Evaluating Robustness of LLMs on Crisis-Related Microblogs across Events, Information Types, and Linguistic Features

Muhammad Imran
Hamad Bin Khalifa University
Qatar Computing Research Institute
Doha, Qatar
mimran@hbku.edu.qa

Abdul Wahab Ziaullah
Hamad Bin Khalifa University
Qatar Computing Research Institute
Doha, Qatar
awahab@hbku.edu.qa

Kai Chen
OpenAI
San Francisco, CA, USA
kaichen@openai.com

Ferda Ofli
Hamad Bin Khalifa University
Qatar Computing Research Institute
Doha, Qatar
fofli@hbku.edu.qa

Abstract

The widespread use of microblogging platforms like X (formerly Twitter) during disasters provides real-time information to governments and response authorities. However, the data from these platforms is often noisy, requiring automated methods to filter relevant information. Traditionally, supervised machine learning models have been used, but they lack generalizability. In contrast, Large Language Models (LLMs) show better capabilities in understanding and processing natural language out of the box. This paper provides a detailed analysis of the performance of six well-known LLMs in processing disaster-related social media data from a large-set of real-world events. Our findings indicate that while LLMs, particularly GPT-4o and GPT-4, offer better generalizability across different disasters and information types, most LLMs face challenges in processing flood-related data, show minimal improvement despite the provision of examples (i.e., shots), and struggle to identify critical information categories like urgent requests and needs. Additionally, we examine how various linguistic features affect model performance and highlight LLMs' vulnerabilities against certain features like typos. Lastly, we provide benchmarking results for all events across both zero- and few-shot settings and observe that proprietary models outperform open-source ones in all tasks.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → **Empirical studies in collaborative and social computing**.

WWW '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1274-6/25/04
https://doi.org/10.1145/3696410.3714511

Keywords

Large language models; social media; disaster response; LLM evaluation; LLM benchmarking

ACM Reference Format:
Muhammad Imran, Abdul Wahab Ziaullah, Kai Chen, and Ferda Ofli. 2025. Evaluating Robustness of LLMs on Crisis-Related Microblogs across Events, Information Types, and Linguistic Features. In *Proceedings of the ACM Web Conference 2025 (WWW '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3696410.3714511

1 Introduction

Microblogging platforms like X (formerly Twitter) are vital during large-scale disasters [30]. They facilitate real-time communication for the public to share firsthand experiences, report damage to infrastructure, and most importantly, seek assistance [2, 23]. Moreover, local governments are increasingly leveraging these non-traditional data sources to enhance their situational awareness and quickly identify humanitarian needs, and inform their response strategies accordingly [20, 28].

Despite their accessibility, data from social media platforms are often highly noisy [13]. During large-scale disasters, the volume of messages can reach millions per day, filled with irrelevant content and chatter [6]. This deluge makes it challenging for local authorities to identify reports critical for humanitarian response. Previous research has addressed this issue by developing supervised machine learning models that filter through the raw data to identify relevant information [16, 24]. However, these models typically struggle with generalizability across different disasters or geographic locations due to the problems of domain shift [14, 19]. Techniques like domain adaptation or transfer learning have been proposed to alleviate these challenges [11, 22]. Nonetheless, when the categories of interest change, training new machine learning models becomes necessary. This process requires human-labeled data, which is time-intensive, and can slow down response efforts.

Large Language Models (LLMs) demonstrate a strong capability to comprehend natural language and generalize across various NLP tasks [32]. Despite numerous studies assessing LLMs' effectiveness with well-structured web data [34] and noisy social media content, mainly in non-humanitarian context [15, 33], no previous

research presents a thorough analysis of their robustness in processing disaster-related social media data. In this paper, we present a comprehensive analysis of social media data collected from 19 major disasters across multiple countries using six well-known LLMs, including GPT-3.5 [5], GPT-4 [1], GPT-4o [25], Llama-2 13B [29], Llama-3 8B [10], and Mistral 7B [18]. We assess the effectiveness of these proprietary and open-source LLMs in handling different disaster types and information categories, and their performance with data from both native and non-native English-speaking countries. We also examine how various linguistic features influence LLMs' performance. Additionally, our study provides benchmarking results for each of the 19 disaster events and evaluates the overall model performance in both zero- and few-shot settings.

Our findings indicate that proprietary models (i.e., GPT-4 and GPT-4o) generally outperform open-source models (i.e., Llama-2 13B, Llama-3 8B, and Mistral 7B) on various tasks. However, GPT models notably struggle with processing data from flood incidents. Moreover, certain information types, such as *requests or urgent needs*, consistently challenge all models, with all GPTs achieving F1 below 0.60. Open-source models also display weaknesses in handling classes like *caution and advice* and *requests or urgent needs*. Additionally, we find that providing models with class-specific examples does not generally enhance their performance.

The rest of the paper is organized as follows. We summarize the related work in Section 2, describe our assessment methodology in Section 3, and present results and discussions in Section 4. Finally, we conclude the paper and provide a future work plan in Section 6.

2 Related Work

In crisis informatics literature, several studies introduced large-scale crisis-related microblog datasets and presented baseline results using both classical machine learning algorithms (e.g., Random Forest, Support Vector Machines, etc.) as well as deep learning models (e.g., RNNs, LSTMs, CNNs, etc.) [17, 24]. Later, researchers undertook an effort to consolidate available datasets and tasks for benchmarking transformer-based models such as BERT [9], DistilBERT [27] and RoBERTa [21], and showed that the transformer-based models typically outperform [4]. A more comprehensive crisis-related dataset along with benchmarking results were presented in [3]. Likewise, a more recent study [31] presented a BERT-based ensemble model, FF-BERT, for the classification of flash flooding messages. Their evaluations examined various BERT-based ensemble models on a specially curated dataset of 21,180 paragraphs of text. Meanwhile, [12] developed QuakeBERT and showed better performance to assess physical and social impacts of an earthquake through microblogs.

Previous research has shown that transformer-based models outperform traditional ML algorithms on various metrics. Recent efforts have focused on using more powerful LLMs across diverse fields and tasks. For instance, LLMeBench has assessed LLMs on multiple NLP tasks such as sentiment analysis and summarization [8]. Additionally, studies like [35] have applied LLMs to crisis-related tasks, evaluating models like Mistral 7B [18] for their ability to analyze disaster-related tweets. Further, Llama-2 and Mistral have been fine-tuned for disaster response guidance, as presented in [26]. This paper builds upon these findings by analyzing LLMs on a crisis-related dataset, exploring LLMs' performance across various disaster types, information types, and the linguistic features of the messages, to identify their capabilities and weaknesses.

3 Assessment Methodology

The increasing complexity and frequency of natural disasters worldwide necessitate AI models, particularly LLMs, that can effectively generalize across various types of disasters (e.g., floods, earthquakes, etc.), languages (English vs. Non-English), and different types of information shared on social media (e.g., warnings, urgent needs, damage reports, etc.). We evaluate the performance of LLMs in processing social media content from different types of disasters in countries that use different languages. Additionally, we investigate how open-source and proprietary models differ in performance and assess the role of few-shot learning, where LLMs are provided with examples, on their effectiveness. For this purpose, this paper addresses the following four key questions.

(1) How do LLMs perform for different types of natural disasters (e.g., floods, wildfires)?
(2) What is LLMs' ability to interpret various types of social media data during disasters?
(3) How do LLMs perform for countries where the native language is not English?
(4) Do certain linguistic features or sentence structures significantly impact LLMs performance?

3.1 Dataset and Models

To answer our research questions, we utilize HumAID [3] dataset comprising 77,196 tweets from 19 different natural disasters that occurred in 11 distinct countries (3 native English-speaking and 8 non-English-speaking) between 2016 to 2019. The tweets in the dataset are labeled by paid crowdsourcing workers into ten distinct information categories (acronyms): *(1) caution and advice (CA)—(2) sympathy and support (SS)—(3) requests or urgent needs (RUN)—(4) displaced people and evacuations (DPE)—(5) injured or dead people (IDP)—(6) missing or found people (MFP)—(7) infrastructure and utility damage (IUD)—(8) rescue volunteering or donation effort (RVDE)—(9) other relevant information (ORI)—* and *(10) not humanitarian (NH)*. We drop the *"other relevant information"* class from our analysis as it mainly contains general event-related information that does not belong to other categories. The dataset is already split into train, development, and test sets. We use the test split (N=15,160) for our experiments. Figure 1 shows various distributions of our dataset.

Models. We select six well-known LLMs (three proprietary and three open-source) for this study. We choose GPT-3.5 [5], GPT-4 [1], and GPT-4o [25] from OpenAI as our proprietary models and Llama-2 13B [29], Llama-3 8B [10] and Mistral 7B [18] as our open-source models. All six models are known for their language understanding capabilities across various NLP tasks.

3.2 Experimental Design

We evaluate the LLMs for the classification task in two settings: zero-shot and few-shot. In the zero-shot setting, the models operate without any class-specific examples, relying solely on their pre-trained capabilities to perform the task. In the few-shot setting,

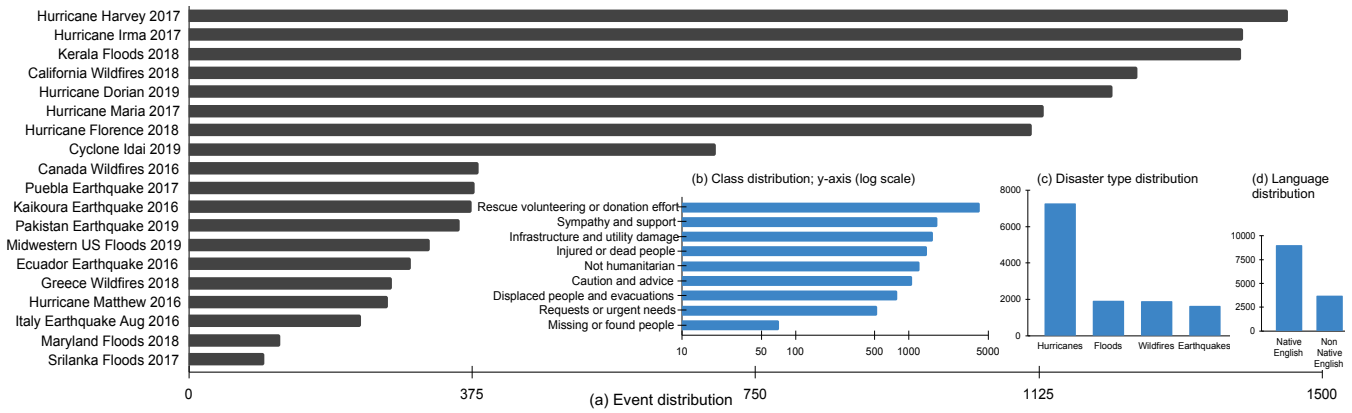
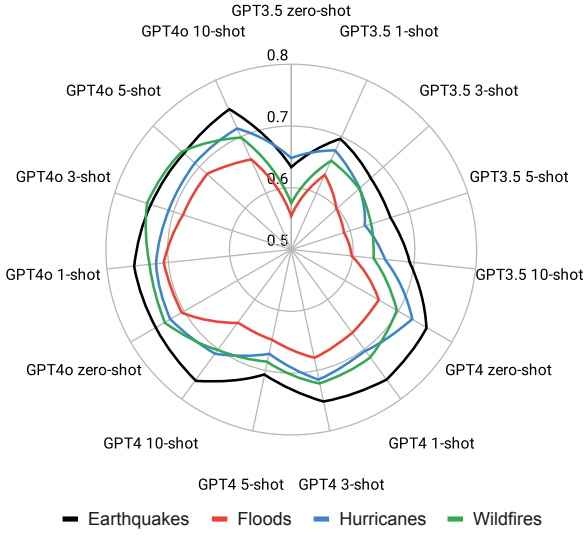
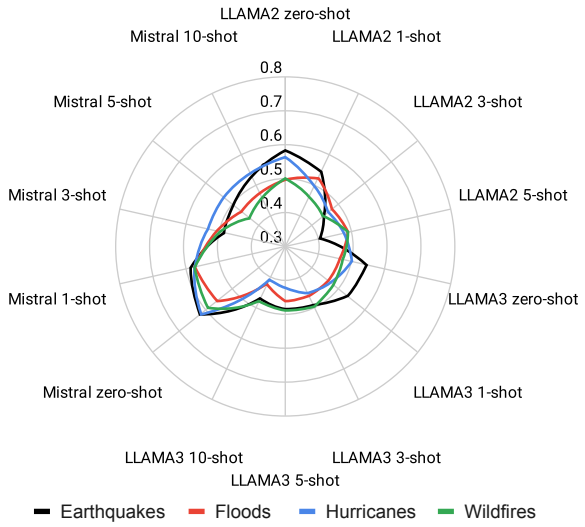


Figure 1: Data distributions for (a) events, (b) information types, (c) disaster types, and (d) native/non-native English countries



(a) Proprietary models



(b) Open-source models

Figure 2: Performance (F1-scores) of LLMs across disaster types and few-shot settings

models receive examples for each class to improve class-specific performance. For instance, in a three-shot experiment, we provide the model with three carefully selected tweets per class from the training set, totaling 30 examples for ten classes. For all experiments, we set the temperature parameter to zero. We use the following prompt across all experiments, except for Llama-2 and Mistral, where we provide additional instructions to control for verbosity.

Prompt: "Read the category names and their definitions below, then classify the following tweet into the appropriate category. In your response, mention only the category name.

Category name: category definition

- Caution and advice: Reports of warnings issued or lifted, guidance and tips related to the disaster.
- Sympathy and support: Tweets with prayers, thoughts, and emotional support.
- Requests or urgent needs: Reports of urgent needs or supplies such as food, water, clothing, money,...
- Displaced people and evacuations: People who have relocated due to the crisis, even for a short time...
- Injured or dead people: Reports of injured or dead people due to the disaster.
- Missing or found people: Reports of missing or found people due to the disaster.
- Infrastructure and utility damage: Reports of any type of damage to infrastructure such as buildings, houses,...
- Rescue volunteering or donation effort: Reports of any type of rescue, volunteering, or donation efforts...
- Not humanitarian: If the tweet does not convey humanitarian aid-related information."

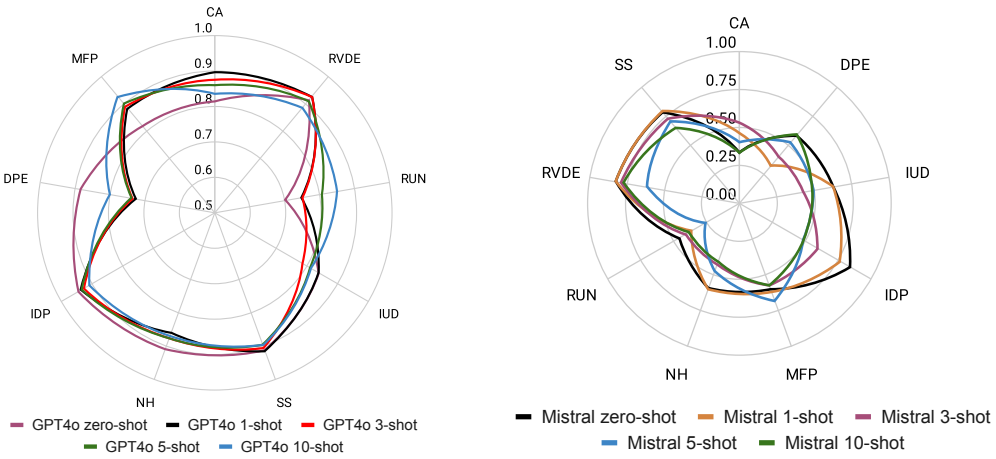
Tweet: {input tweet}

Category:

4 Results and Discussion

4.1 Disaster Type Analysis

Our first research question examines how LLMs perform across different types of disasters. We analyze data from 19 events, grouped into four event types: 5 earthquakes, 7 hurricanes, 3 wildfires, and 4 floods. We present results for both proprietary and open-source models and compare their performance in zero-shot and few-shot (i.e., 1, 3, 5, and 10) settings.



(a) GPT-4o in zero- and few-shot settings **(b) Mistral 7B in zero- and few-shot settings**

Figure 3: Performance (F1 scores) of LLMs across various information types (i.e., classes)

Figure 2(a) shows the macro F1-scores for GPT-3.5, GPT-4, and GPT-4o across various few-shot settings. Notably, all models consistently show high performance for earthquakes, with GPT-4 achieving a maximum F1-score of 0.76 in the 10-shot setting and GPT-3.5 a minimum of 0.63 in the zero-shot setting. Conversely, model performances for floods consistently remain the lowest, with GPT-4o's 1-shot performance reaching the highest F1-score of 0.70, and GPT-3.5's zero-shot the lowest at 0.55. The results for wildfires and hurricanes are less consistent, though GPT-4o outperforms GPT-4 and GPT-3.5 in most cases. Surprisingly, increasing the number of shots does not show plausible performance improvements for all models. For GPT-3.5, there is a noticeable improvement from the zero-shot to other few-shot settings. However, for GPT-4, the performance from zero-shot to 3-shot remains nearly unchanged, and unexpectedly degrades in the 5-shot setting, and then recovers in 10-shot. Similarly, GPT-4o does not exhibit a consistent improvement as the number of shots increases.

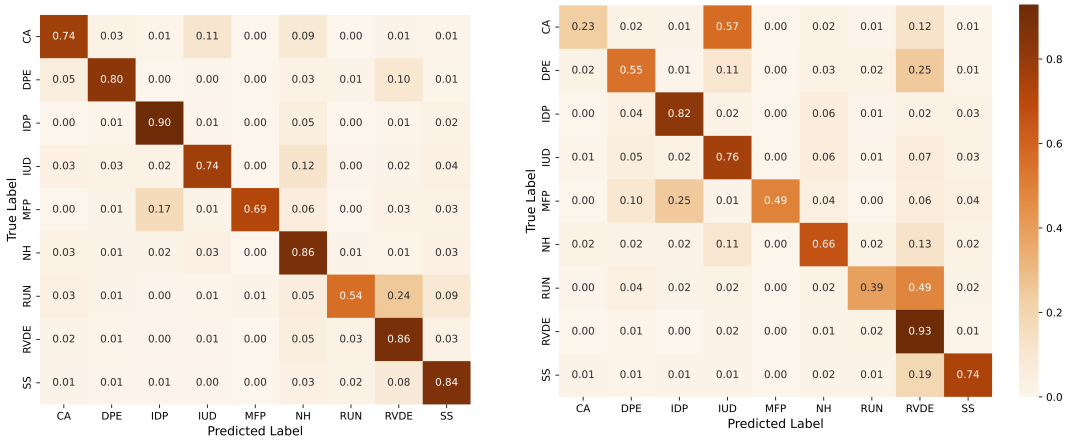
Figure 2(b) presents the F1-scores for the Llama-2 13B, Llama-3 8B, and Mistral 7B models across various few-shot settings, excluding the Llama-2 10-shot due to token limit constraints. Overall, these open-source models perform less effectively than their proprietary counterparts. Specifically, Mistral's zero-shot achieves the highest F1-score of 0.62 for earthquakes and also shows similar results for hurricanes. Mistral consistently outperforms Llama-2 and Llama-3 across most cases. A notable observation is that the zero-shot setting generally yields the best results for both models, and adding more example shots does not significantly enhance performance. Overall, we observe that the open-source models tend to perform better for hurricanes as opposed to the proprietary models' superior performance for earthquakes.

4.2 Information Type Analysis

Our second research question examines LLMs' capabilities in processing diverse types of information related to humanitarian response and situational awareness during disasters. Our analysis contains nine distinct information categories, detailed in Section 3.1, with their cumulative distribution across all events depicted in Figure 1(b). While we conducted experiments across all six models in all few-shot settings (except for Llama-2 10-shot), the following results focus solely on the two top-performing models, GPT-4o and Mistral, from the proprietary and open-source categories, respectively. The complete set of results, including all six models, are provided in Appendix A.

Figure 3(a) shows the class-wise macro F1-scores for GPT-4o models across all few-shot settings. These models consistently achieve F1-scores above 0.80 in all few-shot settings for the classes *rescue volunteering or donation effort* (RVDE), *sympathy and support* (SS) and *injured or dead people* (IDP). In contrast, the *requests or urgent needs* (RUN) and *displaced people and evacuations* (DPE) classes consistently yield low performance, with F1-scores below 0.75, except for higher shots (i.e., 5 and 10). Notably, the *requests or urgent needs* (RUN) class exhibits significant variability in performance across different shots. To understand why certain classes underperformed, we conducted an error analysis using the confusion matrix shown in Figure 4(a). We specifically examined random samples from the *requests or urgent needs* (RUN) class which are confused with the *rescue volunteering or donations effort* (RVDE) class. Our analysis revealed that the model often confused general calls for volunteering and donations with ongoing volunteering efforts. This confusion led to a high rate of misclassification of tweets from *requests or urgent needs* (RUN) as *rescue volunteering or donation effort* (RVDE) (24%), as shown in Figure 4(a).

Figure 3(b) presents the class-wise F1-scores of Mistral 7B across all few-shot settings. Mistral 7B notably under performs in the categories *requests or urgent needs* (RUE) and *caution and advice* (CA). Other instances of low performance include *displaced people and evacuations* (DPE) in the 1-shot setting (F1=0.32), *not humanitarian* (NH) in the 3-shot (F1=0.29), and most critically, *requests or urgent needs* (RUE) in the 5-shot (F1=0.17). However, the model performs relatively well with *rescue volunteering or donation effort* (RVDE) and *injured or dead people* (IDP), especially in zero- and 1-shot scenarios. Overall, this open-source model lags behind its



(a) GPT-4o (b) Mistral 7B

Figure 4: Confusion matrices for GPT-4o (left) and Mistral 7B (right) models under the zero-shot setting

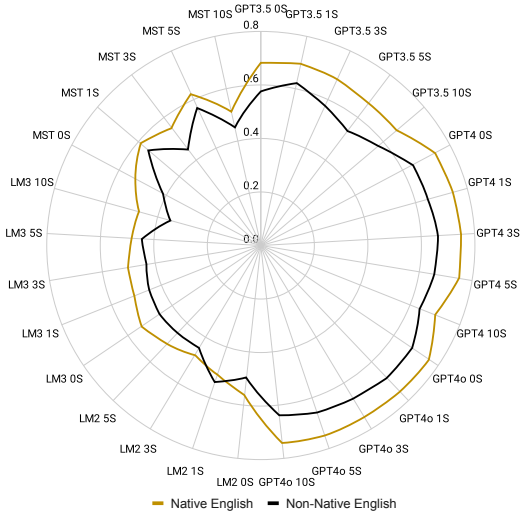


Figure 5: Performance (F1-scores) of LLMs on native-English-speaking vs. non-English-speaking countries. LM2=Llama-2 13B, LM3=Llama-3 8B, MST=Mistral 7B

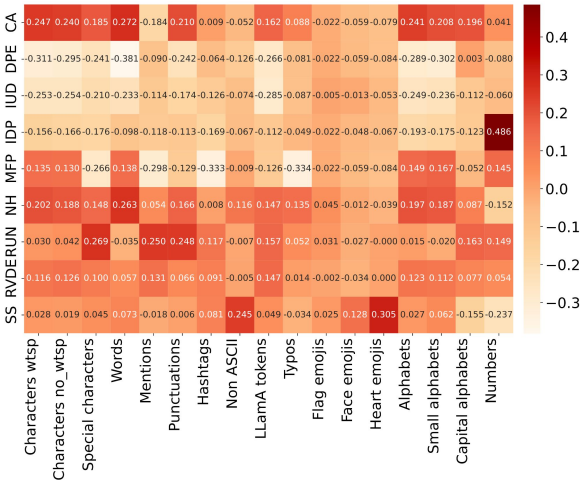


Figure 6: Distribution of z-scores of language features (x-axis) across information classes (y-axis)

proprietary counterpart in information type classification performance. Figure 4(b) shows the confusion matrix of Mistral 7B 0-shot, which we used to perform an error analysis of mistakes made by the model. We observed that open-source models also confuse *requests or urgent needs* (RUE) with *rescue volunteering or donations effort* (RVDE) due to the same reasoning where calls for volunteering or donations were mistaken with the efforts for volunteering or donations. Additionally, we analyzed errors made by the Mistral zero-shot model in classifying *caution and advice* (CA) tweets. We found that the presence of intensity descriptors such as "severe earthquakes" or "category 5 hurricane" led the model to mistakenly label tweets as *infrastructure or utility damage* (IUD). The results from all the models are provided in Appendix A.

4.3 Native vs. Non-Native English Analysis

Our third research question examines the performance of LLMs in processing social media content from native-English-speaking versus non-English-speaking countries. Our dataset includes 11 events from native-English-speaking countries and 8 events from non-English-speaking countries. Figure 1(c) illustrates the distribution of tweets across these two categories.

Figure 5 displays the F1-scores for all models, including both proprietary and open-source. It is evident that all models achieve better performance in processing data from native-English-speaking countries. Proprietary models show a marked advantage in understanding data from these regions across all few-shot settings. However,

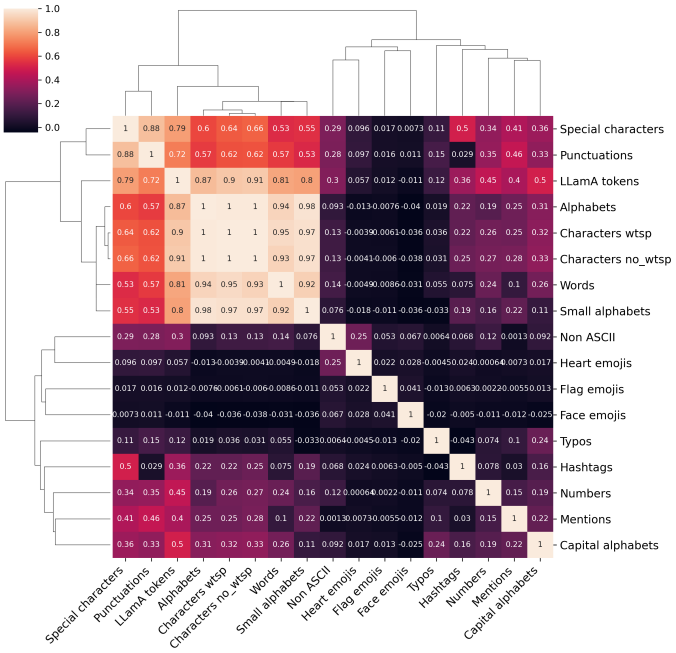


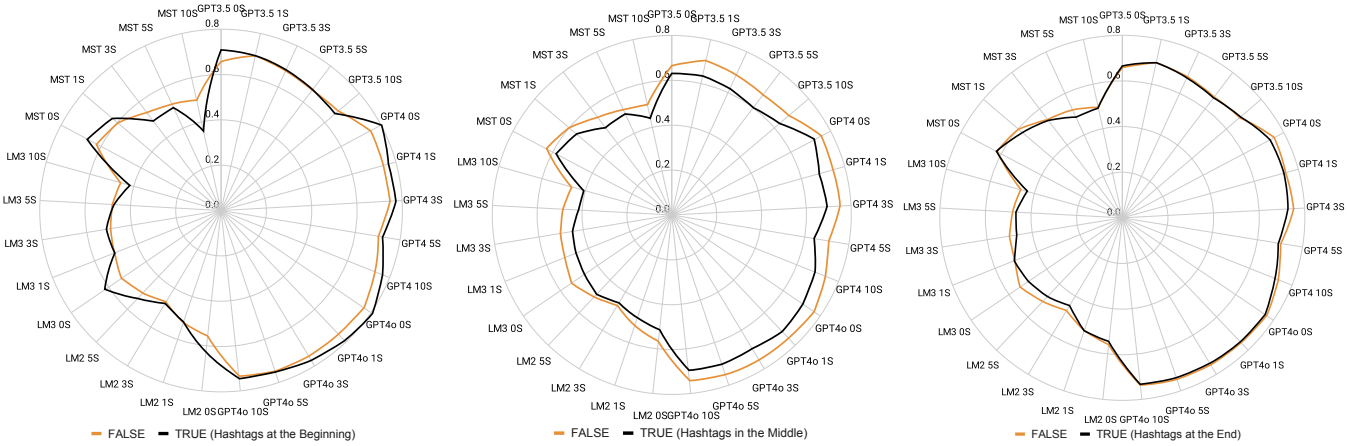
Figure 7: Multicollinearity analysis of linguistic features

Table 1: Logistic regression analysis for Mistral-7B zero-shot

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.3295	0.049	27.288	0.000	1.234	1.425
Typos	-0.0457	0.029	-1.586	0.113	-0.102	0.011
Special characters	-0.0146	0.009	-1.691	0.091	-0.032	0.002
Characters	-0.0037	0.000	-9.562	0.000	-0.004	-0.003
Numbers	0.0239	0.006	3.737	0.000	0.011	0.036
Hashtags	0.0308	0.013	2.291	0.022	0.004	0.057
Mentions	0.0702	0.017	4.185	0.000	0.037	0.103
Face emojis	0.1462	0.224	0.651	0.515	-0.294	0.586
Heart emojis	0.2719	0.143	1.905	0.057	-0.008	0.552

Table 2: Logistic regression analysis for GPT-4o zero-shot

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.6471	0.055	29.719	0.000	1.538	1.756
Typos	-0.0575	0.033	-1.768	0.077	-0.121	0.006
Special characters	-0.0115	0.010	-1.171	0.241	-0.031	0.008
Characters	-0.0017	0.000	-3.974	0.000	-0.003	-0.001
Numbers	0.0125	0.007	1.734	0.083	-0.002	0.027
Hashtags	0.0335	0.016	2.144	0.032	0.003	0.064
Mentions	0.0287	0.018	1.590	0.112	-0.007	0.064
Face emojis	0.0996	0.254	0.391	0.695	-0.399	0.598
Heart emojis	0.2166	0.164	1.324	0.185	-0.104	0.537



(a) Hashtags strictly at the beginning **(b) Hashtags in the middle** **(c) Hashtags strictly at the end**

Figure 8: Impacts of hashtag positioning on LLMs performance (macro F1-scores). LM2=Llama-2 13B, LM3=Llama-3 8B, MST=Mistral 7B

their performance drops when processing data from non-English-speaking countries, although they still outperform open-source models for the same category. Furthermore, in the non-English-speaking category, GPT-4o zero-shot setting leads with an F1-score of 0.76, while Mistral in the 5-shot setting tops among open-source models with an F1-score of 0.62. The remaining open-source models generally score below 0.60, which is a surprising finding.

4.4 Linguistic Feature Analysis

Our fourth research question explores whether various linguistic features, such as word count, hashtag count, and emoji usage in tweets, affect the performance of LLMs. Previous studies have shown that such features significantly influence the performance of traditional machine learning and deep learning models [7]. We aim to determine if this holds true for LLMs, as well. We defined 17 linguistic features and analyzed their frequency distributions across all classes. Figure 6 presents a heatmap of z-scores for these features' presence in each class, revealing notable patterns. For instance, the

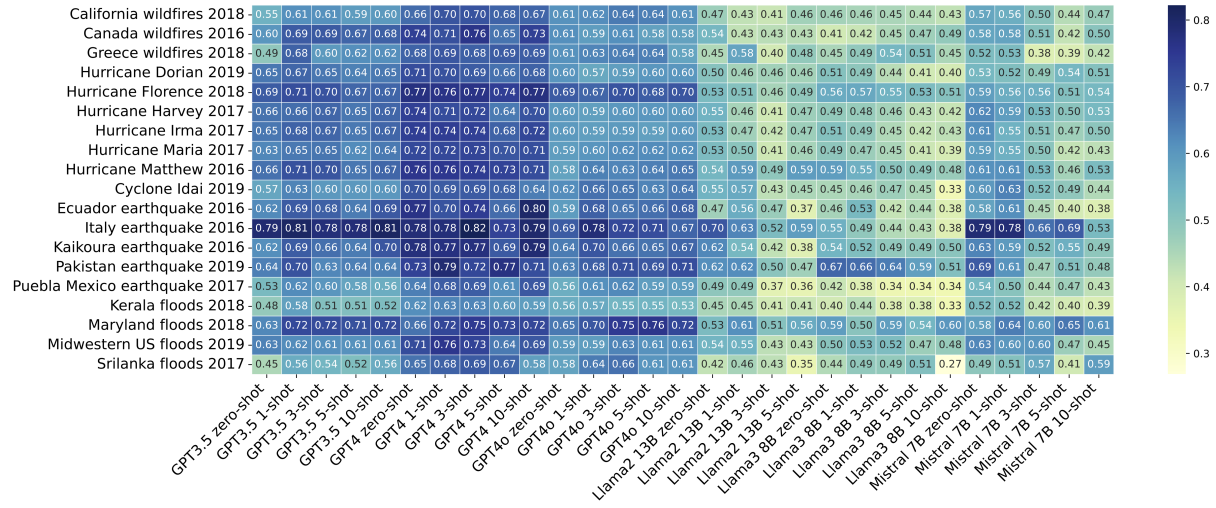


Figure 9: F1-scores for all proprietary and open-source models for 19 events across all k-shot settings

injured and dead people (IDP) class has a high z-score for numbers, likely due to the prevalence of numerical data in such messages reporting casualties or injured people due to the disaster event. Similarly, the *sympathy and support* (SS) class shows a high value for heart emojis, reflecting emotional expressions in such tweets. We observed that tweets discussing *requests and urgent needs* often include more mentions of other user accounts, particularly NGOs and official accounts.

Next, we perform a logistic regression analysis to ascertain how different linguistic features affect model performance. To avoid the undesirable effects of multicollinearity, we exclude highly correlated linguistic features like character, word, and alphabet counts as illustrated in Figure 7 and work with a reduced set of features as our independent variables and consider the binary (correct/incorrect) validation of the predicted class labels with the ground truth as our dependent variable. Table 1 summarizes the analysis results for Mistral zero-shot. We see that numbers, hashtags, mentions, face, and heart emojis have positive correlations with model performance whereas character, special character and typo counts have the opposite effect. For example,—with a relatively small but statistically significant coefficient—, increasing character counts tends to negatively impact model performance. On the contrary,—again with relatively small but statistically significant coefficients—, number, hashtag, and mention counts play positive roles in improving predictive performance. According to Table 2, these observations also hold for GPT-4o zero-shot model with the exception that coefficients for number and hashtag counts are not statistically significant as before. In both cases, it is notable that face and heart emojis have relatively larger coefficients and show more prominence in the regression analysis but lack statistical significance.

Hashtag Positioning Impacts. Next, we investigate whether the placement of hashtags within messages affects LLM performance. We categorize messages into three groups: *(i)* messages with hashtags only (strictly) at the beginning, *(ii)* messages with hashtags in the middle, and *(iii)* hashtags only (strictly) at the end. Figure 8 presents F1-scores for each scenario in separate radar charts.

Interestingly, we observed that hashtags placed in the middle of messages frequently result in higher error rates, as illustrated in Figure 8(b) with the brown circle. Most probably, these errors stem from the disruption in sentence structure caused by mid-sentence hashtags, which can confuse models by introducing unexpected breaks or context shifts. This phenomenon appears more pronounced in proprietary models compared to open-source models. Specifically, proprietary models such as GPT-3.5 (in all shot settings except zero-shot), GPT-4, and GPT-4o consistently exhibited difficulties in accurately interpreting messages with mid-sentence hashtags. The models often misclassify or overlook critical context surrounding the hashtag, leading to erroneous predictions. Additionally, there is a notable difference in performance for GPT-4, GPT-4o, and Llama-3 zero-shot and Mistral 10-shot configurations when hashtags are exclusively at the beginning of messages. Conversely, the positioning of hashtags at the end of messages does not significantly affect LLMs' performance.

4.5 Event-wise and Overall Performance

In addition to our main analyses addressing the specified research questions, we conduct two additional experiments to assess LLMs' performance for individual events and their overall performance. These results also help benchmark the LLMs against this dataset.

Figure 9 shows the event-wise F1-scores for both proprietary and open-source models across various few-shot settings. Notably, the proprietary GPT models consistently outperform the open-source Llama and Mistral models. Among the proprietary models, all few-shot configurations of GPT-4 yield superior results compared to any few-shot setting of GPT-3.5 and GPT-4o. Specifically, GPT-4's zero-shot and 3-shot settings perform comparably and exceed the performance of its 5-shot and 10-shot settings. Interestingly, GPT-4o appears to face challenges in this experiment, particularly with hurricane and wildfire events. However, for earthquakes, GPT-4o's

performance is comparable to or slightly below that of GPT-4. For GPT-3.5, the one-shot variant stands out as the most effective across the majority of events.

The event-wise results for open-source models (Figure 9) highlight Mistral's zero- and one-shot settings as the most effective. In most cases, adding examples—increasing the number of shots—does not typically enhance the model's performance. Notably, larger number of shots, such as 5- or 10-shot, introduce additional tokens to the prompt, which may actually confuse the model rather than help it. However, both Llama 2 and 3 showed underwhelming performance across most events, with the exception of a few earthquake cases. In some instances, the Llama models scored as low as 0.27 (Sri Lanka floods) and 0.33 (Cyclone Idai).

Next, we evaluate the overall performance of LLMs on the entire data, including all events, information types, and language variations. Figure 10 shows the F1-scores for both proprietary and open-source models across all shots and the SOTA supervised baseline (i.e., RoBERTa F1=0.78) as we report in [3]. It is clear that GPT-4o consistently outperforms all other LLMs in all configurations, though it does not outperform the baseline. GPT-4 ranks as the second-best overall, while Llama-2 and Mistral generally underperform across all shots. Notably, there is no consistent trend in performance with the addition of more shots, with the exception of specific instances such as GPT-3.5's progression from zero to various few-shots, and Llama-2's improvement from 3- to 5-shot settings. We summarize the experimental results, including accuracy, precision, and recall of all the models across all shots in Table 3.

5 Ethical Considerations

The datasets used in this study consist of publicly available tweets posted by individuals or organizations during various natural disasters. The data was collected in strict adherence to the terms and conditions set forth by the Twitter (now X) API to ensure ethical compliance. To safeguard individuals' privacy, any personally identifiable information, including names, addresses, phone numbers, or other sensitive details, was systematically anonymized before data processing. Moreover, no attempts were made to infer or store additional demographic or personal information about the users.

6 Conclusion and Future Work

We presented a comprehensive evaluation of prominent large language models in processing social media data from 19 major natural disasters across 11 countries, including 8 non-native and 3 native English-speaking regions. Our findings highlight varying strengths and limitations of LLMs in managing diverse disaster types, information categories, and linguistic complexities. Specifically, the models demonstrated notable difficulties with flood-related data and frequently misclassified critical information categories such as *requests and urgent needs* and *caution and advice*. Furthermore, our analysis identified key factors such as message length, typographical errors, and the presence of special characters as significant challenges that impair model performance. Importantly, we observed that providing few-shot examples yielded limited performance gains for most models. This could be due to the high variability in social media content, even from the same class. Finally, we provided benchmarking results, aiming to inform further research into LLMs' vulnerabilities and assist in developing more robust models for disaster information processing.

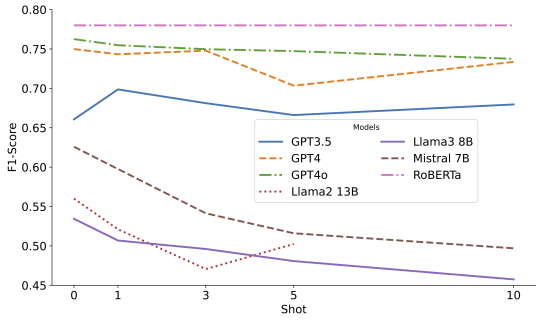


Figure 10: Overall performance of proprietary and open-source models across k-shot settings (k={0, 1, 3, 5, 10}) and RoBERTa (F1=0.78) as a supervised baseline [3].

Table 3: Comparison of LLMs' performance in terms of F1-score, Accuracy, Precision, and Recall

# Shots	LLM Model	F1-score	Accuracy	Precision	Recall
0-shot	GPT-4	0.750	0.785	0.764	0.747
	GPT-4o	0.762	0.801	0.771	0.760
	GPT-3.5	0.661	0.686	0.729	0.644
	Llama-2 13B	0.562	0.554	0.694	0.522
	Llama-3 8B	0.534	0.540	0.621	0.547
	Mistral 7B	0.628	0.697	0.732	0.582
1-shot	GPT-4	0.743	0.769	0.777	0.728
	GPT-4o	0.755	0.80	0.763	0.760
	GPT-3.5	0.699	0.748	0.733	0.675
	Llama-2 13B	0.522	0.522	0.655	0.559
	Llama-3 8B	0.507	0.520	0.603	0.532
	Mistral 7B	0.598	0.682	0.702	0.563
3-shot	GPT-4	0.748	0.760	0.779	0.728
	GPT-4o	0.748	0.787	0.766	0.748
	GPT-3.5	0.681	0.729	0.718	0.666
	Llama-2 13B	0.471	0.430	0.660	0.508
	Llama-3 8B	0.496	0.518	0.620	0.551
	Mistral 7B	0.543	0.592	0.652	0.526
5-shot	GPT-4	0.703	0.726	0.756	0.685
	GPT-4o	0.747	0.784	0.759	0.758
	GPT-3.5	0.666	0.715	0.719	0.638
	Llama-2 13B	0.504	0.457	0.644	0.513
	Llama-3 8B	0.481	0.498	0.623	0.545
	Mistral 7B	0.516	0.513	0.614	0.531
10-shot	GPT-4	0.734	0.730	0.779	0.702
	GPT-4o	0.737	0.769	0.744	0.764
	GPT-3.5	0.680	0.729	0.721	0.660
	Llama-3 8B	0.457	0.463	0.580	0.512
	Mistral 7B	0.521	0.556	0.599	0.523

Future work: We aim to extend our qualitative analyses to understand the reasons behind LLMs' underperformance for specific disaster types and information categories, with a focus on identifying actionable solutions to address these issues. Beyond text-based models, our future research will explore the potential of large vision-language models in processing multimodal social media data, such as combining textual and visual content, to provide a more holistic understanding of disaster events. This exploration is particularly relevant for enhancing emergency management systems in complex real-world scenarios.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv:2303.08774. Retrieved from https://arxiv.org/abs/2303.08774.

[2] Firoj Alam, Ferda Ofli, Muhammad Imran, and Michael Aupetit. 2018. A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. In *Proceedings of the International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. ISCRAM, Rochester, NY, USA, 553–572.

[3] Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. 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*. AAAI, Online, 933–942.

[4] Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021. CrisisBench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the International AAAI conference on web and social media*. AAAI, Online, 923–932.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Carlos Castillo. 2016. *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press, Cambridge, England.

[7] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, New York, NY, USA, 675–684.

[8] Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Ali, et al. 2023. LLMeBench: A Flexible Framework for Accelerating LLMs Benchmarking. arXiv:2308.04945. Retrieved from https://arxiv.org/abs/2308.04945.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805.

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv:2407.21783. Retrieved from https://arxiv.org/abs/2407.21783.

[11] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM computing surveys (CSUR)* 46, 4 (2014), 1–37.

[12] Jin Han, Zhe Zheng, Xin-Zheng Lu, Ke-Yin Chen, and Jia-Rui Lin. 2024. QuakeBERT: Accurate Classification of Social Media Texts for Rapid Earthquake Impact Assessment. arXiv:2405.06684. Retrieved from https://arxiv.org/abs/2405.06684.

[13] Xingsheng He, Di Lu, Drew Margolin, Mengdi Wang, Salma El Idrissi, and Yu-Ru Lin. 2017. The signals and noise: actionable information in improvised social media channels during a disaster. In *Proceedings of the 2017 ACM on web science conference*. ACM, New York, NY, USA, 33–42.

[14] Bohao Huang, Kyle Bradbury, Leslie M. Collins, and Jordan M. Malof. 2020. Do Deep Learning Models Generalize to Overhead Imagery from Novel Geographic Domains? The xGD Benchmark Problem. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, Waikoloa, HI, USA, 1476–1479. doi:10.1109/IGARSS39084.2020.9323080

[15] Oana Ignat, Gayathri Ganesh Lakshmy, and Rada Mihalcea. 2024. Cross-cultural Inspiration Detection and Analysis in Real and LLM-generated Social Media Data. arXiv:2404.12933. Retrieved from https://arxiv.org/abs/2404.12933.

[16] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)* 47, 4 (2015), 1–38.

[17] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (Portoroz, Slovenia, 23-28). European Language Resources Association (ELRA), Paris, France, 1–6.

[18] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825. Retrieved from https://arxiv.org/abs/2310.06825.

[19] Tarun Kalluri, Wangdong Xu, and Manmohan Chandraker. 2023. GeoNet: Benchmarking Unsupervised Adaptation across Geographies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ, USA, 15368–15379.

[20] Peter M Landwehr, Wei Wei, Michael Kowalchuck, and Kathleen M Carley. 2016. Using tweets to support disaster planning, warning and response. *Safety science* 90 (2016), 33–47.

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692. Retrieved from https://arxiv.org/abs/1907.11692.

[22] Yuchi Ma, Shuo Chen, Stefano Ermon, and David B. Lobell. 2024. Transfer learning in environmental remote sensing. *Remote Sensing of Environment* 301 (Feb. 2024), 113924. doi:10.1016/j.rse.2023.113924

[23] Volodymyr V Mihunov, Nina SN Lam, Lei Zou, Zheye Wang, and Kejin Wang. 2020. Use of Twitter in disaster rescue: lessons learned from Hurricane Harvey. *International Journal of Digital Earth* 13, 12 (2020), 1454–1466.

[24] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *Proceedings of the international AAAI conference on web and social media*. ACM, New York, NY, USA, 376–385.

[25] OpenAI. 2024. GPT-4o System Card. https://cdn.openai.com/gpt-4o-system-card.pdf. Accessed on 2024-11-24.

[26] Hakan T Otal and M Abdullah Canbaz. 2024. LLM-Assisted Crisis Management: Building Advanced LLM Platforms for Effective Emergency Response and Public Collaboration. arXiv:2402.10908. Retrieved from https://arxiv.org/abs/2402.10908.

[27] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108. Retrieved from https://arxiv.org/abs/1910.01108.

[28] Simone Splendiani and Antonella Capriello. 2022. Crisis communication, social media and natural disasters–the use of Twitter by local governments during the 2016 Italian earthquake. *Corporate Communications: An International Journal* 27, 3 (2022), 509–526.

[29] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288. Retrieved from https://arxiv.org/abs/2307.09288.

[30] Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, New York, NY, USA, 1079–1088.

[31] Rohan Singh Wilkho, Shi Chang, and Nasir G Gharaibeh. 2024. FF-BERT: A BERT-based ensemble for automated classification of web-based text on flash flood events. *Advanced Engineering Informatics* 59 (2024), 102293.

[32] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data* 18, 6 (2024), 1–32.

[33] Yizhou Zhang, Karishma Sharma, Lun Du, and Yan Liu. 2024. Toward Mitigating Misinformation and Social Media Manipulation in LLM Era. In *Companion Proceedings of the ACM on Web Conference 2024*. ACM, New York, NY, USA, 1302–1305.

[34] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. arXiv:2303.18223. Retrieved from https://arxiv.org/abs/2303.18223.

[35] Abdul Wahab Ziaullah, Ferda Ofli, and Muhammad Imran. 2024. Monitoring Critical Infrastructure Facilities During Disasters Using Large Language Models. Retrieved from https://arxiv.org/abs/2404.14432.

A Class-wise results

Tables 4 and 5 present the class-wise results across various shots for proprietary models (GPT-3.5, GPT-4, and GPT-4o) and open-source models (Llama-2 13B, Llama-3 8B, and Mistral 7B), respectively.

Table 4: F1 Scores for proprietary models (GPT-3.5, GPT-4, and GPT-4o) across classes and k-shots

Class	GPT-3.5					GPT-4					GPT-4o				
	ZS	1S	3S	5S	10S	ZS	1S	3S	5S	10S	ZS	1S	3S	5S	10S
Caution and advice	0.65	0.84	0.89	0.86	0.87	0.91	0.89	0.90	0.88	0.89	0.81	0.90	0.88	0.86	0.84
Rescue volunteering	0.89	0.95	0.94	0.94	0.95	0.92	0.92	0.90	0.88	0.88	0.91	0.93	0.93	0.91	0.89
Requests or urgent needs	0.68	0.58	0.58	0.50	0.55	0.71	0.66	0.74	0.68	0.73	0.70	0.75	0.75	0.81	0.85
Infrastructure damage	0.91	0.77	0.79	0.78	0.80	0.85	0.79	0.80	0.70	0.75	0.84	0.84	0.79	0.81	0.81
Sympathy and support	0.85	0.83	0.79	0.79	0.79	0.91	0.89	0.88	0.90	0.88	0.91	0.91	0.91	0.90	0.90
Injured or dead people	0.80	0.88	0.86	0.84	0.83	0.91	0.88	0.91	0.85	0.85	0.94	0.94	0.93	0.93	0.91
Displaced people	0.83	0.74	0.68	0.62	0.66	0.90	0.74	0.78	0.62	0.81	0.88	0.73	0.74	0.74	0.80
Missing or found people	0.87	0.81	0.86	0.82	0.85	0.85	0.88	0.84	0.83	0.81	0.82	0.88	0.89	0.90	0.93
Not humanitarian	0.46	0.78	0.72	0.77	0.77	0.70	0.88	0.80	0.91	0.81	0.91	0.86	0.88	0.88	0.87

Table 5: F1 Scores for open-source models (Llama-2 13B, Llama-3 8B, and Mistral 7B) across classes and k-shots

Class	Llama-2 13B					Llama-3 8B					Mistral 7B				
	ZS	1S	3S	5S	10S	ZS	1S	3S	5S	10S	ZS	1S	3S	5S	10S
Caution and advice	0.61	0.91	0.82	0.47	-	0.52	0.62	0.62	0.58	0.63	0.36	0.50	0.71	0.88	0.36
Rescue volunteering	0.75	0.53	0.35	0.42	-	0.50	0.47	0.50	0.44	0.35	0.95	0.96	0.87	0.65	0.82
Requests or urgent needs	0.73	0.78	0.84	0.83	-	0.34	0.28	0.27	0.25	0.21	0.56	0.55	0.66	0.70	0.63
Infrastructure damage	0.77	0.59	0.44	0.64	-	0.54	0.46	0.34	0.24	0.30	0.83	0.73	0.45	0.54	0.53
Sympathy and support	0.90	0.93	0.92	0.90	-	0.75	0.70	0.74	0.73	0.64	0.83	0.83	0.79	0.86	0.63
Injured or dead people	0.66	0.75	0.65	0.82	-	0.79	0.78	0.77	0.75	0.74	0.88	0.79	0.60	0.49	0.52
Displaced people	0.58	0.20	0.24	0.47	-	0.36	0.34	0.22	0.27	0.30	0.69	0.36	0.46	0.66	0.68
Missing or found people	0.70	0.78	0.80	0.79	-	0.47	0.34	0.48	0.51	0.44	0.64	0.69	0.65	0.75	0.77
Not humanitarian	0.34	0.73	0.69	0.56	-	0.53	0.57	0.53	0.57	0.50	0.69	0.86	0.87	0.60	0.89