkey–Syria earthquake, affected individuals sought immediate information about aftershock risks, structural safety, and access to emergency aid, often facing delayed communication[4]. In such situations, a conversational system like DAISY could enable users to ask questions and receive rapid, context-specific guidance, supporting more informed decision-making during emergencies.

Existing bots, like the American Red Cross's Clara chat-

[1] *Valley Christian High School*
[2] *University of Tennessee Knoxville*
[3] *L&M STEM Academy*
[4] *Ocean Lakes High School*
[5] *William A. Shine*

bot, can disseminate information, provide emergency response guidelines, and support disaster risk reduction initiatives[5]. However, these chatbots may not provide real-time data or comprehensive coverage. DAISY aims to tackle all these issues. Utilizing diverse, handpicked data sources ensures that DAISY can provide comprehensive and accurate insights into specific events and general knowledge about various natural disasters. Real-time data is fetched from the Global Disaster Alert and Coordination System (GDACS) to give users detailed reports on the casualties, range, location, and severity of current disasters, among other essential characteristics[6].

We created two independent versions of DAISY utilizing the Zephyr and GPT frameworks and measured the effectiveness of each across three metrics: answer length (word count), answer depth (evaluated by human reviewers), and answer accuracy (verified using the ChatGPT-4o model under human supervision). Between these two models, we hypothesized that Zephyr -7B Alpha with RAG would perform better across the three evaluation metrics. This expectation is grounded in the fundamental difference between static fine-tuned models and retrieval-augmented systems. A fine-tuned GPT model relies on patterns learned during training and is limited to the information encoded in its parameters, which may become outdated or incomplete[7]. In contrast, Retrieval-Augmented Generation (RAG) enables the model to retrieve relevant information from external sources at inference time, improving factual grounding and allowing responses to incorporate up-to-date, verifiable content[8]. This capability is particularly important in disaster scenarios, where accurate, real-time data is critical.

Zephyr-7B Alpha outperformed GPT-3.5 Turbo in six of seven categories in this evaluation, although this comparison does not isolate RAG as the causal factor, as the two systems differ in multiple architectural and training-related dimensions.

The analysis is limited to model performance on generated question–answer pairs assessed through human review and model-based verification, which introduces some subjectivity and does not capture real-world user interactions or long-term deployment effects.

Another limitation is the relatively small and structured set of test queries, consisting of five standardized questions per disaster category. While this design enabled controlled comparison across models, it may not fully capture the diversity and unpredictability of real-world user interactions. In practice, users often submit incomplete, urgent, or context-specific queries that differ significantly from well-formed evaluation questions. Increasing the number and diversity of test queries would improve the robustness of the evaluation and better reflect real-world usage conditions. While prior research has explored the use of chatbots and AI systems for disaster communication, DAISY does not aim to introduce the first such system. Existing platforms, including conversational agents developed for emergency response and disaster preparedness, have demonstrated the value of providing accessible, automated information during crises. However, many of these systems are limited in their ability to integrate real-time data or systematically evaluate different modeling approaches. Thus, DAISY focuses on a comparative evaluation of two distinct conversational AI architectures—a fine-tuned GPT-3.5 model and a Zephyr-based Retrieval-Augmented Generation (RAG) system—utilized in natural disaster information dissemination. By analyzing how these approaches differ in response quality, factual accuracy, and depth across multiple disaster scenarios, this work contributes to field-specific assessment of modeling strategies.

Related Work

Research on disaster communication has long emphasized that crises create urgent, dynamic information needs and that digital systems can help distribute and interpret time-sensitive information[9,10].

In the field of crisis informatics, disasters are often described as information-rich events in which timely data sharing and interpretation are essential for effective response[11].

A growing body of research has explored the use of artificial intelligence and data-driven systems to support disaster response. Early systems such as AIDR leveraged machine learning to automatically classify and process social media messages during disasters, improving situational awareness and enabling more efficient information management[12]. Similarly, studies on social media usage during natural disasters have shown that user-generated content can provide valuable real-time information into disaster conditions and public needs, highlighting the importance of systems that can organize and interpret large volumes of information[13]. Furthermore, survey work in this field has emphasized that processing large-scale, rapidly changing data streams remains a central challenge in emergency management systems[14].

Recent work has begun to explore the use of large language models (LLMs) specifically for disaster-related information dissemination and emergency response support. LLM-based systems have been applied to tasks such as summarizing crisis information, organizing large volumes of user-generated content, and assisting with real-time situational awareness during natural disasters. More recent studies suggest that LLMs can support decision-making and communication by synthesizing information from multiple sources and generating accessible, context-specific responses for affected populations[15]. However, these applications also highlight important limitations, particularly in high stakes environments where accuracy and reliability are critical. Prior research has shown that large language models may produce factually incorrect or hallucinated information, raising concerns about their safe deployment in

disaster communication systems[16,17]. These challenges underscore the need for approaches that improve factual grounding and ensure that responses are based on reliable, up-to-date information when LLMs are used in disaster contexts.

Within natural language processing, Retrieval-Augmented Generation (RAG) has emerged as an important approach for improving factual grounding by combining parametric language models with external knowledge retrieval. By incorporating relevant documents at inference time, RAG-based systems can produce more accurate and context-specific responses compared to models that rely solely on static training data[8]. This distinction is especially important for disaster communication, where information is constantly evolving and requires up-to-date contextual awareness.

Existing work underscores the need for robust, domain-specific assessment of AI-driven systems before they can be reliably deployed in real-world emergency contexts[14].

Methods

Data collection and preparation

DAISY uses reliable data from government agencies (e.g., FEMA, USGS), international organizations (e.g., UNDRR, WHO), and research institutions[18–21]. The dataset includes encyclopedia sources that provide general definitions of natural disasters and specific research papers for audiences in various situations. Data is collected through methods such as web scraping, API access, and manual entry, with rigorous data cleaning processes that include deduplication and validation to ensure accurate and high-quality information.

We defined explicit criteria for data cleaning and quality control. Deduplication was performed by identifying documents with highly similar titles, identical metadata, or overlapping text content exceeding approximately 90% similarity, in which case only one representative version was retained. Content validation involved removing documents that met any of the following conditions: (1) incomplete or truncated text (e.g., missing sections due to scraping errors), (2) non-informational content such as navigation pages or index listings, (3) outdated or irrelevant material not directly related to natural disaster characteristics, impacts, or response strategies, and (4) internally inconsistent or factually incorrect information identified through cross-referencing with authoritative sources (e.g., FEMA, USGS, WHO). Additionally, documents with excessive formatting noise (e.g., unreadable OCR outputs or encoding errors) were excluded to ensure clean text input for downstream processing. This structured filtering process ensured that the final dataset consisted of high-quality, relevant, and non-redundant documents suitable for both fine-tuning and retrieval.

Real-time data is scraped from the GDACS Extensible Markup Language (XML) archives to provide the status of natural disasters[22].

Source documents were used to generate question–answer pairs for fine-tuning, while evaluation questions were written independently using standardized templates and were not directly derived from or paired with specific training documents. Care was taken to ensure that evaluation questions were phrased separately from the generated Q&A pairs to avoid lexical or structural similarity.

This setup reduces the likelihood that models are evaluated on memorized training content and instead assesses their ability to generate responses to novel prompts. As a result, performance more accurately reflects generalization to realistic user queries rather than overlap with the training data.

While both models were exposed to similar underlying knowledge sources, the fine-tuned GPT model learned from static Q&A pairs, whereas the Zephyr-RAG system dynamically retrieved information at inference time.

Data categories

The dataset includes 7 distinct types of natural disasters: earthquakes, hurricanes, wildfires, subsidence, volcanic eruptions, floods, and tornadoes. Each disaster type contains anywhere from 20–100 PDF documents categorized by regions such as North America, South Asia, etc. These documents vary in length depending on source type, ranging from short informational briefs (˜1–5 pages) to longer technical reports and research articles (˜10–50+ pages). The dataset thoroughly explains each disaster's characteristics, trends, impact, frequency, and specific events, enhancing the overall quality and depth of DAISY's responses. The wide range of areas and natural disaster types is intended to accurately serve a broader audience of various backgrounds and situations. The dataset used has been made available on Harvard Dataverse (DOI: https://doi.org/10.7910/DVN/8UKTV8).

System architecture

DAISY was implemented with the Python programming language in Google Colab Pro + API for model development, leveraging LLMs to create an interactive chatbot. We explore two different implementations, one using GPT-3.5 Turbo fine-tuned on our dataset, and one using Zephyr-7B Alpha with Retrieval-Augmented Generation (RAG)[8,23,24]. For both implementations, real-time data is fetched, parsed, and merged with the existing 7 natural disaster datasets, then inputted into the respective models. We compared the evaluation results of DAISY using these models. Source code can be found at https://github.com/Alice-zou/DaisyAI-Developers.

Real-time data collection

Real-time data was collected primarily from the GDACS XML archive endpoint, which provides structured, machine-readable disaster event data. Supplementary metadata was occasionally accessed through the GDACS interface and integrated into the same JSON-based processing pipeline. The finalized data was then outputted as an array to be used as input for DAISY.

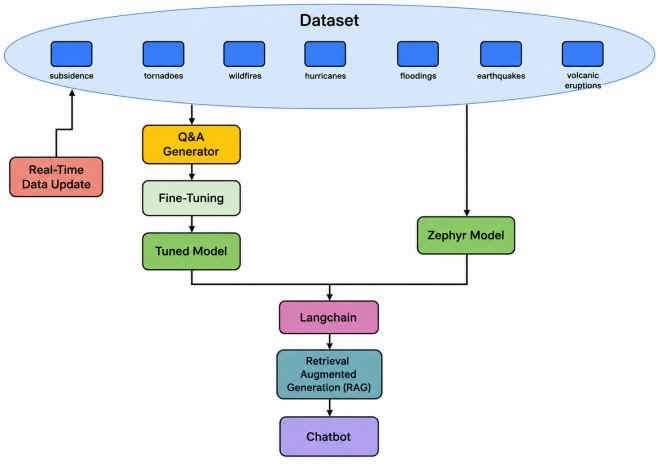

Fig. 1 System architecture implementation.7 natural disaster datasets were input into the two models. Real-time data was fetched, parsed, and merged with existing datasets on the spot. The left chain first used a generator to create question-answer pairs from the data. Hand-selected good-quality Q&A pairs went through Fine-Tuning, generating a Tuned Model. Using Langchain and Retrieval Augmented Generation, the Zephyr Model generated a webpage with the chatbot.

Fine-tuning GPT-3.5 Turbo

For the iteration of DAISY using the GPT-3.5 Turbo model, we fine-tuned the model using our newly collected natural disaster data formatted as JSONL files containing question-and-answer pairs. We first consolidated a set of PDFs from natural disaster datasets into a single document. This document was then segmented into a specified number of chunks for efficient processing. The GPT model was then prompted to generate a specific question for each chunk. The model processed each question and generated corresponding answers, forming key-value pairs that were stored in a dictionary. Finally, the dictionary was serialized into a JSONL file.

Errors and ill-generated pairs were manually removed from each JSONL file, and the remaining high-quality pairs were used as input for the fine-tuning process. With this refined dataset, we invoked OpenAI's fine-tuning APIs to generate the fine-tuned GPT-3.5 Turbo model for DAISY.

Prompt templates for Q&A generation followed a structured format in which the model was instructed to generate one clear, self-contained question per document chunk and a corresponding factual answer grounded only in the provided text. Prompts emphasized clarity, completeness, and avoidance of hallucinated information.

Fine-tuning was performed using OpenAI's GPT-3.5 Turbo fine-tuning API with JSONL-formatted question–answer pairs. The training process used default OpenAI hyperparameters, including automatic epoch selection based on dataset size, with a standard learning rate multiplier and batch configuration. No additional manual hyperparameter tuning was applied.

Zephyr model and RAG

We used RAG and the LangChain framework to utilize the fresh dataset for the iteration of DAISY using the Zephyr model[25]. The data ingestion process involved loading PDFs containing information on natural disasters, then splitting the documents into manageable chunks to facilitate efficient processing and embedding. The chunks were then embedded using a pre-trained model to create vector representations. The embedded document chunks are stored in a vector database using Chroma, allowing for efficient retrieval of relevant documents based on user queries[26].

A conversational retrieval chain is set up to handle user interactions, retrieve relevant documents, and generate appropriate responses. HuggingFacePipeline integrates the Zephyr model for response generation.

Documents were split into chunks of approximately 500–1000 tokens with an overlap of 100–200 tokens to preserve contextual continuity. Text embeddings were generated using a pre-trained sentence-transformer model, producing dense vector representations stored in a Chroma vector database. The embedding dimensionality followed the default configuration of the selected model.

During retrieval, the system used a top-k strategy (k = 3–5) to select the most relevant document chunks for each query. Retrieved context was incorporated into a structured system prompt instructing the model to generate responses grounded strictly in the provided sources. The Zephyr model was executed using a HuggingFace pipeline with a temperature of approximately 0.3–0.7 and a maximum token limit of 256–512 tokens per response.

User interface and experience

The user interface (UI) is designed to be intuitive and accessible. DAISY's UI is developed using Gradio, providing an engaging and user-friendly web page for interacting with the chatbot[27]. The interface allows users to submit queries and

receive responses in a conversational format. It also provides reference texts relevant to the generated answers.

Results

Question-and-answer (Q&A) pairs were generated from raw PDF files for seven natural disaster topics, and all pairs were manually reviewed for quality (Table 1). Inaccurate Q&A pairs were removed, and the remaining 1,440 high-quality pairs (out of 1,552 generated) were used to fine-tune the models. However, differences in dataset quality likely contributed to the variation in pass rates. For example, earthquakes and flash floods had pass rates of 77.04% and 81.03%, respectively; these pass rates fell lower than the rest of the natural disaster types analyzed. Overall, the Q&A generation approach achieved an average pass rate of 92.79% across all seven disaster categories (Table 1). We then evaluated the performance of two versions of the DAISY chatbot: one based on the fine-tuned GPT-3.5 Turbo model and another based on the Zephyr-7B model with Retrieval-Augmented Generation (RAG). Each chatbot was asked the same five questions for each of the seven natural disasters (covering both general definitions and user-response guidance). Performance was evaluated using three parameters: answer length, answer depth, and answer accuracy. Answer length was measured by word count, answer depth by human evaluation, and answer accuracy by the latest ChatGPT-4o model with human supervision; specifically, answer accuracy was assessed using a combination of model-assisted verification (ChatGPT-4o) and manual cross-checking against authoritative sources such as FEMA, USGS, and WHO. The LLM-based evaluation was used as a supportive tool rather than a definitive judge, and final accuracy judgments were made with human oversight to ensure factual correctness.

To reduce subjectivity in the depth evaluation, a standardized scoring rubric was used. Responses were evaluated on a 3-point scale:

(1) Low depth – response is minimal, lacks explanation, or omits key information;

(2) Moderate depth – response provides a basic explanation but lacks detail, examples, or completeness;

(3) High depth – response is detailed, well-structured, and includes comprehensive explanations, relevant context, and actionable insights.

Depth evaluations were conducted by multiple human reviewers to improve reliability. Reviewers were instructed to apply the rubric consistently across all responses. To reduce bias, responses were anonymized so that reviewers were not aware of which model generated each answer. In cases of disagreement, scores were resolved through discussion and consensus.

Overall, the Q&A generation process achieved a high average pass rate of 92.79%, indicating that the majority of generated question–answer pairs were of sufficient quality for training. However, pass rates varied across disaster types, with earthquakes (77.04%) and flash flooding (81.03%) showing notably lower performance compared to others such as tornadoes (99.39%) and wildfires (98.94%). This variation suggests that differences in source data quality or complexity across disaster categories may have influenced the consistency of generated Q&A pairs.

Across all disaster categories, the Zephyr-RAG system consistently produced longer responses and achieved higher scores in both depth and accuracy compared to the fine-tuned GPT-3.5 Turbo model (Table 2). In six of the seven disaster types, Zephyr-RAG outperformed GPT-3.5 Turbo, often winning four or five out of five questions per category. The only exception was the hurricane/typhoon/cyclone category, where the fine-tuned GPT model achieved a higher win ratio (3:2), suggesting that performance may vary depending on dataset characteristics or domain-specific factors.

On average, the Zephyr chatbot generated substantially longer responses (approximately 53–161 words compared to 14–47 words for the GPT model) and achieved higher evaluation scores, with an average of 3.9 out of 5 question-level wins per category compared to 1.1 for GPT-3.5 Turbo. Additionally, Zephyr-generated responses frequently included direct references to source material, whereas the fine-tuned GPT model did not provide explicit citations. Overall, these results indicate a consistent performance advantage for the retrieval-augmented system while highlighting that differences in performance may depend on the specific disaster domain and underlying dataset quality.

To assess whether the observed performance differences were likely due to chance, a two-sided binomial sign test was applied to the per-question win outcomes across all 35 evaluation questions. Across the full evaluation set, the Zephyr-RAG system won 27 of 35 question-level comparisons, whereas the fine-tuned GPT-3.5 Turbo model won 8 of 35. Under the null hypothesis that both systems were equally likely to win any given question, this difference was statistically significant ($p = 0.0019$). Although the sample size remains limited, this result suggests that the observed advantage of the Zephyr-RAG system is unlikely to be explained by random variation alone within this experimental setup.

Discussion

The results of the study support our hypothesis that the Zephyr model integrated with RAG would outperform the fine-tuned GPT-3.5 Turbo model. The Zephyr-RAG system demonstrated stronger performance in six out of seven categories in this experimental setup; however, this comparison does not

Table 1 Q&A generator outputs by natural disaster type; shows total Q&As generated, number of successful versus unsuccessful Q&As, and pass rate (%) for each scenario after human review.

Natural Disaster Name	Total Generated Q&As	Number of Successful Q&As	Number of Unsuccessful Q&As	Pass Rate %
Subsidence	97	93	4	95.87%
Tornado	659	655	4	99.39%
Hurricane, Typhoon, Cyclone	256	233	23	91.02%
Wildfire	95	94	1	98.94%
Earthquake	135	104	31	77.04%
Volcanic Eruption	136	120	16	88.23%
Flash Flooding	174	141	33	81.03%
Total	1552	1440	112	92.79%

Table 2 DAISY chatbots performance comparison; compares Fine-tuned GPT-3.5 Turbo versus Zephyr-7B Alpha (with RAG) on question length ranges, human-evaluated depth ratio, GPT accuracy ratio, win ratio, and winning model for each scenario.

Natural Disaster Name	Length Range Comparison	Depth (Human) Ratio	Accuracy (GPT) Ratio	Win Ratio	Winning Model
Subsidence	16–51 vs 14–115	1:4	2:3	1:4	Zephyr
Tornadoes	15–56 vs 69–96	2:3	2:3	1:4	Zephyr
Hurricane, Typhoon, Cyclone	10–55 vs 52–135	3:2	2:3	3:2	Tuned
Wildfires	15–42 vs 49–339	1:4	4:1	1:4	Zephyr
Earthquake	13–47 vs 56–101	1:4	0:5	0:5	Zephyr
Volcanic Eruption	18–52 vs 54–164	1:4	1:4	0:5	Zephyr
Flash Flooding	14–27 vs 75–175	2:3	2:3	2:3	Zephyr
Average	14–47 vs 53–161	1.6:3.4	1.9:3.2	1.1:3.9	Zephyr

isolate the effect of retrieval alone, as the two systems differ in model architecture, parameterization, training procedure, and inference-time retrieval. The Zephyr-RAG system was able to retrieve up-to-date external information and provide direct source citations in its answers, which may contribute to improved response quality in this configuration. Access to relevant real-world information may contribute to improvements in accuracy and response depth, as reflected in the higher evaluation scores observed in this configuration. Because the two systems differ in multiple factors—including model architecture, parameterization, training procedure, and the use of retrieval—this study does not isolate the causal effect of Retrieval-Augmented Generation (RAG) alone.

In contrast, the fine-tuned GPT-3.5 Turbo model was constrained to the knowledge contained in the datasets and could not draw on outside sources, which likely limited the depth and completeness of its answers.

Notably, however, the fine-tuned GPT-3.5 Turbo model outperformed the Zephyr-RAG system in the hurricane/typhoon/cyclone category, achieving a higher win ratio (3:2). One possible explanation is that the hurricane dataset may have contained more consistent or structured information, allowing the fine-tuned model to effectively internalize relevant patterns during training. In contrast, the retrieval process in the Zephyr-RAG system may have introduced variability depending on which documents were retrieved at inference time. Additionally, hurricanes are relatively well-documented and frequently studied disasters, which may reduce the advantage of real-time retrieval compared to disaster types with more dynamic or less standardized information.

These results suggest that performance differences between architectures may depend on both dataset characteristics and the nature of the disaster domain. While the Zephyr-RAG system outperformed the fine-tuned model in most categories, this advantage was not universal.

Importantly, given the limited evaluation size of five questions per category, the results simply demonstrate a trend toward improved performance with retrieval-augmented systems in this experimental setup, rather than providing definitive evidence of overall superiority.

The variability in Q&A pass rates between different disaster datasets was likely due to differences in dataset quality. Because all training Q&As were handpicked, human error may have introduced variability in quality. This could explain why certain datasets, such as earthquake and flash flooding, showed lower pass rates. Ensuring more consistent dataset preparation—for example, standardizing the Q&A generation and review criteria—could help minimize inconsistencies in future iterations. Ultimately, these results highlight the importance of utilizing chatbots with real-time information retrieval—RAG, for instance—for tasks like natural disaster information dissemination.

While the results highlight the advantages of retrieval-augmented systems for improving response accuracy and depth, deploying such systems in real-world disaster scenarios presents several challenges. One key concern is the risk of misinformation or outdated data being retrieved from ex-

ternal sources, which could lead to incorrect or potentially harmful guidance. Ensuring the reliability and credibility of retrieved information is therefore critical, particularly in high stakes environments[16]. Additionally, retrieval-based systems may introduce latency, as accessing and processing external data sources can increase response time, which may be problematic during time-sensitive emergencies where immediate answers are required[28]. System reliability is another important consideration, as large-scale disasters may involve high user demand, limited internet connectivity, or disruptions to data sources. Addressing these challenges will require robust data validation pipelines, efficient retrieval mechanisms, and infrastructure capable of maintaining performance under stress.

Future work includes replicating this analysis using newer-generation large language models to assess whether the advantages of RAG persist as base model capabilities continue to evolve. Future work should also incorporate real user testing and simulated emergency scenarios to better assess how the system performs in practical settings. For example, controlled user studies could evaluate how effectively individuals are able to obtain critical information under time pressure, while simulation-based experiments could test system performance under high query volume, incomplete data conditions, or rapidly evolving disaster events. These approaches would provide a more comprehensive understanding of the system's usability, reliability, and real-world impact.

Acknowledgments

Our team, DAISY-AI Developers, thanks Ms. Ping Wang for her support through her NASA-funded HiRISE+AI project at Nova77 STEM Workshop (80NSSC22K0440) and her NSF- funded Planet+AI project at University of Tennessee, Knoxville (DRL- 2314155).

References

1 Center for Disaster Philanthropy, 2024 rio grande do sul brazil floods. 2024, https://disasterphilanthropy.org/disasters/2024-rio-grande-do-sul-brazil-floods/.

2 Federal Emergency Management Agency, New mexico south fork fire and salt fire, dr-4795-nm. 2024, https://www.fema.gov/disaster/4795.

3 J. P. Stimpson, S. M. K. Mita, K. G. Minatoya, A. K. Kuwata, A. K. Kobayashi and K. R. Fink, Monitoring public health during the 2023 Maui wildfire using Google Trends. Qeios. 2026, https://pubmed.ncbi.nlm.nih.gov/41852373/.

4 World Health Organization Regional Office for Europe, Situation reports – Türkiye and Syria earthquakes. 2024, https://www.who.int/europe/emergencies/situations/turkiye-and-syria-earthquakes/situation-reports.

5 American Red Cross, Meet clara, the disaster response chatbot. 2024, https://www.redcross.org/get-help/disaster-relief-and-recovery-services/meet-clara.html.

6 Global Disaster Alert and Coordination System, What is gdacs? 2024, https://gdacs.org/About/overview.aspx.

7 R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora and et al., On the opportunities and risks of foundation models. arXiv. Vol. abs/2108.07258, pg. 1–159, 2021, https://arxiv.org/abs/2108.07258.

8 P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel and D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. Vol. abs/2005.11401, pg. 1–21, 2020, https://arxiv.org/abs/2005.11401.

9 J. B. Houston, J. Hawthorne, M. F. Perreault, E. H. Park, M. G. Hode, M. R. Halliwell, S. E. T. McGowen, R. Davis, S. Vaid, J. A. McElderry and S. A. Griffith, Social media and disasters: a functional framework for social media use in disaster planning, response, and research. Disasters. Vol. 39, pg. 1–22, 2014, https://doi.org/10.1111/disa.12092.

10 C. Reuter and M.-A. Kaufhold, Fifteen years of social media in emergencies: a retrospective review and future directions. Proceedings of the ACM on Human-Computer Interaction. Vol. 2, pg. 1–27, 2018, https://doi.org/10.1145/3209978.

11 L. Palen and K. M. Anderson, Crisis informatics—new data for extraordinary times. Science. Vol. 353, pg. 224–225, 2016, https://doi.org/10.1126/science.aag2579.

12 T. Imran, C. Castillo, J. Lucas, P. Meier and S. Vieweg, AIDR: artificial intelligence for disaster response. Proceedings of the 23rd International World Wide Web Conference. Vol. 2014, pg. 159–162, 2014, https://doi.org/10.1145/2567948.2577034.

13 S. Vieweg, A. L. Hughes, K. Starbird and L. Palen, Microblogging during two natural hazards events: what Twitter may contribute to situational awareness. Proceedings of the SIGCHI Conference. Vol. 2010, pg. 1079–1088, 2010, https://doi.org/10.1145/1753326.1753486.

14 S. Imran, C. Castillo, F. Diaz and S. Vieweg, Processing social media messages in mass emergency: a survey. ACM Computing Surveys. Vol. 47, pg. 1–38, 2015, https://doi.org/10.1145/2771588.

15 F. Xu, J. Ma, N. Li and J. C. P. Cheng, Large language model applications in disaster management: an interdisciplinary review. International Journal of Disaster Risk Reduction. Vol. 127, pg. 105642, 2025, https://doi.org/10.1016/j.ijdrr.2025.105642.

16 S. Farquhar, J. Kossen, L. Kuhn and Y. Gal, Detecting hallucinations in large language models using semantic entropy. Nature. Vol. 630, pg. 625–630, 2024, https://doi.org/10.1038/s41586-024-07421-0.

17 J. Maynez, S. Narayan, B. Bohnet and R. McDonald, On faithfulness and factuality in abstractive summarization. Proceedings of ACL. Vol. 2020, pg. 1906–1919, 2020, https://doi.org/10.18653/v1/2020.acl-main.173.

18 Federal Emergency Management Agency, Current disaster declarations. 2024, https://www.fema.gov/disaster/current.

19 U.S. Geological Survey, Core science systems mission area. 2024, https://www.usgs.gov/mission-areas/core-science-systems.

20 United Nations Office for Disaster Risk Reduction, Comprehensive disaster and climate risk management. 2024, https://www.undrr.org/climate-action-and-disaster-risk-reduction/comprehensive-disaster-and-climate-risk-management.

21 World Health Organization, Publications. 2024, https://www.who.int/publications.

22 Global Disaster Alert and Coordination System, GDACS XML archive. 2024, https://www.gdacs.org/contentdata/xml/.

23 OpenAI, GPT-3.5 turbo fine-tuning and API updates. 2023, https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/.
24 L. Tunstall, L. von Werra, N. Lambert and A. Rush, Zephyr: direct distillation of LM alignment. arXiv. Vol. arXiv:2310.16944, pg. N/A, 2023, https://arxiv.org/abs/2310.16944.
25 H. Chase, LangChain. 2022, https://github.com/langchain-ai/langchain.
26 Chroma Core, Chroma: open-source search and retrieval database for AI applications. 2024, https://github.com/chroma-core/chroma.
27 A. Abid, J. Zhang, A. Bagaria and J. Zou, Gradio: hassle-free sharing and testing of ML models in the wild. arXiv. 2019, https://arxiv.org/abs/1906.02569.
28 V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen and W.-T. Yih, Dense passage retrieval for open-domain question answering. arXiv. Vol. abs/2004.04906, pg. 1–24, 2020, https://arxiv.org/abs/2004.04906.

Appendix

To evaluate the performance of the DAISY chatbot systems, a standardized set of 35 evaluation questions was constructed, consisting of five questions for each of the seven natural disaster categories (earthquakes, hurricanes, wildfires, tornadoes, floods, volcanic eruptions, and subsidence).

The questions were designed to reflect two primary categories:

1. Conceptual Understanding Questions

These questions assess whether the chatbot can provide clear, accurate definitions and explanations of disaster phenomena.

2. Action-Oriented Questions

These questions evaluate the chatbot's ability to provide practical, safety-critical guidance that a user might seek during a real-world disaster scenario.

Questions were developed based on common public information needs identified in disaster communication literature, typical queries observed in emergency information platforms (e.g., evacuation guidance, safety procedures), and the structure of the training dataset, which included both descriptive and applied disaster information.

To ensure consistency across categories, the same five question templates were adapted to each disaster type.

The evaluation questions were manually reviewed to ensure: (1) Clarity (questions are easily understandable by a general audience), (2) relevance (questions reflect realistic disaster-related information needs), and (3) non-overlap with training data phrasing (questions were written independently and not copied directly from generated Q&A pairs to reduce data leakage risk)

To further reduce the risk of data leakage, training and evaluation were separated at the document level. Documents used to generate fine-tuning question–answer pairs were not reused during evaluation, and evaluation questions were designed to avoid direct overlap with training content. While both systems draw from similar underlying domains, this separation helps ensure that evaluation performance reflects generalization rather than memorization of specific training examples.

Additionally, each question was checked to ensure it could be answered using credible disaster information sources (e.g., FEMA, USGS, WHO).

Below is the list of 35 questions that were asked, divided into each natural disaster type.

Earthquake

1. What is an earthquake and what causes it?
2. What are the primary hazards associated with earthquakes?
3. What should you do during an earthquake to stay safe?
4. What should you do immediately after an earthquake?
5. How can earthquakes be monitored or predicted?

Hurricanes/Typhoons/Cyclones

6. What is a hurricane and how does it form?
7. What are the main dangers associated with hurricanes?
8. What should individuals do to prepare before a hurricane?
9. What should you do during a hurricane?
10. What actions should be taken after a hurricane passes?

Wildfires

11. What is a wildfire and how does it start?
12. What environmental conditions increase wildfire risk?
13. What should you do if a wildfire is approaching your area?
14. How can individuals reduce wildfire risk around their homes?
15. What are the health impacts of wildfire smoke exposure?

Tornadoes

16. What is a tornado and how does it form?
17. What are the warning signs of a tornado?
18. What should you do during a tornado?
19. Where is the safest place to take shelter during a tornado?
20. What should you do after a tornado has passed?

Floods/Flash Flooding

21. What causes floods and flash floods?
22. What are the dangers of flash flooding?
23. What should you do if you receive a flood warning?
24. Why is it dangerous to drive through floodwaters?
25. What precautions should be taken after a flood?

Volcanic Eruptions

26. What is a volcanic eruption and what causes it?
27. What hazards are associated with volcanic eruptions?
28. What should you do during a volcanic eruption?
29. How does volcanic ash affect health and infrastructure?
30. What should people do after an eruption?

Subsidence

31. What is land subsidence and what causes it?
32. What are the risks associated with subsidence?
33. How can subsidence impact infrastructure and communities?
34. What are warning signs of potential subsidence?
35. What mitigation or prevention measures can reduce subsidence risk?