Multimodal Disaster-related Tweet Classification with Parameter-Efficient Fine-Tuning of Large Language Models

Dongping Guo[1], Anh Tran[1], Xinli Xiao[2], Hongmin Li[1], and Doina Caragea[3]

[1] California State University, East Bay, Hayward CA 94542, USA,
dguo@horizon.csueastbay.edu, anhtranst@gmail.com,
hongmin.li@csueastbay.edu
[2] Arkansas Tech University, Russellville AR 72801, USA,
xxiao@atu.edu
[3] Kansas State University, Manhattan KS 66502, USA,
dcaragea@ksu.edu

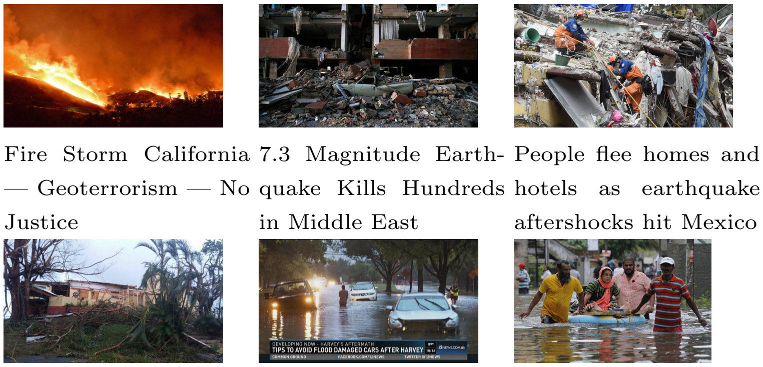
Abstract. This paper investigates the use of multimodal large language models (LLMs) for classifying social media posts in disaster response. Using the CrisisMMD dataset, we evaluate two proprietary models—GPT-4o and GPT-4o mini under zero-shot and few-shot setting, and evaluate the open-source LLaMA 3.2 11B across zero-shot, one-shot, and fine-tuned settings, on two classification tasks spanning seven real-world disaster events. Our results show that zero-shot multimodal LLMs demonstrate reasonable generalization, while one-shot and five-shot prompting do not consistently improve performance. In contrast, fine-tuned open-source models—especially LLaMA 3.2 11B, substantially outperform all zero- and few-shot settings, particularly on complex tasks. We also observe that the LLaMA 3.2 11B model struggles with few-shot multimodal inputs involving multiple images, resulting in a performance drop in one-shot experiments compared to zero-shot. Fine-tuning LLaMA 3.2 11B on both single- and multimodal inputs achieves state-of-the-art results. Moreover, lightweight text-only models such as LLaMA 3.2 1B and 3B, when fine-tuned, can match or surpass previously best-performing approaches. These findings underscore the value of task-specific fine-tuning and offer a cost-effective path for applying optimized multimodal LLMs in real-time disaster response. Our code is available at this link.

Keywords: multimodal large language models, multimodal disaster tweets classification, parameter-efficient fine-tuning

1 Introduction

Social media platforms, such as X (formerly known as Twitter), serve as real-time information hubs during disaster events, providing crucial updates on infrastructure damage, affected populations, and emergency response efforts. The value of such information has been widely recognized by both researchers and practitioners [39,12,29]. Despite its potential, leveraging social media data for real-time decision-making by local and federal agencies (e.g., FEMA, public health, police) and other stakeholders (e.g., NGOs) remains a challenge. The noise and multimodal nature of posts—combining text, images, and videos, often with misinformation—complicate extracting actionable insights.

One major effort to address these challenges involves leveraging machine learning and natural language processing techniques to automatically identify informative social media content and categorize it into relevant types, such as individuals seeking help or infrastructure damage. Various models, such as supervised deep learning models, semi-supervised learning models, and domain adaptation or transfer learning using pre-trained language models, have been explored for disaster-related tweet text classification [16]. In particular, there has been continuous progress towards classifying multimodal social media data for crisis analysis [25,36,32,21], as various multimodal models such as CLIP [31] have been proposed. Such models have often been evaluated on the CrisisMMD dataset [2,25], which consists of labeled tweets containing both text and images for seven disasters, such as California Wildfires. CrisisMMD supports two classification tasks: Informativeness classification and Humanitarian category classification. A few examples of tweets in the CrisisMMD dataset are shown in Fig. 1.



Fire Storm California — Geoterrorism — No Justice

7.3 Magnitude Earthquake Kills Hundreds in Middle East

People flee homes and hotels as earthquake aftershocks hit Mexico

Puerto Rico Temple “Almost Completely Destroyed” In Hurricane Maria

Tips to avoid flood-damaged cars after Harvey

SRI LANKA: Authorities race to rescue flood victims

Fig. 1: Examples of image-text tweets from the CrisisMMD dataset

With the rapid advancement of Small Pre-trained Language Models (e.g., BERT) and Large Language Models (LLMs)(e.g., LLaMA), there has been a surge in research exploring their applications in disaster response and social media crisis data analysis [37,22,38,35,33]. A recent survey on LLMs for disaster management [14] provides a comprehensive review of numerous studies in this field. However, most studies on disaster-related tweet classification focus on a single modality - tweet text alone [34,12,37,38]. Some studies have evaluated

LLMs under zero-shot and few-shot settings [12,37], while others have fine-tuned LLaMA 2 for crisis tweet classification [38], all using text-only data.

Among the few studies that have investigated multimodal LLMs for the two tasks of the CrisisMMD dataset, McDaniel *et al.* [22] evaluated GPT-4o, Gemini 1.5-flash-001, and Claude-3.5 Sonnet in the zero-shot setting but left detailed prompt engineering for future work. Giaccaglia *et al.* [8] proposed leveraging the multimodal LLM LLaVA [18] to generate captions for tweet images, which were then used alongside tweet text to train a classifier based on the RoBERTa model [19]. Their evaluation focused solely on the Informative task, but the model's performance did not surpass that of the fine-tuned CLIP model used by Mandal *et al.* [21].

To the best of our knowledge, no prior work has fine-tuned multimodal LLMs on multimodal disaster-related social media datasets, such as CrisisMMD. To fill this gap, in this paper, we first evaluate the zero-shot and one-shot capabilities of multimodal LLMs on disaster-related tweet classification tasks using the CrisisMMD dataset with some prompt engineering. Specifically, we assess the performance of GPT-4o [27], GPT-4o Mini [26], and LLaMA 3.2 11B [23]. GPT-4o is included due to its strong overall performance in prior studies [22,12], while GPT-4o mini offers a more cost-effective alternative with reduced size. LLaMA 3.2 11B is evaluated as a representative of smaller-scale open-source multimodal LLMs. We also explore the five-shot performance of the GPT-4o and GPT-4o mini models on the multi-class humanitarian category classification task. Note that for LLaMA 3.2 11B, the multimodal few-shot capability is limited as of now, as it is recommended to use only a single image during inference [20,24] and the model may not function reliably with multiple images, as noted by a member of the Meta LLaMA team [4]. Our experiments confirmed this limitation: the model's performance in few-shot settings dropped significantly compared to its zero-shot performance.

With that, we then explore Parameter-Efficient Fine-Tuning (PEFT), specifically Low-Rank Adaptation (LoRA) [11], to fine-tune the LLaMA 3.2 11B [23] and compare its results with the results of prior works. In addition, as Meta has open-sourced two lightweight text models, LLaMA 3.1 1B and 3B, designed for on-device use, we also explore fine-tuning these models with LoRA to assess their potential for disaster response applications on edge devices. Unlike full fine-tuning, PEFT methods such as LoRA and adapters [10] enable efficient adaptation of large models with limited computational resources, making them particularly suitable for real-time disaster response applications that are transparent and accessible to stakeholders. To summarize, our main contributions are as follows:

- We evaluate multimodal LLMs, including GPT-4o [27], GPT-4o mini [26], and LLaMA 3.2 11B [23], in the zero-shot and one-shot settings on the CrisisMMD dataset across two classification tasks. Our results demonstrate that these models perform well under zero-shot setting, with GPT-4o and

[4] Sanyam Bhutani, Meta. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct/discussions/43#66f98f742094ed9e5f5107d4

GPT-4o mini outperforming LLaMA 3.2 11B on both single-modality and multimodal inputs. In some cases, GPT-4o mini even outperforms GPT-4o and approaches the performance of previous best models (i.e., fine-tuned CLIP model in [21]) on this dataset. But one-shot prompting does not always improve performance, which is consistent with prior work findings [12].

- We fine-tune the LLaMA 3.2 11B vision model using LoRA on single-modality (text-only and image-only) as well as multimodal (text and image) data from the CrisisMMD dataset. Our fine-tuned models achieve state-of-the-art performance, surpassing prior models, zero-shot and one-shot proprietary models, which demonstrates that fine-tuning open-source multimodal models allows them to reach or exceed the performance of state-of-the-art proprietary models, while maintaining transparency and accessibility.
- We investigate the performance of fine-tuned lightweight text-only LLMs for disaster-related tweet classification, specifically LLaMA 3.2 1B and 3B fine-tuned with LoRA on the text data of CrisisMMD. These models outperform prior best-performing approaches, further demonstrating the effectiveness of fine-tuning even with smaller models. This makes it feasible to deploy these models in applications that can run on edge devices to aid disaster response and real-time decision-making.

2 Related Work

Disaster-related tweet classification has been extensively studied. Here, we focus on the most relevant works on multimodal models and LLMs in this domain.

Multimodal Disaster-Related Tweet Classification: As one of the most widely used benchmarks for multimodal disaster tweet classification, the CrisisMMD dataset contains tweets with both textual descriptions and accompanying images. Previous work has employed CNN-based architectures such as VGG-16+CNN [25], transformer-based models like BERT [1,13], multimodal models using attention and fusion techniques [6,28,32], and semi-supervised methods [36]. More recently, Mandal *et al.* [21] leveraged pretrained language-vision models, such as CLIP and ALIGN, to align text and image embeddings in a shared latent space. Their fine-tuned CLIP models achieved state-of-the-art performance on two tasks within the CrisisMMD dataset, outperforming all previous approaches. In this study, we directly compare our fine-tuned models with these fine-tuned CLIP baselines.

LLMs for Disaster-Related Tweet Classification: LLMs have shown promising results in disaster-related text classification. Imran *et al.* [12] evaluated six prominent LLMs, including GPT-4, GPT-4o, LLaMA-2 13B, LLaMA-3 8B, and Mistral 7B, using the HumAID dataset [4] under zero-shot and few-shot settings to assess their performance across different disaster types and information categories. Their findings indicate that proprietary models (e.g., GPT-4, GPT-4o) generally outperform open-source models (e.g., LLaMA and Mistral) across various tasks. However, GPT models struggle with flood-related data and all models face difficulties in classifying requests or urgent needs. Providing class-specific examples did not significantly enhance performance. Taghian Dinani *et* *al.* [37] evaluated LLaMA-2 in zero-shot and few-shot (one-shot and five-shot) settings on datasets including the CrisisBench dataset [5]. Their results show that while LLaMA-2 performed well in zero-shot and few-shot settings, fine-tuned models (CapsNet, BERT, and Bi-LSTM) still achieved superior performance, reinforcing the effectiveness of task-specific fine-tuning.

McDaniel *et al.* [22] evaluated GPT-4o, Gemini 1.5-flash-001, and Claude-3.5 Sonnet, with multimodal inputs (text and image) on the CrisisMMD dataset, as well as text-only inputs for other CrisisBench datasets. Their evaluation focused on zero-shot capabilities without detailed prompt engineering. Our experiments demonstrate that with some prompt engineering, GPT-4o achieves significantly improved performance on the informativeness category classification task with multimodal inputs (text and image). They also found that the informativeness classification task generally performed better without additional context, which is consistent with our findings.

Finally, Yin *et al.* [38] introduced CrisisSense-LLM, a fine-tuned LLaMA-2 model trained on CrisisBench for multi-label classification, a task distinct from the single-label classification approach used in our study. Their findings suggest that instruction fine-tuning significantly improves LLMs' ability to classify disaster tweets into multiple categories, such as event type, informativeness, and humanitarian involvement. Recent research has also explored retrieval-augmented generation (RAG) and domain-specific fine-tuning to enhance LLM performance in crisis settings [14]. However, most prior studies have focused solely on text-based models, leaving the multimodal aspect largely unexplored.

Parameter-Efficient Fine-Tuning(PEFT): PEFT [10], originally proposed for pre-trained language models such as BERT, has become an efficient and promising approach for fine-tuning Large Language Models (LLMs), including multimodal LLMs [40]. The goal of PEFT is to achieve performance comparable to that of full fine-tuning while modifying only a fraction of the backbone model's parameters or fine-tuning a small set of external parameters. There are three primary types of PEFT methods for LLMs: *adapter-based methods*, *prefix or prompt-based tuning*, and *reparameterization-based methods* such as LoRA.

In the adapter-based method, additional learnable modules are incorporated into the layers of the backbone model, either in a sequential [10] or parallel [9] manner. In prefix or prompt-based tuning, either a trainable tensor is added as a prefix to the input embeddings [15], or learnable vectors are introduced as prefixes or prompts in the hidden states of the backbone model layers [17,30]. In the reparameterization-based method, a low-rank decomposition technique is applied to transform model weight updates. Low-Rank Adaptation (LoRA) [11] follows this approach by updating the parameters of a weight matrix through the decomposition of its gradients into two low-rank matrices, thereby significantly reducing the number of trainable parameters while maintaining performance comparable to full fine-tuning. In our study, applying LoRA to the LLaMA 3.2 11B vision model enabled training of less than 1% of the model's approximately 11 billion parameters—specifically, about 0.2451% with a rank of 8, 0.4902% with a rank of 16 and 0.7352% with a rank of 24.

3 Fine-Tuning Methodology

We employ Low-Rank Adaptation (LoRA) [11], to fine-tune the LLaMA models in this study.

LoRA. The idea of LoRA is to use low rank matricies as the modifier of the parameters. More specifically, the original parameters $W_0 \in \mathbb{R}^{d \times k}$ is updated through low-rank decomposition using Equation (1), where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$:

$$W_0 + \mathit{\Delta} W = W_0 + BA \tag{1}$$

therefore reducing the number of parameters that are updated, while achieving comparable performance with full fine-tuning. The higher the rank (r value), the more parameters are updated. This approach enables the adaptation of large-scale vision-language models without modifying the weights of the entire model, which makes ideal for resource-constrained settings such as real-time disaster response.

4 Dataset

We utilize the real-world multimodal dataset **CrisisMMD** [2] in this study, which is publicly available and comprising tweets (with both tweets texts and images) collected during seven major natural disasters that happened in 2017: *California Wildfire, Hurricane Harvey, Hurricane Irma, Hurricane Maria, Iraq-Iran Earthquake, Mexico Earthquake, and Sri Lanka Floods.* CrisisMMD supports two classification tasks:

- **Tweet Informativeness Classification:** This task involves determining whether a tweet provides valuable crisis-related information or not. Each tweet is independently labeled for its text and image components as either: 1) *Informative*: contains relevant disaster-related information. 2) *Not-informative*: lacks crisis-related content.
- **Humanitarian Category Classification:** This task categorizes tweets into different humanitarian information types. The text and image of each tweet are annotated independently into one of the following eight categories: 1) *Affected individuals*; 2) *Injured or dead people*; 3) *Missing or found people*; 4) *Rescue, volunteering, or donation efforts*; 5) *Infrastructure and utility damage*; 6) *Vehicle damage*; 7) *Other relevant information*; 8) *Not humanitarian.* We refer readers to the original paper [3] for descriptions of the categories.

For the experiments, we use the same benchmark splits introduced by Ofli *et al.* [25], and only consider tweets where the text and image share the same label for consistency. Additionally, we followed the practice in Mandal *et al.* [21] and merged a few semantically similar classes to streamline classification. Specifically, *"injured or dead people"* and *"missing or found people"* were consolidated into *"affected individuals"*, while *"vehicle damage"* was grouped under *"infrastructure*

		Text				Image			
Task	**Category**	**Train**	**Dev**	**Test**	**Total**	**Train**	**Dev**	**Test**	**Total**
Informative	Informative	5,546	1,056	1,030	7,632	6,345	1,056	1,030	8,431
	Not-informative	2,747	517	504	3,768	3,256	517	504	4,277
	Total	8,293	1,573	1,534	11,400	9,601	1,573	1,534	12,708
Humanitarian	Affected individuals	70	9	9	88	71	9	9	89
	Rescue/Volunteering	762	149	126	1,037	912	149	126	1,187
	Infrastructure damage	496	80	81	657	612	80	81	773
	Other relevant	1,192	239	235	1,666	1,279	239	235	1,753
	Not-humanitarian	2,743	521	504	3,768	3,252	521	504	4,277
	Total	5,263	998	955	7,216	6,126	998	955	8,079

Table 1: Class distribution for *Informativeness, Humanitarian Category* task.

and utility damage". The class distributions across the train/dev/test splits of the dataset for the Tweet Informativness and Humanitarian Category tasks are presented in Table 1.

5 Experimental Setup

As the fine-tuned CLIP models in Mandal *et al.* [21] achieved the best results in prior work, we design our experiments to closely align with their setup for direct comparison. Specifically, we evaluate performances of the models on the two classification tasks of CrisisMMD using three input settings: 1) text-only inputs, 2) image-only inputs, and 3) text-image multimodal inputs. We also compare our zero-shot experiments with the GPT-4o in McDaniel *et al.* [22] for the informativeness task [5]. More specifically, we evaluate the following model settings:

- **Zero-shot and one-shot with GPT-4o**: The version or timestamp of the GPT-4o model is gpt-4o-2024-08-06.
- **Zero-shot and one-shot with GPT-4o mini**: The version or timestamp of the model is: gpt-4o-mini-2024-07-18.
- **Fine-tuned LLaMA 3.2 1B and 3B (Text-only)**: We fine-tune the two single-modality (text-only) models on the text-only versions of the two classification tasks by adding a sequence classification head—a linear layer—on top of the LLaMA transformer.
- **Zero-shot, one-shot and Fine-tuned LLaMA 3.2 11B**: We use the instruction-tuned version of this model.

In the one-shot setting, we randomly selected one example from each category to provide to the models. We also experimented with several simple prompts and used GPT-4o to suggest alternative formulations for prompt engineering. The final prompts were selected based on their clarity, including explicit instructions that defined the classification task and listed the category labels, followed by the

[5] For humanitarian task, since we combined a few categories in our experiments, we may not be able to directly compare with their zero-shot results on this dataset.

input modality (text only, image only, or both). For instance, for the humanitarian task, the instructions included brief explanations of each category to help the model better understand the label semantics. Full prompt details are available in our Github code repository. [6]. To further evaluate the few-shot capabilities of the models on the CrisisMMD dataset, we conducted five-shot experiments for the Humanitarian category classification task using the GPT-4o and GPT-4o mini models. For all experiments, we set temperatures to zero or close to zero.

For fine-tuning, we apply LoRA to all fine-tuned models, with some experiments conducted using the Unsloth library [7] to accelerate the process on an A6000 GPU. Due to time and resource constraints, we did not perform extensive hyper-parameter tuning but instead experimented with a few different configurations. The final models' performances on the test set were reported using hyper-parameters selected based on the development set. The chosen hyper-parameters values are in the Github code repository as well.

6 Experimental Results and Discussion

The main experimental results are presented in Table 2, where we report overall Accuracy as well as weighted Precision, Recall, and F1 scores to compare our models with the CLIP baseline. Results from selected five-shot experiments and per-category performance comparisons are provided in Table 3 and 4. More detailed results are available in our GitHub code repository.

6.1 Zero-shot and One-shot Multimodal LLMs

To assess the effectiveness of prompt engineering, we first compare our zero-shot GPT-4o results with the zero-shot GPT-4o results in McDaniel *et al.* [22] for the informativeness task. As we can see, with some prompt engineering, the model's performance improves significantly especially for the text and image multimodal inputs experiment (F1 score 87.71 vs 76.90).

We then analyze the zero-shot experimental results and compare them with the CLIP baseline. The findings indicate that multimodal LLMs demonstrate strong zero-shot capabilities for disaster-related tweet classification. Among the evaluated models, GPT-4o and GPT-4o mini significantly outperform LLaMA 3.2 11B, particularly in the zero-shot text+image setting. This aligns with the benchmark evaluations of multimodal LLMs which suggest the GPT-4o and GPT-4o mini models significantly outperform LLaMA 3.2 11B [7]. Interestingly, in the zero-shot setting, GPT-4o mini achieves the best overall performance across both tasks and various input modalities, while also being significantly more cost-effective. This contrasts with our earlier preliminary experiments, which showed

[6] https://github.com/deeplearning-lab-csueb/Fine-tune-Multimodal-LLM-for-CrisisMMD

[7] LLaMA 3.2 11B is likely the smallest model among the three. Since GPT-4o and GPT-4o mini are proprietary (closed-source), their exact parameter sizes are not publicly available.

	Informative Task				**Humanitarian Task**			
Text only	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1**
CLIP [21]	86.11	85.95	86.11	85.99	81.26	81.47	81.26	80.70
Zero-shot GPT 4o [22]	-	-	-	79.10	-	-	-	-
Zero-shot GPT 4o	79.47	81.38	79.47	79.93	74.66	75.03	74.66	74.27
Zero-shot GPT 4o-mini	85.98	86.69	85.98	85.2	74.24	77.97	74.24	75.07
Zero-shot LLaMA3.2-11B	83.57	83.28	83.57	83.22	74.45	78.43	74.45	74.09
One-shot GPT 4o	84.16	83.94	84.16	83.99	71.10	76.59	71.10	72.00
One-shot GPT 4o-mini	82.07	85.42	82.07	79.87	69.95	77.64	69.95	70.89
One-shot LLaMA3.2-11B	82.79	84.46	82.79	81.22	74.45	77.21	74.45	75.09
LLaMA3.2-1B-lora	87.48	89.16	92.62	90.86	80.63	80.73	80.63	80.66
LLaMA3.2-3B-lora	**88.72**	**90.39**	**93.11**	**91.73**	83.87	83.71	83.87	83.66
LLaMA3.2-11B-lora	88.33	88.22	88.33	88.24	**85.24**	**85.47**	**85.24**	**85.32**
Image only								
CLIP [21]	90.55	90.48	90.55	90.49	87.43	87.48	87.43	87.14
Zero-shot GPT 4o	85.14	85.58	85.14	85.29	84.71	87.18	84.71	85.28
Zero-shot GPT 4o-mini	87.29	87.21	87.29	87.24	84.50	86.58	84.50	85.12
Zero-shot LLaMA-3.2-11B	86.05	86.57	86.05	85.34	76.54	76.79	76.54	75.66
One-shot GPT 4o	86.38	86.38	86.38	86.38	84.29	87.16	84.29	85.04
One-shot GPT 4o-mini	88.72	88.01	88.72	88.37	84.50	86.68	84.50	85.13
One-shot LLaMA3.2-11B	73.92	79.88	73.92	67.77	71.83	74.45	71.83	69.68
LLaMA3.2-11B-lora	**92.31**	**92.28**	**92.31**	**92.29**	**89.95**	**89.95**	**89.95**	**89.91**
Text + Image								
CLIP [21]	93.15	93.12	93.15	93.13	90.22	90.23	90.22	90.04
Zero-shot GPT 4o [22]	-	-	-	76.90	-	-	-	-
Zero-shot GPT 4o	88.07	88.17	88.07	87.71	75.08	81.77	75.08	75.52
Zero-shot GPT 4o-mini	85.66	87.33	85.66	84.55	77.17	83.48	77.17	78.59
Zero-shot LLaMA-3.2-11B	79.99	83.93	79.99	77.07	74.24	76.05	74.24	73.59
One-shot GPT 4o	86.57	87.39	86.57	85.81	75.29	82.47	75.29	75.97
One-shot GPT 4o-mini	87.81	88.55	87.81	87.20	78.74	84.34	78.74	80.26
One-shot LLaMA3.2-11B	66.23	73.59	66.23	67.21	68.59	68.53	68.59	65.29
LLaMA3.2-11B-lora	**94.78**	**94.77**	**94.78**	**94.77**	**91.62**	**91.63**	**91.62**	**91.62**

Table 2: Performance comparison of models on Tweet Informativeness and Humanitarian Category tasks across different modalities.

GPT-4o to be superior. The observed shift may be due to updates made by OpenAI to the GPT-4o model, as others have also noted changes in its behavior. Nonetheless, the results underscore the effectiveness of proprietary multimodal models—especially when combined with prompt engineering—while also highlighting the need to fine-tune open-source alternatives to narrow the performance gap.

Finally, compared to the zero-shot setting, one-shot prompting does not consistently improve performance—particularly when images are used as input, and especially for the LLaMA 3.2 11B model. In image-only and text + image experiments, the LLaMA 3.2 11B model experiences a significant performance drop in the one-shot setting. This may be due to limitation we mentioned in the Section 1 that is pointed out by the LLaMA team member. For GPT-4o and GPT-4o mini, our results are consistent with the findings of Imran *et al.*[12], which suggest that adding few-shot examples does not always yield performance gains. We conducted five-shot experiments with both GPT-4o and GPT-4o mini on the Humanitarian classification task to further verify this. The results of the five-shot experiments, presented as weighted F1 scores, are shown in Table 3, alongside comparisons with zero-shot and one-shot settings.

	Text Only			**Image Only**			**Text + Image**		
Model	0-shot	1-shot	5-shot	0-shot	1-shot	5-shot	0-shot	1-shot	5-shot
GPT-4o	74.27	72.00	72.17	85.28	85.04	84.53	75.52	75.97	79.02
GPT-4o mini	75.07	70.89	72.76	85.12	85.13	85.46	78.59	80.26	78.63

Table 3: F1 Scores (%) of GPT-4o and GPT-4o mini on the Humanitarian classification task across different modalities and shot settings.

As shown in the results, adding more examples (i.e., one-shot to five-shot) does not consistently improve performance and can, in some cases—such as with GPT-4o mini on text+image inputs—even lead to a slight decline. Exploring strategies for selecting high-quality examples—or choosing the most relevant example for each test instance at inference time—could be a promising direction for improving few-shot performance.

6.2 Fine-Tuned Multimodal LLMs

From the results in Table 2, we can see that fine-tuning open-source multimodal LLMs significantly enhances their performance, surpassing both the baseline and zero-shot multimodal models across all three input modality settings. For example, in the text+image experiments on the informativeness task, LLaMA 3.2 11B-LoRA achieves an F1 score of 94.77, surpassing both the CLIP baseline (93.13) and the best-performing zero-shot model in this setting, GPT-4o (87.71). A similar trend is observed in the Humanitarian Category task, where the fine-tuned LLaMA3.2-11B-lora achieves an F1 score of 91.62, outperforming both the CLIP baseline (90.04) and best performing zero-shot model in this setting, GPT-4o mini (78.59). We believe with some more hyper-parameter tuning the results may further improve.

This demonstrates that fine-tuning open-source multimodal models allows them to reach or exceed the performance of state-of-the-art proprietary models, while maintaining transparency and accessibility. Additionally, these results reinforce the idea that while zero-shot multimodal LLMs show strong generalization, they still benefit significantly from task-specific fine-tuning.

To further analyze the benefits of the fine-tuned multimodal LLM models, we examined selected test set predictions, as shown in Figure 2, where the fine-tuned LLaMA3.2-11B model successfully corrected several misclassifications made by the zero-shot LLaMA3.2-11B model in the multimodal experiments. The improvements observed in these cases highlight the model's enhanced ability to integrate visual and textual information, leading to more accurate disaster related classification.

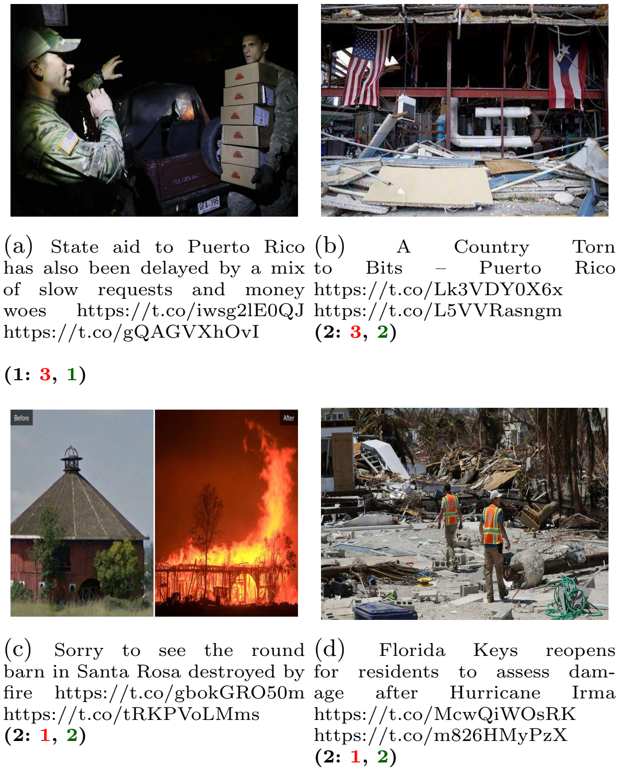
Description	**Zero-shot GPT-4o mini**			**LLaMA3.2-11B-lora**		
	Precision	Recall	F1	Precision	Recall	F1
Affected individuals	15.09	88.89	25.81	75.00	66.67	70.59
Rescue/Volunteering	61.11	96.03	74.69	88.19	88.89	88.54
Infrastructure damage	78.57	81.48	80.00	95.06	95.06	95.06
Other relevant	78.14	82.13	80.08	89.58	91.49	90.53
Not-humanitarian	93.57	69.25	79.59	93.19	92.26	92.72

Table 4: Model performance per category on the Humanitarian classification task using Text + Image multimodal input.

We present the per-category performance of the zero-shot GPT-4o mini model and the LoRA fine-tuned LLaMA 3.2 11B model on the Humanitarian classification task—specifically for the Text+Image setting—in Table 4, highlighting the performance gains achieved through fine-tuning across individual categories.

6.3 Fine-Tuning Lightweight Text-Only LLMs

Our results in Table 2 also show that fine-tuned lightweight text-only LLMs (e.g., LLaMA3.2-1B, 3B) consistently outperform the baseline CLIP model across both classification tasks. In the informativeness task, LLaMA3.2-3B-lora achieves the highest F1 score (91.73), surpassing CLIP (85.99) and all zero-shot and one-shot LLMs. Similarly, for the humanitarian task, LLaMA3.2-3B scores 83.66, outperforming CLIP (80.70). Compared to zero-shot and one-shot multimodal LLMs—fine-tuned lightweight models (LLaMA3.2-1B-lora, 3B-lora) perform better on both tasks, with especially large gains in the humanitarian task. This demonstrates the effectiveness and efficiency of fine-tuning smaller open-source models, which also offer a more cost-effective alternative to proprietary models like GPT-4o. Notably, on the informativeness task, we also tried fully fine-tuning LLaMA3.2-1B (i.e. tuning all 1B parameters), but that yielded worse results than its LoRA-tuned counterpart, highlighting the promise of parameter-efficient fine-tuning. Further experiments are needed to validate this, which we leave for future work.



(a) State aid to Puerto Rico has also been delayed by a mix of slow requests and money woes https://t.co/iwsg2lE0QJ https://t.co/gQAGVXhOvI

(1: 3, 1)

(b) A Country Torn to Bits – Puerto Rico https://t.co/Lk3VDY0X6x https://t.co/L5VVRasngm **(2: 3, 2)**

(c) Sorry to see the round barn in Santa Rosa destroyed by fire https://t.co/gbokGRO50m https://t.co/tRKPVoLMms **(2: 1, 2)**

(d) Florida Keys reopens for residents to assess damage after Hurricane Irma https://t.co/McwQiWOsRK https://t.co/m826HMyPzX **(2: 1, 2)**

Fig. 2: Sample CrisisMMD instances with their respective classes and LLaMA-3.2-11B predictions from the text + image (multimodal) experiments. Correct and incorrect predictions are highlighted in **green** and **red**, respectively. The first number represents the ground-truth category (y_{true}), the second denotes the zero-shot model prediction, and the third corresponds to the fine-tuned model's prediction. The category numbers are: 0 - Affected individuals, 1 - Rescue/Volunteering, 2 - Infrastructure damage, 3 - Other crisis relevant, 4 - Not-humanitarian.

7 Conclusion

In this study, we provide a comprehensive evaluation of multimodal LLMs for disaster-related tweet classification on the CrisisMMD dataset. Our results demonstrate that while zero-shot and one-shot multimodal LLMs, such as GPT-4o, achieve relatively strong performance, fine-tuned models consistently outperform them. Fine-tuning lightweight text-only models (LLaMA 3.2 1B and 3B) yields significant performance gains over prior approaches, making them cost-effective alternatives. Moreover, fine-tuning the vision-enabled LLaMA 3.2 11B on single and multimodal inputs achieves state-of-the-art results, highlighting the effectiveness of open-source models when optimized for specific tasks. These findings emphasize the importance of fine-tuning for real-world deployment and suggest that open-source models can serve as practical, scalable, and cost-effective alternatives to proprietary solutions in crisis-related social media classification.

In the future work, we plan to compare different Parameter-Efficient Fine-Tuning (PEFT) methods, such as LoRA versus adapters-based method, across more disaster-related datasets, particularly multimodal datasets, to better understand their effectiveness in diverse disaster scenarios. A systematic comparison of different PEFT techniques and their impact when tuning specific model components—such as vision-only, language-only, or both—will also provide deeper insights into optimizing model performance while minimizing computational costs.

Another important direction is investigating multimodal alignment in fine-tuned models to understand how textual and visual modalities interact post-fine-tuning and whether independent or joint tuning leads to better disaster classification. Given the limited availability of labeled crisis data, exploring self-training and semi-supervised learning techniques, such as pseudo-labeling with high-confidence predictions, together with multimodal LLMs may further improve model generalization. Additionally, conducting a cost-performance analysis of fine-tuned models versus zero-shot proprietary models will help identify the most practical solutions for real-world deployment. Finally, integrating these models into real-world crisis monitoring systems and gathering feedback from first responders and humanitarian organizations will be essential for assessing usability, reliability, and real-time decision-making impact.

References

1. Abavisani, M., Wu, L., Hu, S., Tetreault, J., Jaimes, A.: Multimodal categorization of crisis events in social media. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020)
2. Alam, F., Ofli, F., Imran, M.: Crisismmd: Multimodal twitter datasets from natural disasters. In: Proceedings of the Twelfth International Conference on Web and Social Media (2018)
3. Alam, F., Ofli, F., Imran, M.: Crisismmd: Multimodal twitter datasets from natural disasters. In: Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM) (2018)
4. Alam, F., Qazi, U., Imran, M., Ofli, F.: Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks. In: Proceedings of the International AAAI Conference on Web and social media (2021)
5. Alam, F., Sajjad, H., Imran, M., Ofli, F.: Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. In: Proceedings of the International AAAI conference on web and social media (2021)
6. Cheung, T., Lam, K.: Crossmodal bipolar attention for multimodal classification on social media. Neurocomputing (2022)
7. Daniel Han, M.H., team, U.: Unsloth (2023). URL http://github.com/unslothai/unsloth
8. Giaccaglia, P., Bono, C.A., Pernici, B., et al.: Enhancing emergency post classification through image information amplification via large language models. In: Proc.

Conf. on Information Systems for Crisis Response and Management (ISCRAM 2024), pp. 1–14 (2024)

9. He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., Neubig, G.: Towards a unified view of parameter-efficient transfer learning. In: International Conference on Learning Representations (2022)
10. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning (2019)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021). URL https://arxiv.org/abs/2106.09685
12. Imran, M., Ziaullah, A.W., Chen, K., Ofli, F.: Evaluating robustness of llms on crisis-related microblogs across events, information types, and linguistic features (2024). URL https://arxiv.org/abs/2412.10413
13. Krawczuk, P., Nagarkar, S., Deelman, E.: Crisisflow: multimodal representation learning workflow for crisis computing. In: 2021 IEEE 17th International Conference on eScience (eScience) (2021)
14. Lei, Z., Dong, Y., Li, W., Ding, R., Wang, Q., Li, J.: Harnessing large language models for disaster management: A survey (2025). URL https://arxiv.org/abs/2501.06932
15. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (2021)
16. Li, H., Caragea, D., Caragea, C.: Combining self-training with deep learning for disaster tweet classification. In: 18th International Conference on Information Systems for Crisis Response and Management (2021)
17. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (2021)
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (2023)
19. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR (2019)
20. Llama Team, Meta: The llama 3 herd of models (2024). URL https://arxiv.org/abs/2407.21783
21. Mandal, B., Khanal, S., Caragea, D.: Contrastive learning for multimodal classification of crisis related tweets. In: Proceedings of the ACM Web Conference 2024 (2024)
22. McDaniel, E., Scheele, S., Liu, J.: Zero-shot classification of crisis tweets using instruction-finetuned large language models (2024). URL https://arxiv.org/abs/2410.00182
23. Meta: Llama 3.2 model card (2024). https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md
24. Meta AI: Llama 3 and more: Advancing ai for vision, edge, and mobile devices (2024). URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Accessed: 2025-02-08
25. Ofli, F., Alam, F., Imran, M.: Analysis of social media data using multimodal deep learning for disaster response. In: 17th International Conference on Information Systems for Crisis Response and Management (2020)

26. OpenAI: Gpt-4o mini: Advancing cost-efficient intelligence (2024). URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2025-02-08
27. OpenAI: Gpt-4o system card (2024). URL https://arxiv.org/abs/2410.21276
28. Pranesh, R.: Exploring multimodal features and fusion strategies for analyzing disaster tweets. In: Proceedings of the Eighth Workshop on Noisy User-generated Text) (2022)
29. Purohit, H., Buntain, C., Hughes, A.L., Peterson, S., Lorini, V., Castillo, C.: Engage and mobilize! understanding evolving patterns of social media usage in emergency management (2025). URL https://arxiv.org/abs/2501.15608
30. Qin, Y., Wang, X., Su, Y., Lin, Y., Ding, N., Yi, J., Chen, W., Liu, Z., Li, J., Hou, L., Li, P., Sun, M., Zhou, J.: Exploring universal intrinsic task subspace for few-shot learning via prompt tuning. IEEE ACM Trans. Audio Speech Lang. Process. (2024)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning (2021)
32. Rezk, M., El-Madany, N.E., Hamad, R.K., Badran, E.F.: Categorizing crises from social media feeds via multimodal channel attention. IEEE Access (2023)
33. Salfinger, A., Snidaro, L.: Probing the consistency of situational information extraction with large language models: A case study on crisis computing. In: 2024 IEEE Conference on Cognitive and Computational Aspects of Situation Management (2024)
34. Sánchez, C., Abeliuk, A., Poblete, B.: Large language models in crisis informatics for zero and few-shot classification. ACM Trans. Web (2025). DOI 10.1145/3736160. URL https://doi.org/10.1145/3736160. Just Accepted
35. Shrestha, T.: Extracting actionable requirements from crisis event tweets for requirements engineers. Master's thesis, University of Calgary, Calgary, Canada (2025). Retrieved from https://prism.ucalgary.ca
36. Sirbu, I., Sosea, T., Caragea, C., Caragea, D., Rebedea, T.: Multimodal semi-supervised learning for disaster tweet classification. In: Proceedings of the 29th International Conference on Computational Linguistics (2022)
37. Taghian Dinani, S., Caragea, D., Gyawali, N.: Disaster tweet classification using fine-tuned deep learning models versus zero and few-shot large language models. In: Data Management Technologies and Applications (2024)
38. Yin, K., Liu, C., Mostafavi, A., Hu, X.: Crisissense-llm: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics (2025). URL https://arxiv.org/abs/2406.15477
39. Zhang, Y., Zong, R., Shang, L., Zeng, H., Yue, Z., Wei, N., Wang, D.: On optimizing model generality in ai-based disaster damage assessment: A subjective logic-driven crowd-ai hybrid learning approach. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (2023)
40. Zhou, X., He, J., Ke, Y., Zhu, G., Gutierrez Basulto, V., Pan, J.: An empirical study on parameter-efficient fine-tuning for MultiModal large language models. In: Findings of the Association for Computational Linguistics: ACL 2024 (2024)