Monitoring Critical Infrastructure Facilities During Disasters Using Large Language Models

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

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

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

ABSTRACT

Critical Infrastructure Facilities (CIFs), such as healthcare and transportation facilities, are vital for the functioning of a community, especially during large-scale emergencies. In this paper, we explore a potential application of Large Language Models (LLMs) to monitor the status of CIFs affected by natural disasters through information disseminated in social media networks. To this end, we analyze social media data from two disaster events in two different countries to identify reported impacts to CIFs as well as their impact severity and operational status. We employ state-of-the-art open-source LLMs to perform computational tasks including retrieval, classification, and inference, all in a zero-shot setting. Through extensive experimentation, we report the results of these tasks using standard evaluation metrics and reveal insights into the strengths and weaknesses of LLMs. We note that although LLMs perform well in classification tasks, they encounter challenges with inference tasks, especially when the context/prompt is complex and lengthy. Additionally, we outline various potential directions for future exploration that can be beneficial during the initial adoption phase of LLMs for disaster response tasks.

Keywords

Large language models, disaster management, social media, information classification, information retrieval

INTRODUCTION

Critical infrastructures, including essential facilities and services vital for societal functioning, span sectors like energy, transport, healthcare, and telecommunications (Labaka et al., 2016; Pescaroli & Kelman, 2017). Ensuring the continuous functioning and accessibility of critical infrastructures is essential in assisting vulnerable populations during disasters and disruptions to these critical infrastructures can result in increased human and economic losses (Auerswald et al., 2005). However, during major disasters, authorities are overwhelmed and thus face challenges in maintaining an updated status of these facilities, and as a consequence, the general public is kept uninformed about the latest status of these facilities.

This paper bridges the gap by exploring an unconventional data source—social media—to monitor Critical Infrastructure Facilities (CIFs) in a given area of interest (AOI). Social media platforms, especially microblogging sites such as X (formerly Twitter), are important data sources for real-time disaster updates. People post information about early warnings, cautions, damages, and their needs almost instantaneously (Olteanu et al., 2015; Qu et al., 2011). Past studies show the use of social media information for several disaster response tasks including damage and needs assessment, among others (Castillo, 2016; Imran et al., 2015).

While social media holds valuable information, it frequently becomes overloaded with noise—comprising casual chatter filled with slangs, abbreviations, and grammatically unconventional sentences. Past studies suggest employing supervised classification techniques to discern social media messages relevant to disaster response (Imran et al., 2015). However, these techniques require human-labeled data for each unique classification task or even when adapting a model trained on past disaster data to a new emergency. For example, Alam et al. (2021) manually labeled around 77K tweets to train deep neural network models such as BERT, RoBERTa, for detecting situational awareness messages. Similarly, Zahra et al. (2020) shows an extensive feature engineering process to train classifiers for eyewitness message detection. These methods are tailored for specific classification tasks, and should a new task/class be needed, retraining of the models would be required. To tackle this challenge, our work explores the use of Large Language Models (LLMs) to perform traditional computational tasks, an effort toward eliminating the requirement for explicit model training.

To monitor CIFs in a specified AOI, our proposed approach starts by acquiring all critical facilities in the area from Open Street Maps (OSM) through automatic APIs. Social media data collected about an ongoing disaster is processed through an LLM to generate embeddings and stored in a vector database.[1] Embedding of a message encapsulates its semantics and thus yields better search results. Therefore, for each CIF obtained from OSM, we query the vector database, where raw embeddings are stored, to retrieve messages pertinent to the specified CIF. The retrieved messages are then analyzed by LLMs to identify *(i)* the impacts reported in each message, *(ii)* the severity of the reported impacts, and *(iii)* the operational status of the CIF. We perform all the tasks in a zero-shot setting and evaluate each step of the proposed methodology using standard evaluation metrics.

Through extensive experimentation, we show both the strengths and limitations of LLMs in handling various computational tasks associated with the processing of social media messages. Notably, we showcase diverse configurations of our CIF retrieval query and the corresponding advantages they offer. For instance, our CIF query consisting of a fixed general term (e.g., "disaster impacts") yields better outcomes compared with a detailed listing of impacts. Furthermore, we note better classification performance of LLMs for individual messages when compared to the classification and inference from a set of messages. Overall, we demonstrate that LLMs, even in the zero-shot setting, hold great potential to replace traditional supervised models. However, there are certain weaknesses as well. Notably, LLMs may struggle with context understanding and misinterpret nuanced or ambiguous language commonly found in social media. Additionally, their performance can be influenced by factors like data biases and the nuances of prompt engineering. Acknowledging these limitations is crucial to refining the application of LLMs in processing social media messages effectively.

The rest of the paper is structured as follows. The following section reviews Related Work. In the Methodology section, we provide details of our proposed approach. Results are presented in the Results section and further deliberated upon in the Discussion section. The paper concludes in the final section.

RELATED WORK

Critical infrastructures are at the heart of economic and social welfare in all countries. Therefore, understanding their risk and resilience to various hazards and disasters have been crucially studied in literature (Yusta et al., 2011). For example, Fekete et al. (2015) explored the use of Geographic Information Systems (GIS) and Remote Sensing (RS) whereas Jovanović et al. (2018) presented an indicator-based approach to assess resilience of critical infrastructures to disasters. To this end, Baloye and Palamuleni (2016) proposed a critical infrastructure-driven spatial database for proactive disaster management. Taking a step further, Qiang (2019) provided a comprehensive assessment of flood exposure of all major critical infrastructures in the United States.

In the last decade, with the advancement of social networks, researchers started to explore the use of real-time social media data during disasters to assess risk and resilience of critical infrastructures. For instance, Mittelstädt et al. (2015) integrated social media data with mobile in-situ data to monitor the disaster impact on critical infrastructure. Later Fan and Mostafavi (2019) developed a graph-based method for detecting credible situation information related to infrastructure disruptions during natural disasters. In another study, Roy et al. (2020) presented a multilabel classification approach to identify hurricane-induced infrastructure disruptions through a sentiment analysis of social media data. They utilized several supervised machine learning models to classify disruptions in social media messages. Alternatively, Heglund et al. (2021) demonstrated the utility of social media data for critical infrastructure resilience based on statistical analysis and forecasting methods. While plethora of these methods utilize conventional pipelines of message filtration, trend detection, and other case-specific algorithms, our methodology leverages more generalized inferential capabilities of the large language models (LLMs).

[1]The list of CIFs, data, and embeddings are available upon request.

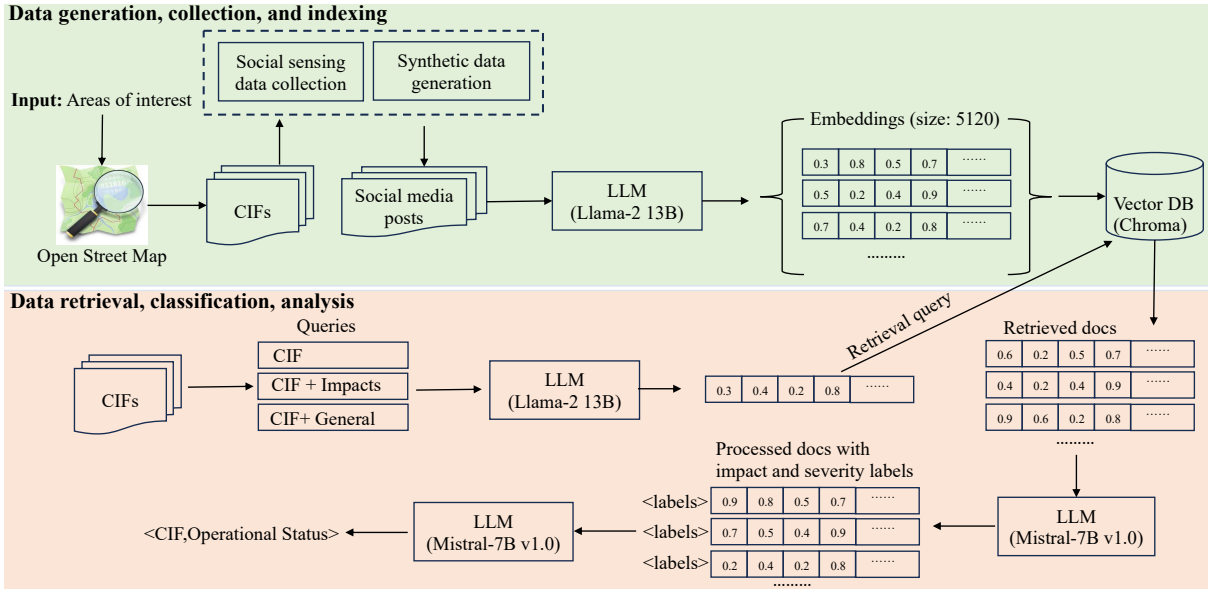


Figure 1. High-level methodology detailing two pipelines: *(i)* data generation, collection, and indexing, *(ii)* data retrieval, classification, and analysis

With the emergence of pre-trained LLMs such as GPT3.5 (Brown et al., 2020), GPT4 (OpenAI et al., 2023), Llama2 (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), more versatile capabilities have become available for research. For instance, Colverd et al. (2023) has recently studied the use of LLMs for generating flood disaster impact reports via a system initiated by a search query phrase, which is then utilized to gather textual data about the disaster from the web. Another recent use of LLM is geotagging of tweets where few-shot learning is utilized for identification of various location descriptors in tweets such as door number addresses, road segments, and road intersections (Hu et al., 2023).

Inspired by this trend, we capitalize on an LLM for real-time monitoring of critical infrastructure facilities (CIFs) during disasters by classifying their situations in three crucial aspects including the type of impact, its severity, and the operational status. Additionally, to compensate for lack of a publicly available dataset with situational information about CIFs, we again revert to an LLM for its generative capabilities to create synthetic data to test our CIF monitoring pipeline.

METHODOLOGY

Overview

The primary objective of this work is to perform real-time monitoring of various Critical Infrastructure Facilities (CIFs) amidst large-scale disaster scenarios. Specifically, we aim to obtain three types of updates for each CIF in the given area of interest, including *(i)* identifying the disaster impact (e.g., damaged, burned) to the CIF, *(ii)* assessing the severity of the impact (e.g., moderate, severe) and *(iii)* determining the operational status of the facility (e.g., closed, partially closed, or open). We employ social sensing data from social media platforms, harnessing its capability to potentially deliver immediate updates, potentially from eyewitnesses or people possessing any relevant information.

Figure 1 depicts the high-level methodology comprising two pipelines: *(i)* data collection and indexing, and *(ii)* data retrieval and processing. The data collection and indexing pipeline performs tasks related to data acquisition, synthetic data generation, embedding generation, and indexing. The pipeline for data retrieval and processing focuses on retrieving pertinent data from the embedding database. Subsequently, the retrieved data are analyzed through LLMs to identify the impact, severity, and operational status of CIFs. Next, we provide detailed descriptions of the important processes of the methodology.

Obtaining CIFs

We utilize Nominatim APIs[2] from Open Street Map (OSM) (OpenStreetMap contributors, 2017) to retrieve information about CIFs within a specified geographical area of interest (AOI). In this work, we focus on two regions

[2]https://nominatim.openstreetmap.org/

prone to natural disasters: Christchurch, New Zealand, susceptible to earthquakes, and Broward County, Florida, USA, susceptible to hurricanes. We note that for both AOIs we also have Twitter data previously collected from these locations during past disasters. To extract CIF data for these AOIs, we construct queries that include the AOI name and the type of infrastructure, including hospitals, fire stations, rail stations, educational facilities, bridges, roads, and tunnels. Example queries include "Christchurch hospitals," "Christchurch fire stations," etc. Employing this approach, we gathered 58 CIFs from Christchurch and 82 CIFs from Broward County, along with associated metadata such as names, addresses, and geographic coordinates.

Data Curation: Real and Synthetic

Social media data about CIFs is central to our study. However, to the best of our knowledge, there is no publicly available social media data containing tagged CIFs and their associated impacts, severity, and operational statuses. To address this gap, we leverage generative AI to create synthetic data (i.e., tweets) that include real-world CIFs from our selected AOIs. In particular, we utilize the open-source Llama-2 13B model (Touvron et al., 2023) to generate tweets reporting diverse impact scenarios on CIFs retrieved from OSM. Moreover, for each generated tweet, we ask the model to provide the reported impact along with its corresponding severity labels. The specific prompt employed for tweet generation for Broward County is as follows:

> *Generate 15 diverse tweets describing the impact of a Category-5 hurricane on [CIF NAME & ADDRESS]. The disaster triggered sub-events such as tornado, storm surge, burst of rain, strong winds, resulting in varied impacts like flooded, collapsed, submerged, damaged, destroyed, cracked, etc.*
>
> *Ensure linguistic diversity in each tweet, providing unique insights into the impact and its severity. In some tweets, include the infrastructure's address. Aim for tweet lengths between 100 to 250 characters, and avoid using emojis. Tag each tweet with the type of impact (e.g., damaged, destroyed) and its severity (e.g., low, mild, severe). These tags will be used for training classifiers.*
>
> *Always include 2 tags at the 'end' of the generated tweet with the following template: (Tags: ***** , *****)*

For instance, the response returned by the LLM for CIF *"Aventura Hospital and Medical Center & 20900, Biscayne Boulevard, Aventura, Miami-Dade County, Florida, 33180, United States"* looks as follows (showing first three tweets for brevity):

> *Aventura Hospital's emergency room is flooded after a burst of rain caused by the Category 5 hurricane. Patients and staff are being evacuated to safer areas. (Tags: Flooded, Mild)*
>
> *Strong winds from the hurricane have caused significant damage to Aventura Hospital's roof and windows, leaving many areas closed off to patients and staff. (Tags: Damaged, Severe)*
>
> *The storm surge from the hurricane has submerged the lower levels of Aventura Hospital, forcing patients and staff to be evacuated to higher floors. (Tags: Submerged, Severe)*

A similar prompt is used for Christchurch where the first paragraph is adapted for an earthquake scenario as follows:

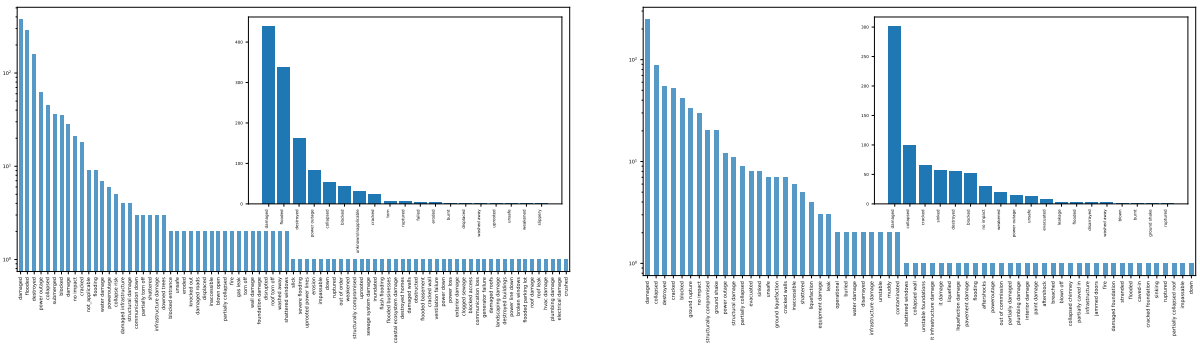
> *Generate 15 diverse tweets describing the impact of a severe earthquake on Christchruch's [CIF NAME & ADDRESS]. The disaster triggered sub-events such as ground shake, landslide, liquefaction, ground rupture, aftershock, resulting in varied impacts like collapsed, cracked, damaged, destroyed, etc.*

Subsequently, we utilize the same LLM employed for tweet generation to assign an operational status label to each generated tweet. The model is restricted to select one of four operational statuses: "open," "closed," "partially open," "partially closed," and "unknown" if no operational status is explicitly mentioned in the tweet. The prompt to obtain the operational status is as follows:

> *Your task is to analyze the provided tweet and determine the operational status of the mentioned infrastructure. The operational status could include descriptors such as open, closed, partially open, partially closed, or unknown.*
>
> *Tweet: [TWEET]*
>
> *Operational status:*

A sample response from the above prompt for the tweet *"The computer lab at Christchurch Girls' High School has been closed due to a collapsed ceiling. Students are using temporary facilities until further notice. #ChristchurchEarthquake"* is as follows:

> *Operational status: closed*



(a) Broward County **(b) Christchurch**

Figure 2. Distribution of impact labels in the synthetic data for (a) Broward County and (b) Christchurch. The outer charts with light blue bars correspond to the LLM-generated raw tags whereas the overlaid smaller charts with dark blue bars show the manually pruned ground-truth tags.

In total, we generated 728 tweets for Christchurch for 58 CIFs and 1,205 tweets for Broward County against 82 CIFs. Although instructed, the model does not always generate 15 tweets for each CIF. Tweets with incorrect tags/labels were manually corrected. As we did not control the prompt to enforce a closed-set of impact types, the resulting list of impact tags is open-ended (N=83), which also allowed for more diversity in the generated tweets. Figure 2 shows the distribution of the resulting tags for each AOI. We noticed multiple similar impact labels (e.g., flooded, flooding, inundated, etc.) in the model responses, which we merged into a unified list of tags, as illustrated in Table 1. The sub-figures in Figure 2 depict the distribution of the refined tag list. Note that, unlike impact tags, we utilized a closed-set of tags for classification of severity (i.e., severe, moderate, mild, and unknown) and operational status (i.e., open, closed, partially open, partially closed, and unknown).

Later, for each CIF, we dispersed its generated tweets across four six-hour time intervals (spanning a 24-hour time frame) by considering their impact order (e.g., "destroyed" followed by "damaged" or "cracked" and so on). The impact ordering ensures a coherent progression of the impact to a CIF. This step is crucial for simulating a real-world setting in our subsequent experiments. Moreover, given that several past studies highlight the prevalence of noise and the limited relevant information in real-world tweets (Grace, 2021; Kumar et al., 2017), we aim to ensure the realism of our dataset. To achieve this, we introduce noise into our generated data with a signal-to-noise ratio of 2% (signal) to 98% (noise). For this purpose, we incorporated tweets collected during past disasters for both AOIs. Specifically, for Christchurch, we used tweets collected after a magnitude-7.1 earthquake struck New Zealand on September 2nd, 2016. For Broward County, we utilized tweets collected after Hurricane Ian hit Florida in September 2022. In total, we prepared 36,286 and 60,062 tweets (including both synthetic and real) for Christchurch and Broward County, respectively. To ensure that CIFs also appear in non-impact tweets, we injected CIF names into 8% of the noisy tweets in each time interval.

Finally, to assign an overall status for a CIF within each time interval, we deduce its operational status. To achieve this, we chronologically order tweets for a CIF and use the operational status of the last tweet as an overall status. If the operational status of the last tweet is "unknown", we use the last valid status.

Generating and Indexing Tweet Embeddings

Next, we employ Llama-2 13B (Touvron et al., 2023) to compute embeddings for all the tweets, including synthetic and real. Each tweet is converted into a 5120-dimensional embedding vector which is the default embedding size of Llama-2 13B. This embedding is then indexed into ChromaDB[3], a well-known open-source vector database. These embeddings serve as the basis for semantic search, enabling the retrieval of tweets specifically mentioning certain CIFs.

Retrieval of CIF Tweets

The first step in determining the impact and operational status of a CIF involves retrieving tweets referencing a specific CIF. To achieve this, we explore three types of retrieval queries: *(i)* a query comprising only the CIF name, *(ii)* a query involving both the CIF name and a set of impact terms such as damaged, cracked, burned, destroyed, submerged, and *(iii)* a query combining the CIF name with the term "disaster impacts." For each query, we first

[3]https://www.trychroma.com/

Table 1. Raw impact labels and their mapping to final impact taxonomy for LLM-based classification

Consolidated labels	Raw model labels
damaged	aftershock damage, damage, damaged, broken windows, damaged infrastructure, damaged roads, damaged roofs, damaged walls, electrical damage, coastal ecosystem damage, foundation damage, landscaping damage, exterior damage, hvac damage, infrastructure damage, roof damage, plumbing damage, sewage system damage, structural damage, wall damage, water damage, equipment damage, damaged foundation, interior damage, IT damage, liquefaction damage, partially damaged, pavement damage, plumbing damage, IT infrastructure damage, equipment damage, paint damage
flooded	flooded basement, flooded businesses, flooded parking lot, flooding, severe flooding, flash flooding, submerged, inundated
destroyed	destroyed, destroyed buildings, destroyed homes, rubble
weakened	weakened, structurally compromised
cracked	cracked, shattered, shattered windows, debris, cracked wall, cracked walls
blocked	blocked access, blocked entrance, impassable, obstructed, clogged sewage, inaccessible, partially blocked, landslide, jammed doors
torn	partially torn off, roof torn off
power outage	power line down, power loss, power outage, power down, communication down, communication loss, down, knocked out, disrupted, downed trees, offline
ruptured	blown open, gas leak, roof leak
collapsed	partially collapsed, collapsed, crushed, collapse risk, collapsed wall, collapsed chimney, partially collapsed roof
failed	generator failure, ventilation failure, out of order
uprooted	uprooted, uprooted power lines
eroded	eroded, erosion
washed away	washed away, muddy
slippery	slippery, slick
displaced	displaced
blown	blown, blown off
burnt	burnt, burning, fire
unsafe	unsafe, unstable, structurally compromised, uninhabitable, contaminated
leakage	leak, gas leak, gasleak
sinked	ground liquefaction, liquefaction, buried, sinked, sinking, caved-in, ground rupture, liquefied
unknown/inapplicable	not applicable, not humanitarian, no impact, unknown

obtain its embeddings from the Llama-2 13B model and subsequently retrieve the top-K (ranging from 5 to 50 with a step size of 5) tweets by querying ChromaDB using cosine similarity.

Top-K results for each query are then used to compute Average Precision (AP) using Equation (1).

$$\text{AP@}K = \frac{1}{\text{GTP@}K} \sum_{\substack{k=5n \\ 1 \le n \le N}} \text{P@}k \times \text{rel@}k, \text{ for } K = 5N \text{ and } N = \{1, \ldots, 10\} \tag{1}$$

where

$$\text{P@}k = \frac{\text{Number of relevant documents retrieved @}k}{k}$$

$$\text{rel@}k = \begin{cases} 1, & \text{if } k^{th} \text{ document is relevant} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{GTP@}K = \min(K, \text{Total number of relevant documents})$$

In Equation (1), we normalize the total sum by the number of 'retrievable' ground-truth documents or *signals* (i.e., Ground Truth Positive or GTP). Finally, to determine the overall retrieval performance of a given query, we

compute mean Average Precision (i.e., mAP) in a time interval $t \in \{0h\text{-}6h, 6h\text{-}12h, 12h\text{-}18h, 18h\text{-}24h, 0\text{-}24h\}$ using Equation (2).

$$\text{mAP@}K^t = \frac{1}{M}\sum_{i=1}^{M} \text{AP@}K_i^t, \text{ for } K = 5N \text{ and } N = \{1, \ldots, 10\} \tag{2}$$

Tweet Classification for Impact, Severity, and Operational Status

We employ the Mistral 7B v1.0 model (Jiang et al., 2023) for classifying the retrieved tweets that are identified as relevant by the CIF queries. These tweets are likely to encompass information about impacts experienced by critical facilities. To perform the impact and severity classification of each tweet, we use the following prompt template:

> *Your task is to analyze the provided tweet and determine the impacts of the disaster on the mentioned infrastructure. Please include only impact descriptors from the list such as blocked, blown, buried, burnt, collapsed, cracked, damaged, destroyed, displaced, disrupted, eroded, failed, flooded, ground liquefaction, ground shake, leakage, muddy, power outage, ruptured, slippery, torn, unsafe, uprooted, washed away, weakened or not_applicable, and severity such as severe, mild, moderate, unknown.*
>
> *Tweet: [TWEET]*
>
> *Infrastructure impact:*
> *Infrastructure severity:*

The response returned by the model to an example tweet such as *"The Christchurch Public Hospital's radiology equipment is malfunctioning due to the earthquake, making it difficult to diagnose patients. #ChristchurchEarthquake"* looks as follows:

> *Infrastructure impact: damaged*
> *Infrastructure severity: moderate*

Finally, tweets that are tagged with an impact type are further analyzed to infer the operational status of CIFs. For this purpose, we use the Mistral 7B v1.0 model with the following prompt:

> *Your task is to analyze the provided tweet and determine the operational status of the mentioned infrastructure. The operational status could include descriptors such as open, closed, partially open, partially closed, or unknown.*
>
> *Tweet: [TWEET]*
> *Operational status:*

The response returned by Mistral 7B v1.0 model to an example tweet such as *"Ground shake from Christchurch earthquake caused significant damage to Wigram Fire Station's foundation, rendering it unstable. #ChristchurchEarthquake"* looks as follows:

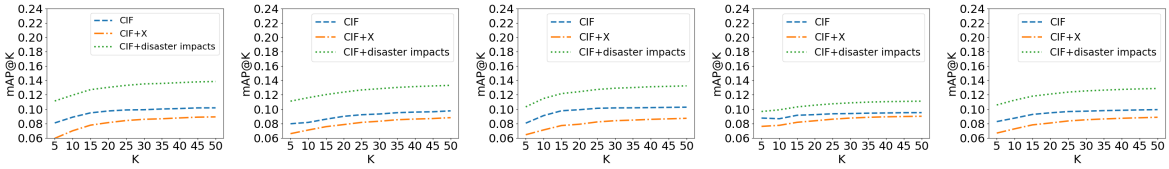
> *Operational status: closed*

To evaluate the performance of impact, severity, and operational status classification, we compare the labels predicted by the Mistral 7B model with the ground-truth labels obtained from the Llama-2 model and report results using standard evaluation metrics such as precision, recall, and F1-score.

Obtaining Overall Operational Status of CIFs

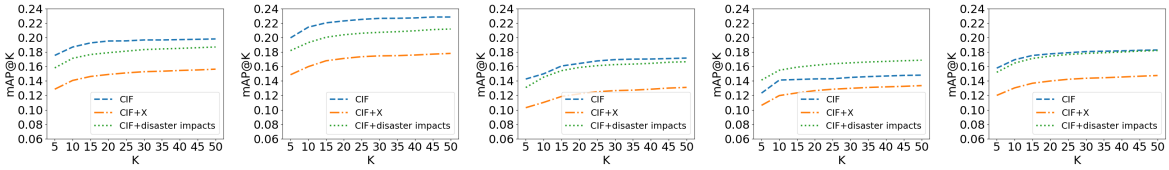
The last step in the tweet classification and analysis pipeline involves extracting the overall operational status of tweets retrieved through a CIF query. The set of tweets resulting from a CIF query may encompass information about the queried CIF, but these tweets may vary in terms of containing impact-related details. Therefore, to determine the overall operational status of a CIF at a specific time, we only consider tweets containing impacts (relying on the impact labels obtained in the previous step). To achieve this, we employ the Mistral 7B v1.0 model with the following prompt to retrieve the overall status.

> *Your task is to analyze the tweets given below and deduce the operational status of a facility, named [CIF]. Since these tweets are retrieved based on the facility name, it's possible that some tweets may not pertain to the given facility. Focus solely on the tweets pertinent to [CIF] and derive the most recent operational status for the facility. Your operational status label must be one of these: open, closed, partially open, partially closed, or unknown.*
>
> *Tweet: [TWEET 1]*
> *Tweet: [TWEET 2]*
> *Tweet: [TWEET 3]*
> *...*
>
> *operational_status:*



(a) 0h-6h **(b) 6h-12h** **(c) 12h-18h** **(d) 18h-24h** **(e) 0h-24h**

Figure 3. Performance comparison of retrieval queries for Broward County



(a) 0h-6h **(b) 6h-12h** **(c) 12h-18h** **(d) 18h-24h** **(e) 0h-24h**

Figure 4. Performance comparison of retrieval queries for Christchurch

RESULTS

In this section we present our results for the retrieval and classification tasks described in the Methodology section.

Retrieval of CIF Tweets

For each AOI, we considered three types of queries to fetch relevant tweets pertaining to each CIF:

(i) *CIF* — a query comprising only the CIF name

(ii) *CIF + X* — a query involving both the CIF name and a set of impact terms such as damaged, cracked, burned, destroyed, submerged, etc. (denoted as X for brevity)

(iii) *CIF + disaster impacts* — a query combining the CIF name with the term "disaster impacts"

Specifically, for Broward County *X* comprises *flooded, submerged, damaged, destroyed, weakened, cracked, blocked, torn, power outage, ruptured, collapsed, failed, uprooted, eroded, burnt, washed away, slippery, displaced, disrupted* whereas for Christchurch it consists of *flooded, destroyed, leak, blocked, cracked, ground liquefaction, power outage, ruptured, buried, collapsed, ground shake, unsafe, muddy*.

We also experimented with K to assess its effect on the retrieval performance. That is, we ran queries for top-K where we changed K from 5 to 50 with a step size of 5. We then computed the overall mAP@K in each time interval using Equation (2) to evaluate the overall performance of each query type. Figures 3 and 4 show the performance plots for all query types across different K values and time intervals for Broward County and Christchurch, respectively. The plots initially show an increasing trend in performance with increasing K but the performance improvement generally slows down after $K = 30$. For Broward County, *"CIF + disaster impacts"* queries outperform the other queries by a big margin whereas for Christchurch *"CIF"* queries hold a slight edge over the other queries in the six-hour time intervals but perform only on par with *"CIF + disaster impacts"* in the 0h-24h time interval.

Tables 2 and 3 summarize the mAP@K scores for $K = 50$ for Broward County and Christchurch, respectively. Again we observe that the scores for the *"CIF + disaster impacts"* queries are highest in all time intervals for Broward County whereas *"CIF"* and *"CIF + disaster impacts"* queries perform on par for Christchurch. However, it is important to note that, for both AOIs, *"CIF + disaster impacts"* queries retrieve the highest number of relevant tweets among all query types in 0h-24h time interval, i.e., 534 out of 1205 relevant tweets in Broward County and 402 out of 728 relevant tweets in Christchurch. Therefore, we conclude that *"CIF + disaster impacts"* query type yields the best overall retrieval performance, and hence, we use it as the default query type in the proposed pipeline.

We note that the resulting mAP scores are generally low and the effective hit rate (i.e., ratio of relevant retrieved tweets to total number of retrieved tweets) is only slightly above 3% for both AOIs. To investigate this further, we color coded and plotted the distribution of relevant and irrelevant tweets in Figure 5. We additionally sub-divided irrelevant tweets belonging to either noise or other CIF that we generated. From Figure 5, we observe that although a large portion of the distribution for each CIF contains irrelevant tweets, a significant portion of irrelevant tweets is

Table 2. Retrieval performance (mAP@50) of the Broward County queries

	Time Intervals					
Queries	0h–6h	6h–12h	12h–18h	18h–24h	0h–24h	Y
CIF	0.102	0.098	0.103	0.095	0.099	303
CIF + X	0.089	0.088	0.087	0.090	0.089	484
CIF + "disaster impacts"	**0.139**	**0.133**	**0.133**	**0.111**	**0.129**	**534**

X = flooded, submerged, damaged, destroyed, weakened, cracked, blocked, torn, power outage, ruptured, collapsed, failed, uprooted, eroded, burnt, washed away, slippery, displaced, disrupted
Y = Number of relevant tweets retrieved out of 1205 in 0-24h time interval

Table 3. Retrieval performance (mAP@50) of the Christchurch queries

	Time Intervals					
Queries	0h–6h	6h–12h	12h–18h	18h–24h	0h–24h	Y
CIF	**0.198**	**0.229**	**0.172**	**0.148**	**0.183**	288
CIF + X	0.156	0.178	0.131	0.134	0.148	326
CIF + "disaster impacts"	0.187	0.212	0.167	0.169	0.182	**402**

X = flooded, destroyed, leak, blocked, cracked, ground liquefaction, power outage, ruptured, buried, collapsed, ground shake, unsafe, muddy
Y = Number of relevant tweets retrieved out of 728 in 0-24h time interval

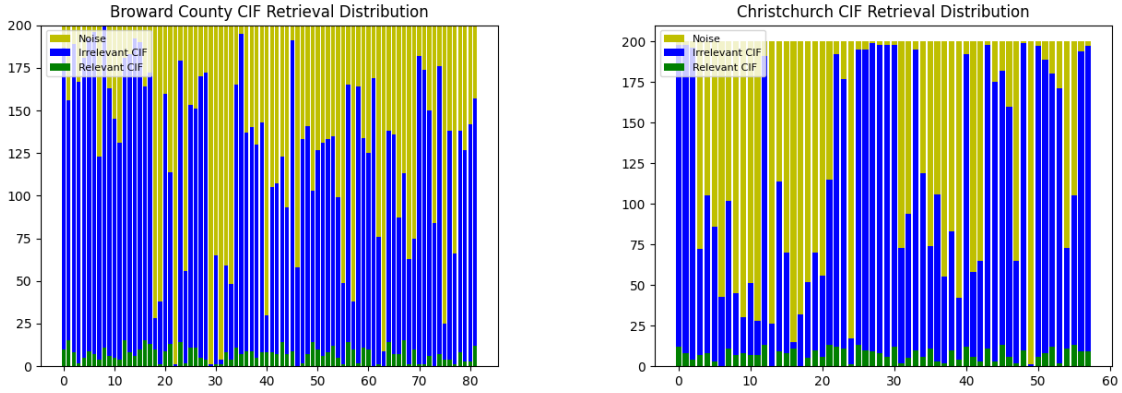
occupied by tweets describing impacts to other CIFs and share some semantic similarity with the relevant tweet. For each query on average roughly 60% of the retrieved results belong to a certain CIF (not necessarily the correct one) and the remaining contains pure noise. Although achieving a higher hit rate is always desirable, we defer further improvements for the retrieval process to our future work.

Tweet Classification for Impact, Severity, and Operational Status

Here we compare the ground-truth impact, severity, and operational status labels obtained from the Llama-2 13B model in the dataset generation stage with the labels inferred by the Mistral 7B v1.0 model in the data analysis and classification stage. Because we do not expect the classification models to depend on a specific CIF or time interval, we evaluate the classification performance over all tweets across all CIFs and time intervals using standard evaluation metrics such as Precision, Recall, and F1-score. Table 4 summarizes the results of tweet-level impact, severity, and operational status classification for both AOIs. The results indicate that the proposed LLM-based classification models achieve reasonable performance for the task at hand. In general, the models achieve better precision than recall in all three tasks. However, the differences between precision and recall are bigger in the operational status classification task, which leads to the lowest F1-scores among all the tasks.

To take a closer look at the impact classification performance on the synthetic data, we provided confusion matrices in Figure 6. For Broward County, the most prominent impact classes such as *damaged* and *flooded* are correctly classified most of the time, but it is notable that around one third of the *flooded* tweets are classified as *damaged* (N=112). This is understandable because *damaged* class has a broader meaning and coverage that overlaps with many other impact types including *flooded*. Similarly, *blocked* (N=20), *destroyed* (N=56), and *power outage* (N=58) tweets are usually classified as *damaged*. However, the opposite is not correct, and in line with this expectation, we see much fewer *damaged* tweets getting labeled as *flooded* (N=2), *blocked* (N=13), *destroyed* (N=2), and *power outage* (N=1). We also observe that two thirds of the *unknown/inapplicable* tweets (N=20) are confused as *damaged*, which is not very desirable and needs closer investigation. For Christchurch, we observe similar trends where, as the most prominent impact class, *damaged* tweets are correctly classified most of the time (N=302), but many other impact tweets such as *blocked* (N=17), *collapsed* (N=51), *cracked* (N=55), *destroyed* (N=15), and *weakened* (N=15) are also tagged as *damaged*. Moreover, majority of the *unknown/inapplicable* tweets (N=23) are confused as *damaged* tweets. Altogether these observations suggest that it is challenging to get the Mistral 7B v1.0 model (or any other LLM for that matter) distinguish between fine-grained impact types.

We then also analyze the impact classification performance on the output of the retrieval system which contains noisy data due to imperfect nature of the retrieval process. Here we assume that ground-truth labels for the noise tweets belong to *unknown/inapplicable* class for all classification tasks. Table 5 summarizes the scores for impact,



(a) Broward County **(b) Christchurch**

Figure 5. Signal-to-noise distribution of retrieved tweets for each CIF in (a) Broward County and (b) Christchurch

Table 4. Performance scores for impact, severity, and operational status classification on the synthetic data

	Impact			Severity			Operational Status		
AOI	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Broward County	0.660	0.563	0.565	0.668	0.512	0.566	0.805	0.349	0.445
Christchurch	0.603	0.570	0.569	0.550	0.458	0.490	0.655	0.280	0.342

severity, and operational status classification on the retrieved tweets. These scores are higher thanks to the model's ability to correctly identify the (incorrectly retrieved) noise tweets as "unknown/inapplicable."

Figure 7 visualizes confusion matrices for further analysis of the classification results of the retrieved tweets. For both AOIs, the confusion matrices are dominated by the correctly identified noise tweets as *unknown/inapplicable* (i.e., N=4987 for Broward County and N=1178 for Christchurch). Besides, for Broward County, majority of the tweets are correctly classified as *damaged* (N=1536), *destroyed* (N=1336), *flooded* (N=1781), *blocked* (N=381), and *power outage* (N=365). However, many ground-truth *damaged* tweets were also labeled incorrectly as *torn* (N=467) or *power outage* (N=273) and many *flooded* tweets as *damaged* (N=308) or *power outage* (N=285). Whereas for Christchurch, while *damaged* (N=252), *destroyed* (N=90), *collapsed* (N=108), and *blocked* (N=47) tweets are classified correctly in general, we see that most *cracked* tweets were confused as *damaged* (N=84) and *damaged* tweets as *collapsed* (N=34). We also see that impacts such as *burnt, ruptured, uprooted, sink,* and *weakened* are oftentimes misclassified. Another interesting observation is that so many tweets are classified (incorrectly more than half of the time) as *power outage* in Broward County whereas almost no tweet gets classified as *power outage* in Christchurch.

Overall Operational Status Classification of CIFs

Tweets retrieved with the impact labels were processed using Mistral 7B v1.0 to infer the overall operational status of a target CIF. The results are provided in Table 6. While exhibiting a reasonable precision, the model struggled with the recall. This could be due to the retrieval process, which incorporates messages from other CIFs and noise. Even though we mitigated the noise by focusing solely on impact-related messages, messages about other pertinent CIFs still pose a challenge for the model.

DISCUSSION

Besides illustrating the opportunities LLMs have to offer for real-time monitoring of CIFs by analyzing social media data during disasters, this study also reveals various challenges to be tackled along the way. For instance, an important challenge is how to instruct LLMs to generate a realistic timeline of short messages (i.e., tweets) about CIFs during disasters that are coherent and consistent both spatially and temporally. Another important challenge is how to guide LLMs to eliminate irrelevant messages successfully in the retrieval process. Yet another important challenge is how to prompt LLMs in a zero-shot setting to discern fine-grained details about disaster impact types and severity affecting CIFs as well as to infer the operational status of CIFs from a set of impact and severity reports.

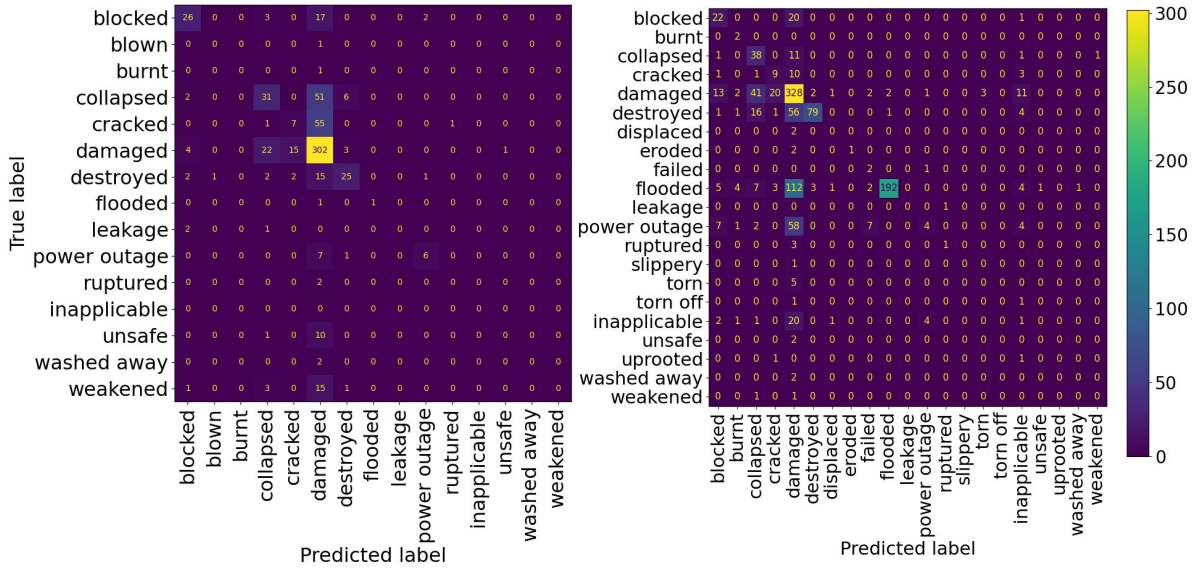


Figure 6. Confusion matrices for all signal impact classification (left) Broward County and (right) Christchurch

Table 5. Performance scores for impact, severity, and operational status classification on the retrieved tweets

	Impact			Severity			Operational Status		
AOI	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Broward County	0.821	0.747	0.769	0.687	0.6445	0.6442	0.517	0.352	0.399
Christchurch	0.841	0.805	0.777	0.817	0.794	0.797	0.919	0.779	0.809

On the data side, we opted for open-ended generation of impact types. This allowed us to generate a quite diverse set of fine-grained impact types (N=83), a property sought by stakeholders to make key decisions. However, this strategy also resulted in a long-tail distribution of impact types where the generic *damaged* became the most prominent impact type by a large margin while more than two thirds of the impact types appeared only once or twice in the entire dataset (see Figure 2). An alternative strategy could be to impose more control over the distribution and ask LLM to generate tweets for a specific impact type at a time (potentially at the expense of reduced diversity).

On the retrieval side, our current approach experienced difficulty in eliminating the irrelevant content for some CIFs more than others (see Figure 5). This could be attributed partly confusing nature of similar CIF or partial/shortened CIF names and partly to weaknesses of the retrieval method employed. In future studies, we plan to explore more sophisticated retrieval techniques.

On the classification side, we observed weaknesses of LLMs in distinguishing granular details about disaster impact types as analyzed through confusion matrices in Figures 6 and 7. There is also the fact that there are multiple correct answers many times and formulating the problem as a single-label classification becomes problematic as illustrated by the examples below:

> *Case 1: The Category 5 hurricane has caused significant damage to the South Area Alternative School's electrical system, leaving the building without power. The school's future is uncertain.*
>
> *Llama-2 13B (ground truth): Power outage*
> *Mistral 7B v1.0 (classified): Damaged*
>
> *Case 2: The hurricane has caused a tree to fall on top of a building near Broward Health Imperial Point, causing significant damage and blocking access to the hospital.*
>
> *Llama-2 13B (ground truth): Damaged*
> *Mistral 7B v1.0 (classified): Blocked*
>
> *Case 3: The University of Florida Field Laboratory's interior is flooded, causing significant damage to research equipment and infrastructure. The facility is closed indefinitely.*
>
> *Llama-2 13B (ground truth): Flooded*
> *Mistral 7B v1.0 (classified): Damaged*

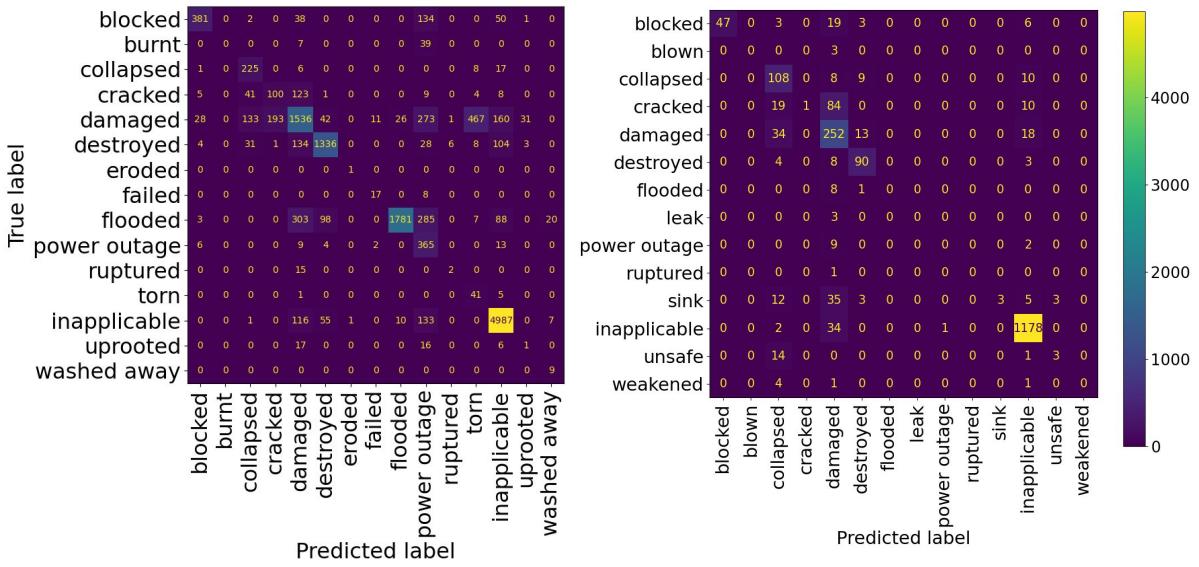


Figure 7. Confusion matrices for impact classification of retrieved tweets (left) Broward County and (right) Christchurch

Table 6. Performance scores for overall operational status classification on the retrieved results

	Overall Operational Status		
AOI	Precision	Recall	F1-score
Broward County	0.522	0.305	0.216
Christchurch	0.470	0.248	0.197

Case 4: Honey Hill Fire Station's front entrance crushed by fallen tree. Emergency responders working to clear debris and reach those trapped.

Llama-2 13B (ground truth): Collapsed
Mistral 7B v1.0 (classified): Blocked

In all cases, both ground-truth and classified labels seem reasonable with subtle differences stemming from how each LLM interprets the overall situation.

Among all the tasks performed by LLMs, the most challenging proved to be inferring the overall operational status of a CIF. Given that the retrieval process may include messages that are semantically akin to a CIF query, potentially encompassing both, messages about other CIFs and noise, the model tends to face confusion when deducing the overall operational status. We remark that enhancing the retrieval process will significantly improve the performance of the inference task.

LIMITATIONS AND FUTURE WORK

In this section, we describe the biases and limitations of this study. Our main source of bias comes from the dataset utilized in this study. Due to the unavailability of real-world curated datasets describing impacts to CIFs, we utilized Llama-2 13B to generate the relevant tweets as well as the matching labels. In Figure 2, we note that the distributions of impacts generated by Llama-2 13B mainly cover broader impacts such as *damaged, destroyed, flooded and collapsed*. Although Llama-2 13B also generated fine-grained impacts such as *shattered windows, ruptured pipe, HVAC damage and muddy road etc*, such distribution might not be representative of the distribution of impacts in real-world disaster. Likewise, the class/label imbalance in the dataset also limits us from fully accessing the capabilities of Mistral 7B v1.0 for the classification of various low-frequency classes. Moreover, the progression of impacts to CIFs in 24-hour time window was done by manually assigning the precedence of impacts to a CIF by the authors. This in part is also a bias that could have affected the prediction of the overall operational status performed in this study.

Among the limitations, the most predominant ones are the language usage, the choice of LLM models and the nature of disasters selected for this study. Unlike real-world tweets, the synthetic tweets generated by Llama-2 13B

were free from any typos, slang or grammatical errors. This study, therefore, does not fully assess the sensitivity of LLM-based classification to such inconsistencies. Moreover, we limited the type of disasters as well the choice of AOI used in this study to only two. This significantly limits the vocabulary of impacts as well as the nature of the content of the tweets in the generated dataset. The choice of AOI used, i.e., Christchurch and Broward County, also resulted in the generation and classification of only English medium tweets.

Likewise, this study is also limited to one particular LLM for the data generation i.e., Llama-2 13B and another LLM for the classification task i.e., Mistral 7B v1.0. Therefore, further analysis of the quality of data generation and classification of crisis-relevant tweets by other LLMs is deferred to our future work.

CONCLUSION

In this study, we investigated the application of Large Language Models to monitor the operational status of critical facilities in an area during large-scale disasters. We utilized social sensing data (both synthetic and real-world) from X (formerly Twitter) for two distinct locations and performed their analysis using LLMs for various computational tasks, including retrieval, classification, and inference. The outcomes are presented using standard evaluation metrics and further discussed to highlight the strengths and weaknesses of LLMs. In summary, we emphasize that the use of LLMs in the zero-shot setting, as compared to supervised models requiring training data, demonstrates reasonable performance. However, challenges arise for LLMs when dealing with complex contextual information provided through lengthy prompts. Additionally, we outline various future directions that we anticipate to be beneficial for the community to build upon.

REFERENCES

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

Auerswald, P., Branscomb, L. M., La Porte, T. M., Michel-Kerjan, E., & MICHEL-KERJAN, E. (2005). The challenge of protecting critical infrastructure. *Issues in Science and Technology*, *22*(1), 77–83.

Baloye, D. O., & Palamuleni, L. G. (2016). Modelling a critical infrastructure-driven spatial database for proactive disaster management: A developing country context. *Jàmbá: Journal of Disaster Risk Studies*, *8*(1), 1–14.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Castillo, C. (2016). *Big crisis data: Social media in disasters and time-critical situations*. Cambridge University Press.

Colverd, G., Darm, P., Silverberg, L., & Kasmanoff, N. (2023). Floodbrain: Flood disaster reporting by web-based retrieval augmented generation with an llm. *NeurIPS Workshop on Artificial Intelligence for Humanitarian Assistance and Disaster Response*.

Fan, C., & Mostafavi, A. (2019). A graph-based method for social sensing of infrastructure disruptions in disasters. *Computer-Aided Civil and Infrastructure Engineering*, *34*(12), 1055–1070.

Fekete, A., Tzavella, K., Armas, I., Binner, J., Garschagen, M., Giupponi, C., Mojtahed, V., Pettita, M., Schneiderbauer, S., & Serre, D. (2015). Critical data source; tool or even infrastructure? challenges of geographic information systems and remote sensing for disaster risk governance. *ISPRS International Journal of Geo-Information*, *4*(4), 1848–1869.

Grace, R. (2021). Toponym usage in social media in emergencies. *International Journal of Disaster Risk Reduction*, *52*, 101923.

Heglund, J., Hopkinson, K. M., & Tran, H. T. (2021). Social sensing: Towards social media as a sensor for resilience in power systems and other critical infrastructures. *Sustainable and Resilient Infrastructure*, *6*(1-2), 94–106.

Hu, Y., Mai, G., Cundy, C., Choi, K., Lao, N., Liu, W., Lakhanpal, G., Zhou, R. Z., & Joseph, K. (2023). Geo-knowledge-guided gpt models improve the extraction of location descriptions from disaster-related social media messages. *International Journal of Geographical Information Science*, *37*(11), 2289–2318.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, *47*(4), 1–38.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jovanović, A., Øien, K., & Choudhary, A. (2018). An indicator-based approach to assessing resilience of smart critical infrastructures. *Urban Disaster Resilience and Security: Addressing Risks in Societies*, 285–311.

Kumar, A., Singh, J. P., & Rana, N. P. (2017). Authenticity of geo-location and place name in tweets.

Labaka, L., Hernantes, J., & Sarriegi, J. M. (2016). A holistic framework for building critical infrastructure resilience. *Technological Forecasting and Social Change*, *103*, 21–33.

Mittelstädt, S., Wang, X., Eaglin, T., Thom, D., Keim, D., Tolone, W., & Ribarsky, W. (2015). An integrated in-situ approach to impacts from natural disasters on critical infrastructures. *2015 48th Hawaii International Conference on System Sciences*, 1118–1127.

Olteanu, A., Vieweg, S., & Castillo, C. (2015). What to expect when the unexpected happens: Social media communications across crises. *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 994–1009.

OpenAI, : Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., . . . Zoph, B. (2023). Gpt-4 technical report.

OpenStreetMap contributors. (2017). Planet dump retrieved from https://planet.osm.org.

Pescaroli, G., & Kelman, I. (2017). How critical infrastructure orients international relief in cascading disasters. *Journal of Contingencies and Crisis Management*, *25*(2), 56–67.

Qiang, Y. (2019). Flood exposure of critical infrastructures in the united states. *International Journal of Disaster Risk Reduction*, *39*, 101240.

Qu, Y., Huang, C., Zhang, P., & Zhang, J. (2011). Microblogging after a major disaster in china: A case study of the 2010 yushu earthquake. *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 25–34.

Roy, K. C., Hasan, S., & Mozumder, P. (2020). A multilabel classification approach to identify hurricane-induced infrastructure disruptions using social media data. *Computer-Aided Civil and Infrastructure Engineering*, *35*(12), 1387–1402.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yusta, J. M., Correa, G. J., & Lacal-Arántegui, R. (2011). Methodologies and applications for critical infrastructure protection: State-of-the-art. *Energy policy*, *39*(10), 6100–6119.

Zahra, K., Imran, M., & Ostermann, F. O. (2020). Automatic identification of eyewitness messages on twitter during disasters. *Information processing & management*, *57*(1), 102107.