A Multimodal, Multilingual, and Multidimensional Pipeline for Fine-grained Crowdsourcing Earthquake Damage Evaluation

Zihui Ma
New York University
New York, USA
zihuima@nyu.edu

Lingyao Li*
University of South Florida
Tampa, Florida, USA
lingyao@usf.edu

Juan Li
Google LLC
USA
leahlijuan1@gmail.com

Wenyue Hua
University of California, Santa Barbara
Santa Barbara, California, USA
wenyue.hua@ucsb.edu

Jingxiao Liu
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
jingxiao@mit.edu

Qingyuan Feng
New York University
New York, USA
qf2093@nyu.edu

Yuki Miura*
New York University
New York, USA
yuki.miura@nyu.edu

Abstract

Rapid, fine-grained disaster damage assessment is essential for effective emergency response, yet remains challenging due to limited ground sensors and delays in official reporting. Social media platforms offer a rich, real-time source of human-centric observations that enable rapid crowdsourced information collection, making them an ideal resource for emergency management assessment. However, the multimodal and unstructured nature of social media content presents significant challenges for traditional analytical methods. In this study, we propose a structured Multimodal, Multilingual, and Multidimensional (3M) pipeline that leverages multimodal large language models (MLLMs) to assess disaster impacts. We evaluate three foundation models across two major earthquake events using both macro- and micro-level analyses. Results show that MLLMs effectively integrate image-text signals and demonstrate a strong correlation with ground-truth seismic data. However, performance varies with language, epicentral distance, and input modality. This work highlights the potential of MLLMs for disaster assessment and provides a foundation for future research in applying MLLMs to real-time crisis contexts.

ACM Reference Format:

Zihui Ma, Lingyao Li, Juan Li, Wenyue Hua, Jingxiao Liu, Qingyuan Feng, and Yuki Miura. 2018. A Multimodal, Multilingual, and Multidimensional Pipeline for Fine-grained Crowdsourcing Earthquake Damage Evaluation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 18 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding authors.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Efficient and comprehensive disaster damage assessment is critical for informing emergency operations and disaster relief [28, 34, 45].Conventional techniques such as hazard models, expert inspections, and ground-based instruments have supported the characterization of post-disaster conditions [5, 49, 52]. Recently, social media crowdsourcing has emerged as an additional source of information [19, 28], offering large volumes, near-real-time insights from those affected communities [23, 27]. More importantly, social media offers passive human observations, often capturing nuanced perspectives such as emotional reactions, indoor damage, and first-hand observations [22, 28]. These human-centric signals add a layer of damage representation to the conventional methods.

However, earlier machine learning methods frequently relied on hand-crafted features and domain-specific models, which required significant manual effort to extract structured insight [10, 28, 38]. Moreover, they often lack the generalizability to apply across multiple disasters occurring in different locations with different languages, or involving varying damage levels, as models trained on one dataset (*e.g.*, data from a specific disaster or spoken language) may not perform well on another. Additionally, diverse multimodal inputs pose challenges for analysis.

Recent advances in foundation MLLMs have demonstrated potential for cross-modal and multilingual understanding across diverse data sources. They are pretrained using multiple languages, comprehensive text, being good at analyzing text, images, and multiple languages. Though promising, it is unclear whether MLLMs can support fine-grained damage assessment, including structural and environmental impacts, interior damage, and human experiences across different language regions. Moreover, their scalability and generalizability across disasters and geographies have not been systematically evaluated, as this could be critical for supporting disaster managers in implementing quick disaster relief.

To address these gaps,we propose a structured "Multimodal, Multilingual, and Multidimensional" (3M) pipeline integrating data collection, multimodal damage classification, and model evaluation. Our pipeline relies on the reasoning abilities of MLLM to extract interpretable, event-relevant insights from large-scale social media streams. The framework aligns spatial, temporal, and semantic cues from both text and imagery to infer when and where what happened. We evaluate this pipeline using two sudden-onset earthquake events: the 2019 Ridgecrest earthquake in California and the 2021 Fukushima earthquake in Japan. Across these two case studies, we assess three top-performing foundation models, including Gemini-2.5-Flash (hereafter *Gemini*) [50], LLaVA 3-8B (hereafter *LLaVA*) [20], and Qwen 2.5-VL-7B (hereafter *Qwen*) [41], to explore their ability to understand multilingual content, reason across modalities, and generate consistent damage-level predictions. The study aims to answer the following questions through macro- and micro-perceptions:

- Can MLLMs provide reliable and fine-grained damage assessments of textual and image information posted on social media after disasters?
- To what extent do MLLMs generalize across disaster contexts, with respect to factors such as input modality and prompt sensitivity?

Our findings suggest that MLLMs exhibit strong capabilities in event localization, image-text fusion, and perceptual damage estimation. The models correlate near-moderate positive (r=~0.5) to high (r=0.78) with ground-truth seismic intensity data and demonstrate interpretable reasoning patterns. However, we also observe variations in performance depending on linguistic context and event proximity. These findings highlight both the promise and limitations of current LLMs, and point toward future directions for model adaptation and disaster-specific fine-tuning.

2 Related Work

2.1 Earthquake Damage Assessment

Recent advances in earthquake damage assessment span physics-based models, machine learning, and new sensing modalities, each balancing trade-offs in accuracy, scalability, and timeliness. Traditional approaches, such as FEMA's HAZUS and P-58 frameworks [14, 43], rely on structural mechanics to estimate probabilistic damage and losses. While interpretable and robust, these methods are computationally demanding, depend on expert input, and often lack the spatial resolution and speed needed for rapid, localized assessments. Their reliance on coarse, regional building inventories and categorical outputs (e.g., "moderate" or "extensive" damage) limits their utility in dynamic, real-world disaster response. Moreover, their reliance on physical instrumentation limits deployment coverage and often excludes human-centered perspectives on impact.

Building on machine learning advances, researchers have begun exploring novel data sources, such as crowdsourced social media. Existing literature has used sentiment analysis [1, 24, 35, 48], topic modeling [27, 30, 32], and text classification [12, 60, 61] to support hazard monitoring, communication, damage assessment, and behavioral analysis [28]. Yet despite their promise, these sources are often used in isolation. Most existing frameworks do not integrate these diverse inputs into a unified pipeline. They are commonly limited to a single data type (text or image), rely heavily on English-language content, and lack systematic incorporation of damage granularity aligned with MMI levels. This leads to a fragmented understanding of earthquake impacts, with missed opportunities for timely, contextualized, and community-aware responses.

2.2 Multi-modal LLMs Applications

Multimodal foundation models have emerged as powerful tools for integrating diverse data types, revolutionizing capabilities across scientific domains. Models such as GPT-4V [59], Gemini [50], and Claude 3 [18] are capable of understanding and reasoning over multimodal data, including text, images, video, and numerical data, demonstrating remarkable performance in tasks requiring cross-modal understanding. These models have shown effectiveness in analyzing complex scientific imagery alongside textual annotations, enabling new approaches to data fusion in fields ranging from bioinformatics [25, 26, 57] to astronomy [33, 42].

The application of multimodal foundation models has expanded beyond traditional scientific domains to critical social applications, particularly in disaster response [15, 21, 37] and social media analysis [9, 51]. These models are leveraged to interpret structural damage by aerial imagery [17] and social media post analysis [46] to prioritize emergency response resources [62], using both visual and textual contents to achieve a nuanced understanding of real-time information during crisis events.

3 3M Pipeline

To achieve fine-grained earthquake damage assessment from social media, we develop the 3M pipeline, illustrated in Figure 1. The pipeline consists of three primary stages, and each component is detailed in the following subsections.

Data Preparation. Twitter (now rebranded as X) is a microblogging and social networking platform that allows users to share short messages known as "tweets." Since the data in this study are collected prior to the rebranding, we refer to the platform as "Twitter" and use the term "tweets" for consistency. This study focuses on two representative earthquake events: (1) the 2019 Ridgecrest earthquake in California and (2) the 2021 Fukushima earthquake in Japan. These cases are selected because they occurred in seismically active regions with established disaster response systems. For instance, the 2021 Fukushima earthquake struck an area with a history of major seismic events and a well-developed emergency infrastructure. Additionally, residents in both regions are known to frequently use Twitter to share real-time observations during such events.

Then, tweets are collected using the Twitter Search API in "near-real-time" with the keyword "earthquake." For the Ridgecrest event, tweets are collected from July 4 to 10, 2019; for the Fukushima event, from February 13 to 17, 2021. It is noted that no other significant seismic events were reported in the news during these periods. Therefore, we assume the collected tweets are primarily associated with or posted in response to the respective target events, although some unrelated messages may be present in the dataset. Following the compilation of the initial dataset, a filtering process is applied to identify tweets containing damage-related content. Guided by prior

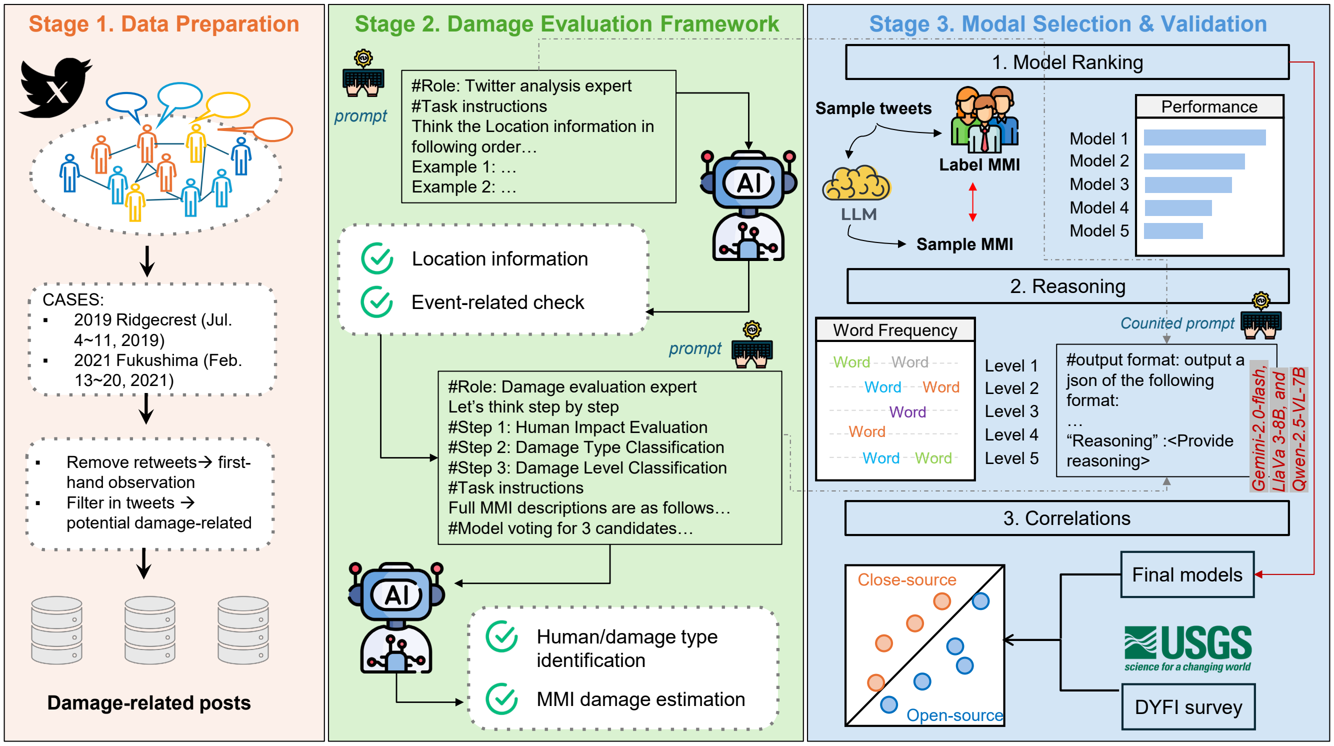

Figure 1: Proposed 3M pipeline, which integrates data preparation, damage evaluation framework, and model selection and validation for social media–based earthquake assessment.

research [22], we construct a library of filter terms (*e.g.*, "damage," "injury," "hurt," "die," "kill"), accounting for common word variants (*e.g.*, "damage," "damages," "damaged"). This filtering yields a refined dataset, referred to as the "damage-related dataset." After applying these criteria, the final dataset consists of 41,431 damage-related tweets for the 2019 Ridgecrest earthquake and 49,539 for the 2021 Fukushima earthquake, which are used for the subsequent analysis. The full list of filter terms is provided in the Appendix 3.

Damage Evaluation Framework. The evaluation of earthquake damage through social media content necessitates a structured and multi-stage analytical framework. For any given Twitter post, the assessment initially establishes event relevance through a two-fold verification process. First, spatial contextualization is conducted using a tiered approach that incorporates (1) geotag metadata, (2) content-based geographic references, and (3) user profile registration information. Among these, we prioritize geotagged metadata, which provides the most precise spatial signal [11, 47]. Among these, geotag metadata is prioritized for its spatial precision [11, 47].

When geotag data is unavailable or ambiguous, we rely on content-based inference (*e.g.*, mentions of place names or landmarks) and, subsequently, on user profile location. In cases where multiple geographic scales are mentioned (*e.g.*, city and neighborhood), the framework returns to the most granular location available. Second, the framework verifies the targeted seismic event to ensure analytical specificity.These two steps are operationalized through the `LOCATION_PROMPT` and `EVENT_PROMPT` modules, respectively.

Upon confirmation of relevance, the damage assessment protocol follows a hierarchical classification approach. The primary analysis differentiates between human-impact scenarios and non-human structural consequences. This bifurcation enables specialized examination of non-human impacts, which are further categorized into interior non-structural damage (*e.g.*, cracked interior walls, broken windows) and exterior structural damage (*e.g.*, building façade collapse, fallen infrastructure). It employs MLLMs to synthesize both textual narratives and visual documentation from social media posts. Based on the aggregated damage indicators, each post is assigned a Modified Mercalli Intensity (MMI) level. The MMI scale is a qualitative, ten-point system that characterizes earthquake intensity based on human perception and observable environmental and structural effects. Unlike instrumental magnitude scales, MMI provides a human-centered measure of impact, making it a widely adopted standard in post-earthquake reporting and risk communication. Detailed descriptors of the MMI scale used in this study are provided in the Appendix A.2. The use of MMI levels enables standardized comparisons of seismic impacts across geographic regions and disaster events. Damage-level classification is supported by two prompt modules tailored to input modality: the `IMAGE_ONLY_PROMPT` processes visual inputs exclusively, while the `TEXT_IMAGE_FUSION_PROMPT` integrates both textual and image-based information for multimodal assessment.

To guide model behavior throughout the pipeline, we incorporate two prompting strategies. First, we apply **chain-of-thought (CoT) prompting** [58] in modules that require structured reasoning, such as `LOCATION_PROMPT` and `TEXT_IMAGE_FUSION_PROMPT`

(Appendix A.6). These prompts instruct the model to "Let's think step by step", which helps it sequentially process location clues or synthesize human impact and structural damage evidence before assigning an MMI level. Second, we employ **few-shot learning** [4] in the `EVENT_PROMPT` module (Appendix A.6) to help the model discern earthquake-related tweets from irrelevant or sarcastic ones. The prompt includes three positive examples and two negative examples, forming a 5-shot configuration. These examples are selected to reflect diverse tweet styles, damage descriptions, and linguistic subtleties. Such diversity enhances the model's generalization and robustness when applied to noisy, real-world social media content.

Model Selection and Validation. This stage involves both quantitative and interpretive evaluation. We evaluate eight state-of-the-art multimodal foundation models, including leading commercial and open-source systems: GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, GPT-4o, GPT-4o-mini, Gemini-2.5-Flash, LLaVA 3–8B, and Qwen 2.5-VL-7B. These models are selected based on their reported performance in vision-language tasks and their accessibility for benchmarking [13, 54].

Using a randomly selected sample of damage-related tweets, each model generated MMI levels through the previous stage. Human-labeled ground-truth classes are based on the agreement of two independent annotators using the same damage framework. Pearson correlation scores are used to rank each model's performance in terms of alignment with official seismic intensity data. The full comparative results are provided in the Appendix A.3. Based on this analysis, Gemini, LLaVA, and Qwen were selected for following analysis, considering computational efficiency and practical deployment constraints. To assess the overall accuracy, we perform a correlation analysis between the model-generated MMI levels and ground-truth labels derived from the USGS "Did You Feel It?" (DYFI) survey [53], a crowdsourced platform that collects public reports of perceived shaking intensity following an earthquake.

Following quantitative validation, we further investigate the reasoning transparency of the top-performing models to understand how MLLMs estimate MMI levels. Specifically, we analyze the textual justifications generated by each model, focusing on the lexical features that underlie their classification decisions. To this end, we conduct a unigram-level TF-IDF analysis to identify high-weighted terms associated with different MMI levels. This analysis reveals the most influential words contributing to the model's classification decision. By analyzing the alignment between high-weighted terms and relevant damage descriptors, we assess whether a model's internal logic aligns with human-interpretable features.

4 Experiments and Results

In this section, we present the main experimental results and analysis. The first part focuses on macro-level evaluation at the pipeline level, including two earthquake case studies and an assessment of epicentral distance effects on model performance. The second part provides a micro-level analysis at the model level, examining impact of input modality, model prompt sensitivity, and detailed analysis of MLLM reliability based on CoT outputs.

4.1 Macro-level Analysis

2019 Ridgecrest Earthquake. Figure 2(a) shows the spatial distributions of social media-derived locations identified by three selected MLLMs in comparison to DYFI MMI scales. Overall, the results suggest that the models are capable of extracting relevant location and event information from tweets, as evidenced by the clustering of identified points near the earthquake epicenter (35.766°N 117.605°W). The spatial patterns for LLaVA and Gemini are especially consistent with the DYFI data, showing concentrated and geophysically plausible coverage in Southern California. Among the three models, Qwen demonstrates relatively weak performance in spatial coverage, with fewer identified points and reduced geographic spread. This may be due to its pretraining focus on Chinese-language data. In contrast, LLaVA and Gemini leverage more robust English-language foundations, resulting in improved extraction and localization performance for this event.

We further assess the models' ability to infer earthquake damage levels. Figure 3(a) presents the city-level correlations between model-estimated average damage levels and DYFI MMI data. All models show near-moderate to high positive agreements, as measured by Pearson correlation coefficients. Interestingly, Qwen achieves the highest correlation ($r = 0.78$), suggesting that although its spatial recall is limited, it may still be effective at identifying intensity-related cues from text and imagery. Meanwhile, LLaVA and Gemini exhibit wider coverage across MMI levels, highlighting their strength in detecting nuanced gradations in earthquake damage from multimodal content.

2021 Fukushima Earthquake. Similarly, we apply 3M pipeline to the 2021 Fukushima Earthquake in Japan with predominantly Japanese social media content. All three tested models demonstrate strong capabilities in extracting location- and event-related information. Most of the identified data points cluster near the earthquake epicenter (37.730°N, 141.595°E), and their spatial distributions align closely with the DYFI MMI data (Figure 2 (b)). All three models capture nearly the full range of earthquake-affected locations, indicating effective event localization and multilingual comprehension, even in a non-English setting. Their performance diverges when it comes to fine-grained damage level assessment. As shown in Figure 3 (b), Gemini exhibits a weak correlation between model-inferred damage levels and DYFI MMI scores ($r = 0.04$), suggesting limited effectiveness in interpreting Japanese-language damage cues. In contrast, LLaVA and Qwen achieve near-moderate correlations ($r = 0.47$ for both), reflecting a better understanding of MMI-scale damage in Japanese content. Although the overall correlation values for LLaVA and Qwen are similar, their strengths differ by intensity range. Qwen demonstrates a more precise differentiation between MMI levels 3 and 4, indicating sensitivity to moderate damage. LLaVA, on the other hand, performs more reliably in the lower MMI range (levels 1 to 3), suggesting its strength lies in detecting lighter damage events.

Epicenter Distance. We examine the correlation between estimated MMI levels and epicentral distance to assess the spatial sensitivity of model estimations, using results from the best-performing models for each case: Gemini for the Ridgecrest event and LLaVA

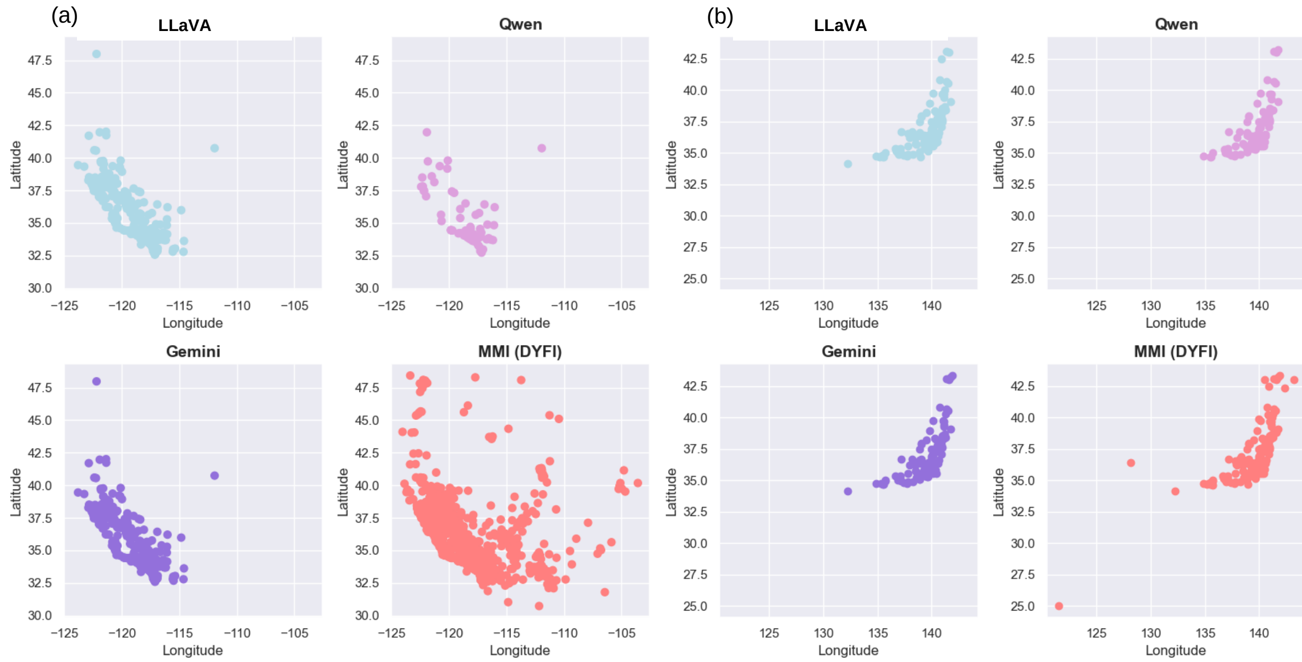

Figure 2: Spatial correlation of data points for (a) Ridgecrest and (b) Fukushima, as identified by LLaVA, Qwen, and Gemini, in comparison to DYFI MMI reports.

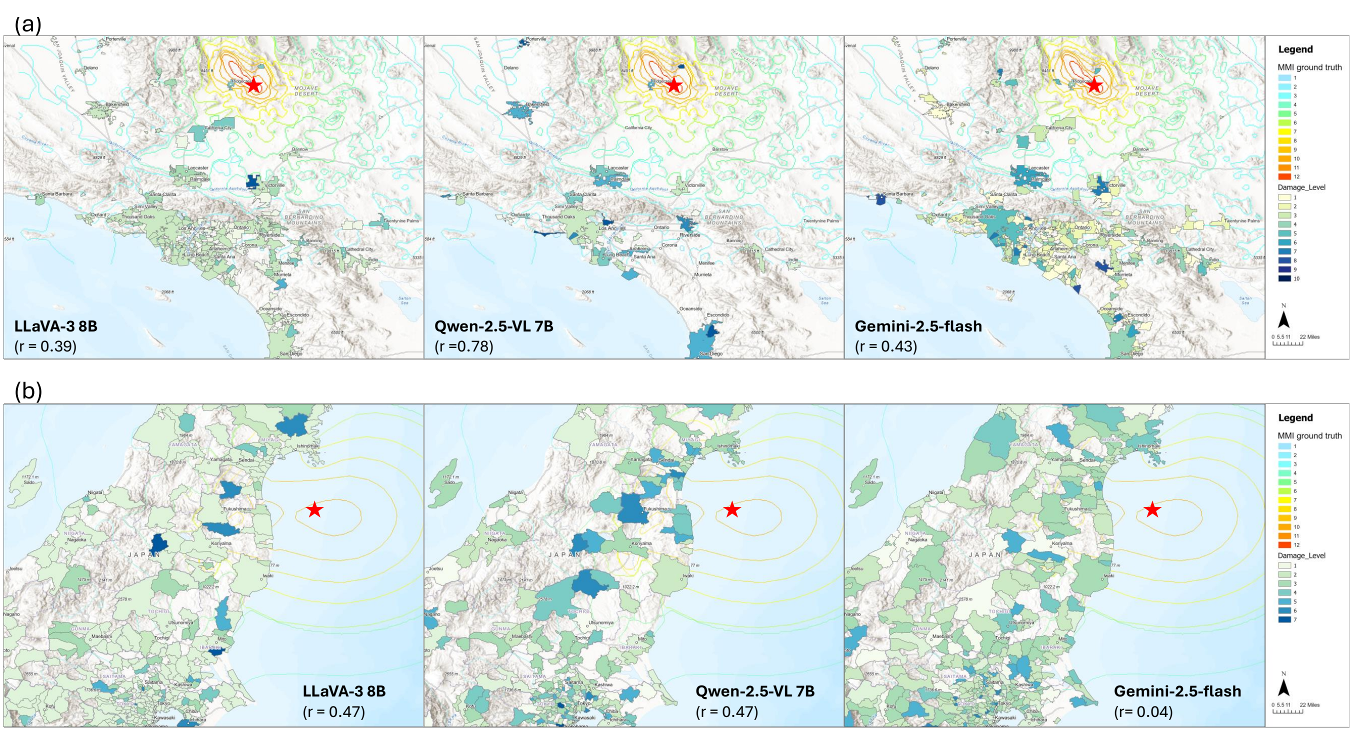

Figure 3: Spatial distribution of model-estimated average damage levels and DYFI MMI levels for (a) Ridgecrest and (b) Fukushima earthquakes, along with corresponding Pearson r scores. The red star symbol indicates the earthquake epicenter.

for the Fukushima event (Figure 4). In both cases, a negative correlation was observed, consistent with the principle of seismic attenuation, where shaking intensity typically decreases with increasing distance from the epicenter. The trend was stronger in the English-language Ridgecrest case, suggesting language familiarity may influence a model's ability to learn physically grounded patterns. Notably, Gemini identified a concentration of high-MMI predictions within 200 km of the epicenter, especially in densely populated areas (*e.g.*, Los Angeles), indicating its ability to focus on high-risk urban zones.

4.2 Micro-level Analysis

Input Modality. The choice of input modality directly influences the framework's evaluation performance. While social media platforms are primarily text-driven, the effectiveness of visual information and its combination with text for damage assessment remains underexplored. Thus, we evaluate model performance across three input configurations: text-only, image-only, and text-image fusion, as implemented in 3M pipeline. The `VOTE_PROMPT` module within Appendix A.6 is used to support ensemble-based selection of the most informative modality for final prediction. Figure 5 presents a line plot with error bars comparing predicted MMI levels against observed DYFI values across the three input modalities. In both earthquake cases, models that fused textual and visual content exhibited a strong, near-linear correlation with observed MMI, reinforcing prior findings in multimodal literature that show the benefit of cross-modal integration [29, 31, 56]. Conversely, models relying solely on visual inputs show diminished performance, particularly in the non-English Fukushima dataset, where image-only analysis was often based on non-damage-related content visuals, such as selfies, emojis, or screenshots, which lacked direct evidence of structural damage or event relevance. Text-only models tended to underestimate severity in cases where visual cues were absent.

Prompt Sensitivity. Given that variations in prompt phrasing could impact model performance and output quality in a non-trivial manner, it is crucial to evaluate the sensitivity of MLLMs to different prompt formulations [6, 44, 64]. This section builds on our earlier results by examining whether slight variations in prompts affect the models' outputs. To explore this, we randomly selected 50 tweets. These examples were processed using the three models selected from our previous experiments to ensure consistency in comparison.

For each tweet, we created seven paraphrased versions of the original prompt using GPT-4o. Since the primary goal of this work is to enable rapid and accurate earthquake damage estimation, and prior sections have shown that the text-image fusion configuration outperforms text-only or image-only alternatives, we focus our prompt sensitivity analysis on `TEXT_IMAGE_FUSION_PROMPT`. These paraphrases reworded the task instructions while preserving their semantic intent. All rewritten prompts were manually checked to ensure clarity and correctness. Modular prompting frameworks such as CRISPE [63], LangGPT [55], and CO-STAR [39] were not explored due to their domain-agnostic formulation and dependence on specialized prompt construction schemes that are less commonly used in rapid disaster response pipelines. We analyzed the impact of prompt variation across four output types: damage level, confidence

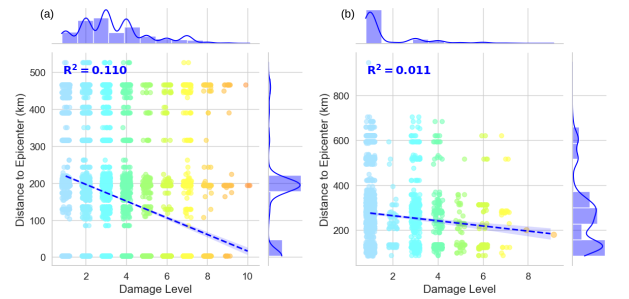

Figure 4: Epicentral distance vs. model estimated MMI values for the (a) Ridgecrest earthquake;(b) Fukushima earthquake.

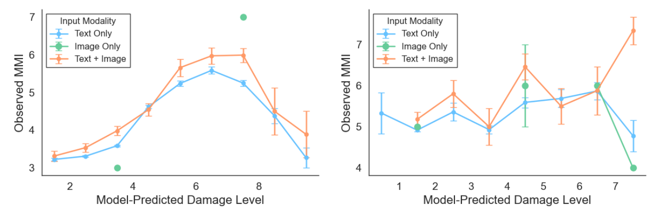

Figure 5: Correlation between model-estimated and DYFI MMI levels across input types for Ridgecrest (left) and Fukushima (right) earthquakes.

score, and categorical judgments such as damage type and human impact. For numerical outputs, we used the mean and standard deviation to measure variability. For categorical outputs, we used Cramér's V [8] (Equation 1) to quantify the frequency of label shifts across prompts, where values closer to 0 indicate low sensitivity and values closer to 1 indicate high sensitivity.

$$V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}} \tag{1}$$

where (1) V is Cramér's V coefficient (2) χ^2 is the chi-squared statistic derived from the contingency table (3) n is the total number of observations, and (4) k is the number of categories in the smaller of the two variables, *i.e.*, $\min(\text{rows}, \text{columns})$.

As shown in Table 1 and Table 2, models evaluated on the Fukushima dataset exhibit greater sensitivity than those tested on Ridgecrest data. This pattern was particularly pronounced for Gemini, which demonstrates substantial response variability when processing Japanese-language tweets. Conversely, Qwen displays the most stable performance, showing minimal variation in damage level assessments and confidence scores, though it exhibits greater inconsistency in damage type classification. LLaVA demonstrated moderate sensitivity across all output dimensions.

Despite variations in categorical classifications, most models maintain relatively stable MMI level predictions, with standard deviations typically ranging between 1 and 2. This indicates that while prompt formulation can influence specific classification details, overall assessments remain reasonably consistent. A similar pattern is observed in confidence scores, suggesting that models

Table 1: Cramér's V scores for human impact and damage type across prompt versions and models

Event	Prompt	Gemini		Qwen		LLaVA	
		Human Impact	Damage Type	Human Impact	Damage Type	Human Impact	Damage Type
2019 Ridgecrest earthquake	v1-v7	0.170	0.225	0.218	0.464	0.636	0.511
2022 Fukushima earthquake	v1-v7	0.502	0.771	0.224	0.624	0.578	0.587

Table 2: Damage level and confidence scores across prompt versions and models

Event	Prompt	Gemini				Qwen				LLaVA			
		DL_mean	DL_std	Conf_mean	Conf_std	DL_mean	DL_std	Conf_mean	Conf_std	DL_mean	DL_std	Conf_mean	Conf_std
2019 Ridgecrest earthquake	v1	3.333	1.325	0.786	0.073	3.600	1.694	0.850	0.059	1.867	1.548	0.853	0.090
	v2	2.795	1.490	0.847	0.098	2.588	2.311	0.887	0.054	0.769	1.945	0.915	0.141
	v3	2.317	1.572	0.906	0.101	2.606	2.304	0.911	0.066	0.769	2.026	0.962	0.070
	v4	2.683	1.980	0.894	0.094	3.100	1.919	0.883	0.091	1.867	1.388	0.915	0.097
	v5	3.095	1.923	0.875	0.091	3.152	2.224	0.809	0.109	2.104	2.479	0.840	0.184
	v6	2.683	1.559	0.848	0.082	3.212	2.162	0.800	0.080	1.940	0.752	0.876	0.072
	v7	1.930	1.486	0.900	0.105	4.188	2.583	0.883	0.066	2.720	2.777	0.832	0.118
2022 Fukushima earthquake	v1	3.838	1.041	0.722	0.062	3.222	1.502	0.850	0.069	3.333	1.325	0.786	0.073
	v2	3.563	1.105	0.731	0.098	2.667	2.075	0.837	0.099	3.167	1.324	0.782	0.100
	v3	2.333	1.875	0.942	0.063	2.444	2.470	0.815	0.151	3.167	1.998	0.915	0.060
	v4	4.000	1.732	0.797	0.077	3.124	2.378	0.805	0.076	2.625	2.042	0.863	0.078
	v5	4.630	1.245	0.687	0.071	2.667	1.637	0.873	0.070	2.167	2.681	0.633	0.334
	v6	2.214	1.528	0.786	0.063	3.955	2.645	0.763	0.127	0.167	0.817	0.852	0.260
	v7	1.243	0.760	0.801	0.092	4.239	2.508	0.787	0.239	1.000	2.703	0.850	0.269

maintain comparable levels of certainty regardless of instructional phrasing.

Reasoning Reliability Evaluation. To better understand how models arrive at their predictions, we conduct an analysis of the language used in their free-text justifications for estimated MMI levels, presenting a taxonomy of lexical patterns associated with different intensity levels. For the Ridgecrest earthquake (Figure 6 (a)), Gemini exhibits a progression in reasoning. At lower MMI levels (0–3), the model frequently uses terms such as "minimal," "preparation," "indoors," and "worries," suggesting a focus on psychological response and perceived safety. As the MMI increases to moderate levels (4–5), emotionally charged terms like "shock" and "fearful" become more common. At higher intensity levels (6–9), the model increasingly references concrete environmental and structural cues, using terms like "rockslides," "cracked," and "roadway." It later shifts toward cascading impact language with words like "fires" and "burned."

For the Fukushima earthquake (Figure 6(b)), LLaVA centers on perceived safety and emotional state, with terms like "visual," "safe," and "scary" at a lower MMI level. At moderate to higher MMI levels, the model references physical objects with increasing specificity, such as terms of "building," "ground," and "chair." For severe impacts, the model incorporates stronger terms such as "injured," "suspend," and "severely." Interestingly, LLaVA often uses hedging terms (e.g., 'possible', 'indicating'), suggesting a more cautious or probabilistic reasoning style.

5 Discussion

Can MLLMs provide reliable and fine-grained damage assessments using multilingual textual and image information posted on social media after disasters? Our experimental results demonstrate that state-of-the-art MLLMs possess substantial potential for fine-grained earthquake damage assessment. The effectiveness of the 3M pipeline across both English- and non-English-language contexts further demonstrates the multilingual capabilities of MLLMs. With appropriate language-aligned foundation models, the pipeline can be generalized to additional languages and extended to other disaster types (*e.g.*, wildfires, hurricanes) through prompt adaptations. This flexibility underscores the scalability of our approach across geographic and hazard domains.

Despite promising results, we observed some model-level performance variation. Qwen demonstrated the most consistent performance across languages, making it suitable for multilingual contexts, while Gemini and LLaVA excelled in urban, English-dominant settings. All models were more reliable at low to moderate damage levels, with reduced accuracy at higher intensities. It is likely due to training and data sparsity. Additionally, model estimation were influenced by epicentral distance, with better performance in densely populated urban areas. This pattern suggests that MLLMs capture attenuation effects but are also shaped by spatial disparities in social media activity. For real-world applications, decision-makers should account for these biases and consider complementary data sources or localized calibration when applying the 3M pipeline beyond densely populated regions.

To what extent do MLLMs generalize across disaster contexts, with respect to factors, such as input modality and prompt sensitivity? Our micro-level analysis further guides for the deployment of MLLMs in disaster contexts. First, modality analysis confirms that multimodal input fusion improve both accuracy and robustness in damage classification. We recommend extending

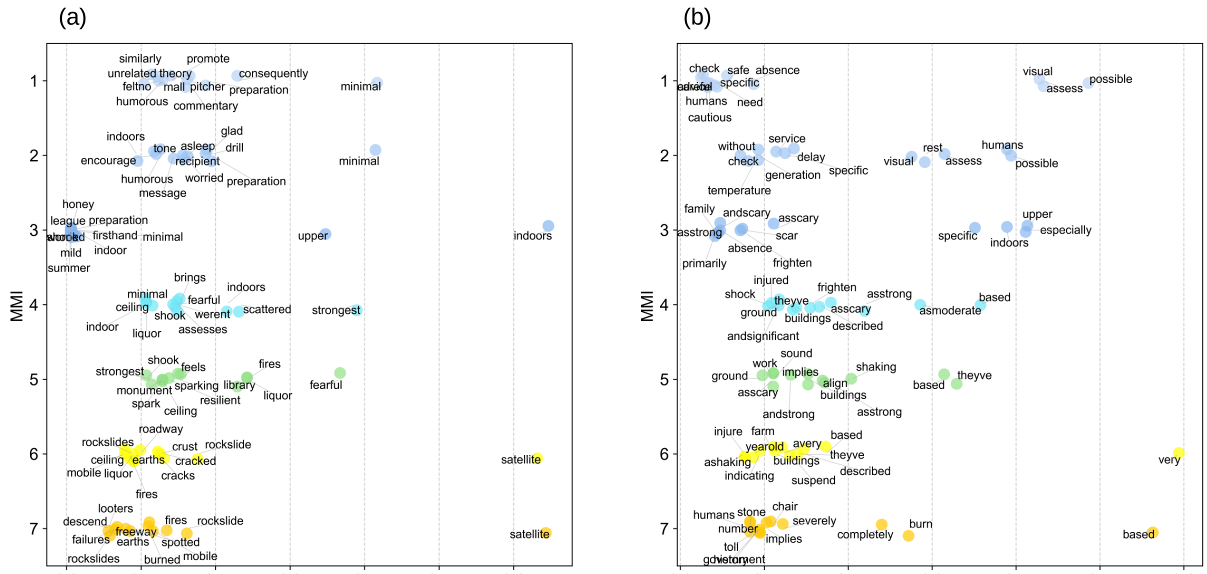

Figure 6: Reasoning reliability evaluation for (a) Ridgecrest and (b) Fukushima.

this approach to include cross-modal fusion of additional modalities such as video, audio, and geospatial data (*e.g.*, satellite imagery, street-level views). Second, prompt sensitivity evaluation reveals that current MLLMs exhibit variability in multilingual contexts, especially in response to subtle changes in instruction phrasing. While categorized classification outputs (*e.g.*, damage type, human impact) are relatively stable, inconsistencies may arise in edge cases. We recommend prompt standardization, pre-deployment testing, and ensemble prompting strategies to reduce sensitivity in multilingual or low-resource environments. Lastly, our reasoning analysis highlights differences in model interpretability and internal logic. For example, Gemini shifts from emotional to structural and cascading-impact cues as damage severity increases, while LLaVA adopts a more visually grounded but cautious reasoning style. These patterns suggest that decision-makers should consider not only performance metrics but also reasoning transparency and alignment with operational needs when selecting models for deployment.

6 Conclusion

This study introduces a structured 3M pipeline for social media–based earthquake damage assessment. The pipeline systematically integrates data preparation, multimodal classification, and model evaluation, providing a scalable framework for rapid and fine-grained disaster analysis. Applied to two real-world earthquake events, the pipeline demonstrates its effectiveness across languages, geographies, and damage dimensions. We also evaluate leading MLLMs and find that they effectively localize events, integrate text and image inputs, and produce damage estimates aligned with seismic data. However, performance varies by language, modality, and prompt design, highlighting the need for further adaptation and robustness testing in real-world deployments. Our findings provide the first step toward globally scalable, cross-lingual disaster sensing with foundation models, and the released codes and prompts to support replication and future research.

Broader Impact

Societal Relevance and Intended Use. This work presents a scalable and multilingual framework for fine-grained earthquake damage assessment that leverages social media and MLLMs. By incorporating both textual and visual data, the system captures dimensions of disaster impact such as interior structural damage or personal distress, that are difficult to observe with conventional sensing systems. Our pipeline offers a lightweight, extensible tool for situational awareness, particularly in the early hours of a crisis when actionable information is limited. Through evaluations in Japan and the United States, we demonstrate the framework's potential for global applicability. The methods and code are intended to be adaptable for other hazards (*e.g.*, floods, wildfires) and use cases (*e.g.*, infrastructure monitoring, rapid needs assessment).

Inclusivity and Linguistic Diversity. Disaster communication varies significantly across languages and cultures. Our framework is intentionally designed to support multilingual and multimodal inputs, allowing for more inclusive analysis across different user populations and platforms. The case study in Japan highlights the feasibility of applying foundation models beyond English-only settings, contributing to the growing body of work on equitable and linguistically diverse NLP applications. We encourage further development toward supporting low-resource languages and culturally grounded interpretations of crisis content.

Interpretability and Human-AI Collaboration. We use prompting strategies such as chain-of-thought and few-shot examples to improve the transparency of multimodal model outputs. In addition to comparing predictions with official ground-truth seismic data (*e.g.*, MMI levels), we include qualitative reasoning traces to assist with human interpretation. These steps enhance trust and traceability in model behavior, while positioning the system as a decision-support tool, not a replacement for expert review. This approach supports the responsible integration of LLMs into high-stakes environments like emergency management.

Ethics

Responsible Data Use and Privacy. All data used in this study are drawn from publicly available social media posts, accessed via Twitter's API under permitted use. Recognizing that disaster-related content is often shared under emotional duress, we employ several safeguards: no direct quotes or images are reproduced, user identifiers are removed, and results are reported only in aggregated spatial formats. Future deployments may benefit from further privacy-preserving measures such as differential privacy or on-device inference, particularly in operational settings.

Robustness and Misinformation Risks. Crisis-related social media can contain misinformation, rumors, or manipulated content. Our pipeline currently includes relevance filtering and heuristics for disaster-date alignment, but does not yet implement automated credibility detection. We view this as a key future direction, and recommend integration with source trustworthiness scoring and stance detection models for robust performance in noisy environments. These safeguards are particularly important in deployments where system outputs influence resource allocation or public messaging.

Scope of Use and Deployment Guidance. This pipeline is developed exclusively for public-interest applications such as disaster response, risk analysis, and resilience planning. It is not intended for use in surveillance, punitive actions, or insurance investigations. Responsible deployment requires human oversight, transparency about model limitations, and collaboration with emergency professionals and affected communities. We advocate for community-informed design and transparent documentation as this framework is adapted for real-world use.

A Appendices

Limitations

While this study provides a scalable and generalizable pipeline for multimodal earthquake damage assessment, it has several limitations that should be considered when interpreting the results. First, the use of social media introduces inherent sampling biases. Prior studies have shown that Twitter users are disproportionately younger, more educated, urban, and male, which limits the demographic representativeness of the data [40]. This population bias can reduce the generalizability of findings, particularly in contexts where equitable disaster response is critical. Moreover, disparities in internet access and digital infrastructure further constrain data coverage. Global digital divides and infrastructural disruptions in disaster-affected regions may result in missing or delayed social media signals. These conditions reduce the utility of social media as a ground-level information source during large-scale disasters.

Second, the data retrieval process itself imposes restrictions. In our study, tweets were collected using a single keyword ("earthquake") and filtered using a manually defined set of damage-related terms. While this approach provides a focused dataset, it may miss relevant posts that use alternative vocabulary or regional expressions. Consequently, reliance on fixed keyword libraries can limit recall and introduce topic filtering bias, especially across languages and local dialects.

Third, our dataset is naturally skewed toward urban and high-connectivity regions such as Ridgecrest and Fukushima, where social media activity is higher. This urban-centric bias reflects broader challenges in social sensing, where digitally underconnected or rural communities are often underrepresented. Consequently, the current model may generalize less effectively to areas that may experience equal or greater need for rapid post-disaster assessment and aid. To address this, future work will explore weak supervision and embedding-based retrieval to capture indirect or culturally grounded language. We also plan to incorporate alternative data streams such as SMS-based reports, sensor networks, and multilingual sources to improve rural coverage, and will develop geographic fairness metrics to evaluate and mitigate representational bias.

Fourth, we utilize foundation MLLMs without domain-specific fine-tuning. While this showcases the models' zero-shot and few-shot capabilities, fine-tuning on curated disaster-specific corpora, particularly multilingual datasets, may further enhance robustness, contextual understanding, and prediction accuracy.

Fifth, our full-scale evaluation is limited to three selected models (from an initial pool of eight), chosen to balance performance and computational feasibility. This reflects practical considerations for real-time deployment in resource-constrained settings, such as embodied agents or edge devices. Nevertheless, broader evaluations with larger or instruction-tuned models may yield different insights and should be explored in future work.

Sixth, the current study only focuses on earthquake events rather than multi-disaster generalization (e.g., floods, wildfires, hurricanes). This focus is intentional, as earthquakes provide a relatively well-structured testbed with quantifiable impacts such as shaking intensity and access to benchmark datasets like DYFI. However, the underlying framework is designed to be scalable and adaptable to other disaster scenarios in future work.

Seventh, a key challenge in social media-based sensing is filtering misinformation and low-quality signals. To mitigate this, our framework integrates prompt-based relevance filtering (`EVENT_PROMPT` and `LOCATION_PROMPT`), along with consistency checks across modalities (text, image, and metadata). Posts lacking coherent damage context or containing emotionally charged but irrelevant content (e.g., memes, selfies) are deprioritized or excluded. While this approach does not guarantee full immunity to misinformation, it increases robustness against noise by emphasizing posts with structurally or semantically aligned disaster evidence. Future versions will explore integrating credibility signals, community-based fact-checking cues, and external knowledge graphs to enhance misinformation resistance.

Finally, we use human-reported DYFI data as the primary ground truth for evaluation. While widely adopted in earthquake research, these crowd-sourced labels are inherently subjective and may vary due to perceptual or reporting inconsistencies. Incorporating additional validation sources—such as structural damage inspections, seismic sensor data, or post-event engineering assessments—would provide a more robust benchmark for future evaluations.

A.1 Damage-related filtering terms

The damage-related term list shown in Table 3 serves as an initial heuristic for filtering potentially relevant tweets. While this static keyword approach offers high precision, we acknowledge that it may fail to capture damage reports expressed through colloquial, metaphorical, or culturally specific language. To address this limitation, our framework incorporates multiple downstream filtering stages that rely on multimodal large language models (MLLMs). These include the `EVENT_PROMPT` and `LOCATION_PROMPT` modules, which perform semantic and visual reasoning to identify earthquake relevance beyond keyword matching.

A.2 MMI description

Table A.2 presents the Modified Mercalli Intensity (MMI) scale, which qualitatively describes earthquake shaking intensity based on observed effects on people, furnishings, buildings, and the natural environment. The scale ranges from I to X, with increasing severity reflected in stronger human reactions, greater object displacement, and more extensive structural and environmental damage. Each level provides a narrative combining perceptual reports and physical observations, offering a ground-level view of earthquake impacts.

A.3 Model comparison

Two annotators independently labeled a randomly selected sample of 50 tweets to evaluate inter-annotator reliability. We used Krippendorff's Alpha (α) (Equation 2) to measure agreement, as it is a robust metric capable of handling multiple annotators, various data types (*e.g.*, nominal, ordinal), and missing data [2]. It also adjusts for chance agreement based on observed versus expected disagreement. The final alpha score was 0.67, indicating substantial agreement. This level of consistency is considered reasonable for subjective tasks involving nuanced, fine-grained classification.

$$\alpha = 1 - \frac{D_o}{D_e} \tag{2}$$

Observed disagreement D_o is calculated as:

$$D_o = \frac{1}{N} \sum_{i=1}^{N} \delta(a_{i1}, a_{i2}) \tag{3}$$

where:

- N is the number of items,
- a_{i1}, a_{i2} are the annotations by two coders for item i,
- $\delta(a, b) = 1$ if $a \neq b$, and 0 if $a = b$.

Expected disagreement D_e is computed from the marginal frequencies:

$$D_e = \sum_{c_1 \neq c_2} p(c_1) \cdot p(c_2) \tag{4}$$

where:

- $p(c) = \frac{n_c}{2N}$ is the proportion of annotations assigned to category c,
- n_c is the total number of times category c is used by both annotators,
- the denominator $2N$ is the total number of annotations across both coders.

Interpretation:

- $\alpha = 1$: perfect agreement
- $\alpha = 0$: agreement equals chance
- $\alpha < 0$: worse than chance

To assess the cost-effectiveness of closed-source MLLMs, we monitored pricing across all eight evaluated models. Among them, Gemini-2.5-Flash was the most economically efficient and also demonstrated high alignment with human annotations. As a result, it was selected as the preferred closed-source multimodal model for our damage estimation tasks. For large-scale processing, we utilized the New York University High Performance Computing (NYU HPC) infrastructure, specifically the Greene cluster, which offers GPU-enabled nodes with NVIDIA Tesla V100 GPUs [36]. Within this environment, the complete analysis was executed in 2 to 3 days per event dataset.

A.4 Example of LLMs outputs

The following table 6 presents a representative example of how the three selected MLLMs—LLaVA, Qwen, and Gemini—analyze a tweet containing both text and image information. All three models accurately identify the location (El Monte, CA) and confirm the tweet's event relevance. While their MMI level estimates and confidence scores are similar, the models differ slightly in how they classify damage type (interior vs. exterior) and assess human impact. The reasoning outputs provide further insight into each model's interpretive process, revealing how text and image inputs are integrated to support the final prediction. This example highlights the overall consistency of model outputs while also illustrating subtle differences in how damage is inferred from multimodal content.

A.5 Satellite image vs MLLMs results

Remote sensing offers critical supplementary insights into the environmental repercussions of seismic events. Building upon our prior reasoning analysis, we conduct an in-depth evaluation of environmental impacts, with a particular emphasis on damage typologies, to further assess the model's inferential capabilities.

We acquire remote sensing data through the utilization of the Scene Classification Map (SCM) derived from Sentinel-2 Level-2A products[7]. The SCM is generated via the Sen2Cor processor, which implements a series of threshold-based assessments on top-of-atmosphere reflectance data across multiple spectral bands to categorize each pixel into predefined classes, including vegetation, water, soil/desert, snow, clouds, and shadows[3]. This classification facilitates the differentiation of land cover types and the detection of alterations attributable to seismic disturbances. The SCM is available at spatial resolutions of 20 m and 60 m and encompasses quality indicators for cloud and snow probabilities[16].

The correlation between tweets referencing exterior damage and the ground-truth MMI values is presented in Table 7. We first

Table 3: A list of terms used to filter "damage-related" tweets.

Language	Damage-related words
English	blackout, broke, broken, burn, burned, burning, burns, catastrophe, catastrophes, catastrophic, chaos, collapse, collapsed, collapses, crack, cracked, cracking, cracks, crash, crashed, crashes, cripple, cripples, crumble, crush, crushed, crushes, damage, damaged, damaging, dead, death, deaths, deform, deformed, deforms, demonish, destruct, destructed, destructing, destructs, destroy, destroyed, destroying, destroys, devastate, devastated, devastates, devastating, die, died, dies, displace, displaced, disrupt, disrupted, disrupting, disrupts, fatalities, fatality, fissure, fissures, fire, flood, flooded, flooding, hurt, hurting, hurts, injuries, injured, injury, kill, killed, killing, leak, leaked, leaking, leaks, massive, outage, rockslide, rubble, rupture, ruptures, safe, safety, scatter, scattered, scatters, severe, shatter, shattered, shatters, smash, smashed, smashes, smashing, suffer, suffered, suffering, suffers, trauma, warp, warps, wreck, wrecked, wrecks
Japanese	停電(blackout，poweroutage), 壊れた(broke, broken), 燃える(burn), 燃えた(burned), 燃えている(burning), 大災害(catastrophe, catastrophes), 壊滅的(catastrophic), 混乱(chaos), 崩壊(collapse, collapsed, collapses), ひび(crack, cracked, cracking, cracks), 墜落(crash, crashed, crashes), 無力(cripple, cripples, helpless), 崩れる(crumble), 押しつぶす(crush, crushed, crushes), 損傷(damage, damaged, damaging), 死んだ(dead, died, die, dies), 死亡(death, deaths), 変形する(deform, deformed, deforms), 破壊(destruct, destructed, destructing, destructs), 破壊する(destroy, destroyed, destroying, destroys), 壊滅させる(devastate, devastated, devastates, devastating), 死ぬ(die, died, dies), 避難する(displace, displaced), 混乱する(disrupt, disrupted, disrupting, disrupts), 死者(fatalities, fatality), 裂け目(fissure, fissures), 火事(fire), 洪水(flood, flooded, flooding), 傷つく(hurt, hurting, hurts), けが(injuries, injury), 負傷した(injured)，殺す(kill, killed, killing), 漏れ(leak, leaked, leaking, leaks), 巨大な(massive), がけ崩れ(rockslide), 土砂崩れ(Landslide), 瓦礫(rubble), 破裂(rupture, ruptures), 安全(safe), 散らす(scatter, scattered, scatters), 厳しい(severe), 粉々にする(shatter, shattered, shatters), 打ち砕く(smash, smashed, smashes, smashing), 苦しむ(suffer, suffered, suffering, suffers), トラウマ(trauma), ゆがむ(warp, warps)

Table 4: MMI Intensity

MMI	Description
I	**People's Reactions:** Not felt. **Natural Environment:** Changes in level and clarity of well water are occasionally associated with great earthquakes at distances beyond which the earthquakes felt by people.
II	**People's Reactions:** Felt by a few. **Furnishings:** Delicately suspended objects may swing.
III	**People's Reactions:** Felt by several; vibration like passing of truck. **Furnishings:** Hanging objects may swing appreciably.
IV	**People's Reactions:** Felt by many; sensation like heavy body striking building. **Furnishings:** Dishes rattle. **Built Environment:** Walls creak; windows rattle.
V	**People's Reactions:** Felt by nearly all; frightens a few. **Furnishings:** Pictures swing out of place; small objects move; a few objects fall from shelves within the community. **Built Environment:** A few instances of cracked plaster and cracked windows within the community. **Natural Environment:** Trees and bushes shaken noticeably.
VI	**People's Reactions:** Frightens many; people move unsteadily. **Furnishings:** Many objects fall from shelves. **Built Environment:** A few instances of fallen plaster, broken windows, and damaged chimneys within the community. **Natural Environment:** Some fall of tree limbs and tops, isolated rockfalls and landslides, and isolated liquefaction.
VII	**People's Reactions:** Frightens most; some lose balance. **Furnishings:** Heavy furniture overturned. **Built Environment:** Damage negligible in buildings of good design and construction, but considerable in some poorly built or badly designed structures; weak chimneys broken at roof line, fall of unbraced parapets. **Natural Environment:** Tree damage, rockfalls, landslides, and liquefaction are more severe and widespread with increasing intensity.
VIII	**People's Reactions:** Many find it difficult to stand. **Furnishings:** Very heavy furniture moves conspicuously. **Built Environment:** Damage slight in buildings designed to be earthquake resistant, but severe in some poorly built structures. Widespread fall of chimneys and monuments.
IX	**People's Reactions:** Some forcibly thrown to the ground. **Built Environment:** Damage considerable in some buildings designed to be earthquake resistant; buildings shift off foundations if not bolted to them.
X	**Built Environment:** Most ordinary masonry structures collapse; damage moderate to severe in many buildings designed to be earthquake resistant.

observe that Gemini demonstrates a stronger capability in identifying exterior damage impacts, as it detected the highest number of relevant tweets. Additionally, Gemini achieved relatively higher performance in estimating exterior damage severity, showing weak to moderate positive Pearson correlation scores, whereas the other

Table 5: Model comparison

Model Name	Open Source	Accuracy	Price ($)
GPT-4.1	✗	0.694	0.45
GPT-4.1-mini	✗	0.145	0.02
GPT-4.1-nano	✗	-0.841	0.02
GPT-4o	✗	0.957	0.85
GPT-4o-mini	✗	0.706	0.15
Gemini-2.5-Flash	✗	0.775	0.15
LLaVA 3-8B	✓	0.113	0.00
Qwen 2.5VL-7B	✓	0.791	0.00

two models exhibited only weak correlations. These findings suggest that future research should consider integrating remote sensing data, particularly for assessing damage to natural and built environments, to further enhance the accuracy of MMI scale estimation.

A.6 Prompt design

```
LOCATION_PROMPT = """
Task:
You are a location identification expert. Your task is to determine whether a tweet is from a U.S.-based location, based on all available metadata and the tweet content.
Use the information below to infer the most granular geographic scale location if possible. Your output results must be generated after reasoning through  textual information.

Input:
Longitude: {longitude}
Latitude: {latitude}
Tweet Text: {tweet}
Location: {location}

Instruction:
Let's think step by step:
 Step 1: Check if Longitude, or Latitude exist. If so, infer the location and return it. Otherwise, move to Step 2.
 Step 2: Analyze the Tweet Text to find any explicit or implicit mention of a location (\emph{e.g.}, city, county, state, street, neighborhood, national park). If found, use it as the final location and return the most granular geographic information available. if not, move to step 3.
 Step 3: If neither one found in Step 1 and Step 2, use location fields from the input to infer location.

Output Instructions:
If a U.S. location can be confidently identified, return it in plain text (\emph{e.g.}, "San Francisco, CA"). Avoid including non-physical locations (\emph{e.g.}, Earth, Galaxy).
If the tweet is not within the U.S. or the indeterminable, return "No".
If the tweet contains multiple locations, return the most granular geographic information.
If the final location information is abbreviated (\emph{e.g.}, "LV" for Las Vegas), return the full location name.
If the final location information contains distance information (\emph{e.g.}, "10 miles from LA"), or other vague details (\emph{e.g.}, "38th floor of hotel"), return "No".
Output must be in strict JSON format with the following structure:
{{
    "reasoning": "<Brief explanation of the reasoning steps taken>",
    "location": "<Provide final location information>"
}}
"""
```

```
EVENT_PROMPT = """
Task:
You are an earthquake engineer. Your task is to determine whether an input tweet is related to <2019 ridgecrest> earthquake in any meaningful way, such as their impact, damage, or aftermath.
Please read the tweet carefully and decide if it is about an earthquake.

Input:
Tweet Text: {tweet}

Instruction:
Examples of tweets related to earthquakes:
-Last night she said that I needed to not stack all these shoe boxes up so high because an earthquake will happen and they will all fall on me! I am more worried about damaging the boxes and not being able to pass as Deadstock TBH than falling on me.
-My outdoor pillows fell and my pancake is now burnt. This is the extent of the damage of the earthquake in Vegas for me.
- Devi Bhujel, making tea in her kitchen in her village in Nepal. #water here is very hard. I take one jerrycan in a basket, it's about 10 liters maybe. The usual walking road is destroyed by the earthquake and construction. WaterAid/Sibtain Haider #July4th
Examples of tweets not related to earthquakes:
-we were watching CNN when they broke the news about the earthquake and the weather dude was like it "originated here" and circled the area near Tehachapi  which is where I'm going today and staying for the next couple days.
-I knew those Trump tanks would cause damage.  #earthquake

Restrictions: Exclude input tweet information if it solely contains magnitudes <\emph{e.g.},6.4 magnitudes>, distances from the epicenter <\emph{e.g.}, 10km> or other standard seismological data.
Your output results must be generated after reasoning through extual and/or visual information.

Output:
Respond only with Yes if the tweet is related to an earthquake.
Respond only with No if the tweet is not related to an earthquake.
Output must be in strict JSON format with the following structure:
{{
    "reasoning": "<Brief explanation of the reasoning steps taken>",
    "is_event_related": "<Yes | No>"
}}
"""
```

```
IMAGE_ONLY_PROMPT:
f"""
Task:
You are the earthquake damage assessment experts. Your task is to identify the damage level align with Modified Mercalli Intensity(MMI) levels from a given tweet.
Your output must be generated based on evidence from the given tweet content.

```

Table 6: Comparison of model responses to a tweet example.

Tweet Example	Model Responses	LLaVA	Qwen	Gemini
	Location Response	El Monte, CA	El Monte, CA	El Monte, CA
	Event Response	Yes	Yes	Yes
	Human-impacted	No	No	Yes
	Damage Type	Exterior	Interior	Interior
	MMI levels	3	4	4
	Voted candidate	Text and Image	Text and Image	Text and Image
	Confidence	0.8	0.9	0.9
Reasoning for *LLaVA*	The tweet describes a 6.4 earthquake that damaged a friend's window in El Monte. The image shows a window with visible damage, but there is no indication of human impact or damage to the interior of the building. The damage is classified as exterior damage, and the damage level is classified as MMI 3, which corresponds to a light earthquake felt indoors, especially on upper floors; no significant structural damage			
Reasoning for *Qwen*	The image shows a broken window, indicating interior damage. The text does not mention any human emotional or physical reaction to the earthquake, so the human impact is considered 0. The broken window suggests moderate damage, which aligns with the MMI scale category of 'Moderate' (4)			
Reasoning for *Gemini*	Both the text and image contribute to the assessment. The text indicates that a window was damaged, implying human impact as the earthquake was felt. The image confirms that the damage is to the interior of the building, specifically a broken window. This level of damage aligns with MMI level 4, which includes minor damage to buildings such as broken windows or minor cracks. The combined evidence gives a more reliable assessment			

Table 7: Validation Results between MMI Ground Truth (Exterior) for Fukushima and Ridgecrest Events

Event	Metric	Gemini	Qwen	LLaVA
Fukushima	Pearson R	0.54	0.05	0.12
	Number of Tweets	1207	86	24
Ridgecrest	Pearson R	0.38	0.09	0.26
	Number of Tweets	1185	264	53

```
Input:

Image Description:
Please analyze the image to assess the severity of the
      earthquake's damage.

Instructions:

1. Human Impact Evaluation:
   Look for language or visual evidence suggesting
        that people experienced or emotionally reacted
         to the earthquake. Indicators may include
        expressions or signs of: fear(\emph{e.g.}, "
        people were terrified", "panic in the streets
        "), shock or confusion(\emph{e.g.}, "people
        didn't know what to do"), physical presence or
         impact (\emph{e.g.}, "people ran outside", "
        rescue teams helping trapped residents"),
        sensation reporting (\emph{e.g.}, "I felt the
        floor shake", "it was the strongest I've ever
        felt"), etc. Then return:
    1: if there is any mention or evidence of human
         emotional or physical experience of the
         earthquake.
    0: if there is no indication that humans were
         present or affected emotionally/physically.

2. Damage Type Classification:
   Classify the damage type as either:
    Interior: Damage that is clearly observed inside a
          building (e,g, cracked or collapsed interior
          walls, broken windows or glass, displaced or
          fallen indoor furniture, ceiling or floor
         cracks, shaking fixtures (\emph{e.g.}, light
         fixtures, shelves)).
    Exterior: Damage that is clearly observed on the
         outside of buildings or in the surrounding
         environment (\emph{e.g.}, Collapsed buildings
         , shifts in building foundation or roof
         collapse, partial structural failure, cracked
          roads/sidewalks/bridges, fallen trees or
         utility poles, visible debris or rubble
         outside).
```

```
    Both: Evidence of damage is present both inside and outside of structures. The content includes clear indicators of both categories listed above.
    None: The input does not provide enough information to determine whether the damage is interior, exterior, or both.

3. Damage Level Classification (MMI Scale):
   After identifying the damage type (Interior, Exterior, Both, or None) and human impact ("1" or "o"), classify the earthquake damage level align with MMI scale.
   If human impact is 1 from the previous step (human can feel the earthquake), consider both human impact and damage level classification.
   If human impact is 0 from the previous step (human can't feel the earthquake), proceed based solely with damage level classification.

   Damage Level Categories (MMI Scale):
   1 - Not felt: No noticeable damage.
   2 - Weak: Felt by only a few people at rest; no damage to buildings.
   3 - Light: Felt indoors, especially on upper floors; no significant structural damage.
   4 - Moderate: Felt by most people; some damage to buildings, such as minor cracks.
   5 - Strong: Felt by everyone; damage to buildings, minor cracks, but no collapse.
   6 - Very Strong: Damage to buildings, visible structural deformation.
   7 - Severe: Significant damage, some collapses or structural failures.
   8 - Very Severe: Many buildings collapse or are severely damaged.
   9 - Violent: Total destruction in some areas, severe damage.
   10 - Extreme: Complete destruction of all structures in the affected area.


Output:
Output must be in strict JSON format with the following structure:
{{
    "human_impact": <1 or 0>,
    "damage_type": "<Interior | Exterior | Both | None>",
    "damage_level": <1-10>,
    "reasoning": "<Explain how you get the human_impact, damage_type, damage_level based on the input information>",
    "confidence": "<Return how confident (scale 0-1) you are in the final MMI damage level>"
}}
"""
```

```
TEXT_IMAGE_FUSION_PROMPT:
f"""
Task:
You are the earthquake damage accessment experts. Your task is to identify the damage level align with Modified Mercalli Intensity(MMI) levels from a given tweet.
Your output must be generated based on evidence from the given tweet content.

Input:
Text Description:
{tweet}

Image Description:
Please analyze the image to assess the severity of the earthquake's damage based on MMI Scale.

Instructions:

1. Human Impact Evaluation:
   Look for language or visual evidence suggesting that people experienced or emotionally reacted to the earthquake. Indicators may include expressions or signs of: fear(\emph{e.g.}, "people were terrified", "panic in the streets"), shock or confusion(\emph{e.g.}, "people didn't know what to do"), physical presence or impact (\emph{e.g.}, "people ran outside", "rescue teams helping trapped residents"), sensation reporting (\emph{e.g.}, "I felt the floor shake", "it was the strongest I've ever felt"), etc. Then return:
    1: if there is any mention or evidence of human emotional or physical experience of the earthquake.
    0: if there is no indication that humans were present or affected emotionally/physically.

2. Damage Type Classification:
   Classify the damage type as either:
    - Interior: Damage that is clearly observed inside a building (e,g, cracked or collapsed interior walls, broken windows or glass, displaced or fallen indoor furniture, ceiling or floor cracks, shaking fixtures (\emph{e.g.}, light fixtures, shelves)).
    - Exterior: Damage that is clearly observed on the outside of buildings or in the surrounding environment (\emph{e.g.}, Collapsed buildings, shifts in building foundation or roof collapse, partial structural failure, cracked roads/sidewalks/bridges, fallen trees or utility poles, visible debris or rubble outside).
    - Both: Evidence of damage is present both inside and outside of structures. The content includes clear indicators of both categories listed above.
    - None: The input does not provide enough information to determine whether the damage is interior, exterior, or both.

3. Damage Level Classification (MMI Scale):
   After identifying the damage type (Interior, Exterior, Both, or None) and human impact ("1" or "o"), classify the earthquake damage level align with MMI scale.
   If human impact is 1 from the previous step (human can feel the earthquake), consider both human impact and damage level classification.
   If human impact is 0 from the previous step (human can't feel the earthquake), proceed based solely with damage level classification.

   Damage Level Categories (MMI Scale):
   1 - Not felt: No noticeable damage.
   2 - Weak: Felt by only a few people at rest; no damage to buildings.
   3 - Light: Felt indoors, especially on upper floors; no significant structural damage.
   4 - Moderate: Felt by most people; some damage to buildings, such as minor cracks.
   5 - Strong: Felt by everyone; damage to buildings, minor cracks, but no collapse.
   6 - Very Strong: Damage to buildings, visible structural deformation.
   7 - Severe: Significant damage, some collapses or structural failures.
   8 - Very Severe: Many buildings collapse or are severely damaged.
   9 - Violent: Total destruction in some areas, severe damage.
   10 - Extreme: Complete destruction of all structures in the affected area.

```

```
Output:
Output must be in strict JSON format with the following structure:
{{
    "human_impact": <1 or 0>,
    "damage_type": "<Interior | Exterior | Both | None>",
    "damage_level": <1-10>,
    "reasoning": "<Explain how you get the human_impact, damage_type, damage_level based on the input information>",
    "confidence": "<Return how confident (scale 0-1) you are in the final MMI damage level>"
}}
"""
```

```
VOTE_PROMPT =
f"""
Task:
You are an expert in AI systems and earthquake damage assessment. Your task is to identify the damage level described in a given tweet, aligned with the Modified Mercalli Intensity (MMI) scale.
Your output must be based on evidence from the tweet content. You should evaluate results from three candidate models, and select the one that provides the best performance. Include reasoning to justify your choice.

Input:
Text Description:
{tweet}

Image Description:
Please analyze the image to assess the severity of the earthquake's damage.

Instructions:
Your goals including:
        1.      Carefully compare the three outputs for accuracy, completeness, and consistency with the input tweet and image.
        2.      Vote for the most reliable and accurate prediction, considering all available evidence.
        3.      Briefly explain your reasoning.
    4.  Your output results must be generated after reasoning through  extual and/or visual information.

1. Human Impact Evaluation:
   Look for language or visual evidence suggesting that people experienced or emotionally reacted to the earthquake. Indicators may include expressions or signs of: fear(e.g., "people were terrified", "panic in the streets"), shock or confusion(e.g., "people didn't know what to do"), physical presence or impact (e.g., "people ran outside", "rescue teams helping trapped residents"), sensation reporting (e.g., "I felt the floor shake", "it was the strongest I've ever felt"), etc. Then return:
    1: if there is any mention or evidence of human emotional or physical experience of the earthquake.
    0: if there is no indication that humans were present or affected emotionally/physically.

2. Damage Type Classification:
   Classify the damage type as either:
    Interior: Damage that is clearly observed inside a building (e,g, cracked or collapsed interior walls, broken windows or glass, displaced or fallen indoor furniture, ceiling or floor cracks, shaking fixtures (e.g., light fixtures, shelves)).
    Exterior: Damage that is clearly observed on the outside of buildings or in the surrounding environment (e.g., Collapsed buildings, shifts in building foundation or roof collapse, partial structural failure, cracked roads/sidewalks/bridges, fallen trees or utility poles, visible debris or rubble outside).
    Both: Evidence of damage is present both inside and outside of structures. The content includes clear indicators of both categories listed above.
    None: The input does not provide enough information to determine whether the damage is interior, exterior, or both.

3. Damage Level Classification (MMI Scale):
   After identifying the damage type (Interior, Exterior, Both, or None) and human impact ("1" or "o"), classify the earthquake damage level align with MMI scale.
   If human impact is 1 from the previous step (human can feel the earthquake), consider both human impact and damage level classification.
   If human impact is 0 from the previous step (human can't feel the earthquake), proceed based solely with damage level classification.

   Damage Level Categories (MMI Scale):
   1 - Not felt: No noticeable damage.
   2 - Weak: Felt by only a few people at rest; no damage to buildings.
   3 - Light: Felt indoors, especially on upper floors; no significant structural damage.
   4 - Moderate: Felt by most people; some damage to buildings, such as minor cracks.
   5 - Strong: Felt by everyone; damage to buildings, minor cracks, but no collapse.
   6 - Very Strong: Damage to buildings, visible structural deformation.
   7 - Severe: Significant damage, some collapses or structural failures.
   8 - Very Severe: Many buildings collapse or are severely damaged.
   9 - Violent: Total destruction in some areas, severe damage.
   10 - Extreme: Complete destruction of all structures in the affected area.

Candidate Responses:
        1.      Text Only: {text_response}
        2.      Image Only: {image_response}
        3.      Text + Image: {text_image_response}

Output:
Output must be in strict JSON format with the following structure:
{{
    "voted": "<Text Only | Image Only | Text + Image>",
    "human_impact": <1 or 0>,
    "damage_type": "<Interior | Exterior | Both | None>",
    "damage\_level": 1--10
    "reasoning": "<Explain the reasoning behind your selected model>",
    "confidence": "<Return how confident (scale 0-1) you are in the final MMI damage level>"
}}
"""
```

A.7 Prompt rewritten versions

```
PROMPT_V2: """
Task:
```

```
You are an earthquake damage assessment expert. For
     each tweet, follow these three steps to classify
     the damage:

Step 1: Describe any human emotional or physical
     reactions mentioned in the tweet or shown in the
     image.
Step 2: Describe any structural or environmental
     damage observed in the image.
Step 3: Based on both observations, classify the
     earthquake's Modified Mercalli Intensity (MMI)
     level.

Input:
Text Description:
{tweet}

Image Description:
Please analyze the image to assess visible earthquake
     damage.

Output:
Respond in JSON format:
{{
    "human_impact": <1 or 0>,
    "damage_type": "<Interior | Exterior | Both | None
        >",
    "damage\_level": 1--10
    "reasoning": "<Step-by-step breakdown>",
    "confidence": "<0.0-1.0>"
}}
"""
```

```
PROMPT_V3: f"""
Task:
Your primary role is to assess earthquake damage using
      visual cues in the image provided. Use the tweet
      text only if needed to resolve ambiguities.

Input:
Image Description:
Analyze for any visible earthquake damage-structural
     collapse, debris, road cracks, etc.

Text Description:
{tweet}

Output:
Return the damage classification in JSON:
{{
    "human_impact": <1 or 0>,
    "damage_type": "<Interior | Exterior | Both | None
        >",
    "damage\_level": 1--10,
    "reasoning": "<Visual evidence used to support the
         output>",
    "confidence": "<0.0-1.0>"
}}
"""
```

```
PROMPT_4: f"""
Analyze the tweet and associated image to determine
     the earthquake damage level according to the MMI
     scale.

Input:
Text: {tweet}
Image: [image provided]

Output:
Strictly return JSON:
{{
    "human_impact": <1 or 0>,
    "damage_type": "<Interior | Exterior | Both | None
        >",
    "damage\_level": 1--10,
    "reasoning": "<Why each field was chosen>",
    "confidence": "<0.0-1.0>"
}}
"""
```

```
PROMPT_5: f"""
Task:
Please answer the following questions based on the
     tweet and image:

1. Did people seem to experience or react to the
     earthquake?
2. Where did the damage occur-inside, outside, both,
     or unclear?
3. What is the MMI level based on the human and
     structural impact?

Tweet: {tweet}
Image: [Analyze the image]

Output:
Output must be in strict JSON format with the
     following structure:
{{
    "human_impact": <1 or 0>,
    "damage_type": "<Interior | Exterior | Both | None
        >",
    "damage\_level": 1--10,
    "reasoning": "<Explain how you get the
         human_impact, damage_type, damage_level based
          on the input information>",
    "confidence": "<Return how confident (scale 0-1)
         you are in the final MMI damage level>"
}}
"""
```

```
PROMPT_6: f"""
Task:
Review the following examples and then analyze the new
      tweet and image.

Example 1:
Tweet: "People ran outside screaming after their house
      walls cracked."
Image: [shows rubble and collapsed roof]
Output:
{{
    "human_impact": 1,
    "damage_type": "Both",
    "damage\_level": 7,
    "reasoning": "Clear human fear and both interior (
         walls) and exterior (roof) damage.",
    "confidence": "0.85"
}}

Now classify:
Tweet: {tweet}
Image: [Analyze the image]

Output:
Output must be in strict JSON format with the
     following structure:
{{
    "human_impact": <1 or 0>,
    "damage_type": "<Interior | Exterior | Both | None
        >",
    ""damage\_level": 1--10,
    "reasoning": "<Explain how you get the
         human_impact, damage_type, damage_level based
          on the input information>",
    "confidence": "<Return how confident (scale 0-1)
         you are in the final MMI damage level>"
}}
"""
```

```
PROMPT_7: f"""
Task:
```

```
Classify the tweet and image below according to the
     following strict schema.

Input:
Tweet Content: {tweet}
Image Content: [image provided]

Output Format:
All fields must match format:
- human_impact: (0 or 1)
- damage_type: "Interior", "Exterior", "Both", or "
     None"
- damage\_level: Integer from 1 to 10
- reasoning: Text, <400 characters
- confidence: Float between 0 and 1

Output:
Output must be in strict JSON format with the
     following structure:
{{
    "human_impact": <1 or 0>,
    "damage_type": "<Interior | Exterior | Both | None
         >",
    "damage\_level": 1--10,
    "reasoning": "<Explain how you get the
         human_impact, damage_type, damage_level based
          on the input information>",
    "confidence": "<Return how confident (scale 0-1)
         you are in the final MMI damage level>"
}}
"""
```

References

[1] Daniyar Amangeldi, Aida Usmanova, and Pakizar Shamoi. 2024. Understanding environmental posts: Sentiment and emotion analysis of social media data. *IEEE Access* (2024).

[2] Ron Artstein. 2017. Inter-annotator agreement. *Handbook of linguistic annotation* (2017), 297–313.

[3] C Aybar. 2022. DCloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2. *Sci Data* (2022). doi:10.1038/s41597-022-01878-2

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

[5] Matthias Butenuth, Daniel Frey, Allan Aasbjerg Nielsen, and Henning Skriver. 2011. Infrastructure assessment for disaster management using multi-sensor and multi-temporal remote sensing imagery. *International Journal of Remote Sensing* 32, 23 (2011), 8575–8594.

[6] Anwoy Chatterjee, HSVNS Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A Prompt Sensitivity Index For Large Language Models. *arXiv preprint arXiv:2410.02185* (2024).

[7] Copernicus. 2022. Sentinel-2.

[8] Harald Cramér. 1999. *Mathematical methods of statistics*. Vol. 9. Princeton university press.

[9] Irene de Zarzà, Joachim de Curtò, Gemma Roig, and Carlos T Calafate. 2023. LLM multimodal traffic accident forecasting. *Sensors* 23, 22 (2023), 9225.

[10] Ashwin Devaraj, Dhiraj Murthy, and Aman Dontula. 2020. Machine-learning methods for identifying social media-based requests for urgent help during hurricanes. *International Journal of Disaster Risk Reduction* 51 (2020), 101757.

[11] Derek Doran, Swapna Gokhale, and Aldo Dagnino. 2014. Accurate local estimation of geo-coordinates for social media posts. *arXiv preprint arXiv:1410.4616* (2014).

[12] Paula García-Tapia-Mateo, Andrés Bueno-Crespo, Irene Garrigos, Jose-Norberto Mazón, José Cecilia, and Juan Morales-García. 2025. Explainable deep learning for early detection of natural disasters through social media text classification. *Available at SSRN 5113748* (2025).

[13] Pranav Guruprasad, Harshvardhan Sikka, Jaewoo Song, Yangyue Wang, and Paul Pu Liang. 2024. Benchmarking Vision, Language, & Action Models on Robotic Learning Tasks. *arXiv preprint arXiv:2411.05821* (2024).

[14] RO Hamburger, C Rojahn, J Heintz, and MG Mahoney. 2012. FEMA P58: Next-generation building seismic performance assessment methodology. In *15th world conference on earthquake engineering*, Vol. 10.

[15] Amanda Lee Hughes and Holden Clark. 2025. Seeing the Storm: Leveraging Multimodal LLMs for Disaster Social Media Video Filtering. In *Proceedings of the International ISCRAM Conference*.

[16] J Jelének and Kopačková-Strnadová. 2021. Synergic use of Sentinel-1 and Sentinel-2 data for automatic detection of earthquake-triggered landscape changes: A case study of the 2016 Kaikoura earthquake (Mw 7.8), New Zealand. *RemoteSensing* (2021). doi:10.1016/j.rse.2021.112634

[17] Hongxiang Jiang, Jihao Yin, Qixiong Wang, Jiaqi Feng, and Guo Chen. 2025. EagleVision: Object-level Attribute Multimodal LLM for Remote Sensing. *arXiv preprint arXiv:2503.23330* (2025).

[18] Darioush Kevian, Usman Syed, Xingang Guo, Aaron Havens, Geir Dullerud, Peter Seiler, Lianhui Qin, and Bin Hu. 2024. Capabilities of large language models in control engineering: A benchmark study on gpt-4, claude 3 opus, and gemini 1.0 ultra. *arXiv preprint arXiv:2404.03647* (2024).

[19] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. 2016. Rapid assessment of disaster damage using social media activity. *Science advances* 2, 3 (2016), e1500779.

[20] Intel Labs. 2024. LLaVA LLaMA-3 8B. https://huggingface.co/Intel/llava-llama-3-8b Accessed: 2024-05-18.

[21] Zhenyu Lei, Yushun Dong, Weiyu Li, Rong Ding, Qi Wang, and Jundong Li. 2025. Harnessing Large Language Models for Disaster Management: A Survey. *arXiv preprint arXiv:2501.06932* (2025).

[22] Lingyao Li, Michelle Bensi, and Gregory Baecher. 2023. Exploring the potential of social media crowdsourcing for post-earthquake damage assessment. *International Journal of Disaster Risk Reduction* 98 (2023), 104062.

[23] Lingyao Li, Zihui Ma, and Tao Cao. 2021. Data-driven investigations of using social media to aid evacuations amid Western United States wildfire season. *Fire Safety Journal* 126 (2021), 103480.

[24] Rong Li, Lei Zhao, ZhiQiang Xie, Chunhou Ji, Jiamin Mo, Zhibing Yang, and Yuyun Feng. 2025. Mining and analyzing the evolution of public opinion in extreme disaster events from social media: A case study of the 2022 yingde flood in china. *Natural Hazards Review* 26, 1 (2025), 05024015.

[25] Tianyu Liu, Yijia Xiao, Xiao Luo, Hua Xu, W Jim Zheng, and Hongyu Zhao. 2024. Geneverse: A collection of Open-source Multimodal Large Language Models for Genomic and Proteomic Research. *arXiv preprint arXiv:2406.15534* (2024).

[26] Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Massimo Hong, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2024. Biomedgpt: An open multimodal large language model for biomedicine. *IEEE Journal of Biomedical and Health Informatics* (2024).

[27] Zihui Ma, Lingyao Li, Libby Hemphill, Gregory B Baecher, and Yubai Yuan. 2024. Investigating disaster response for resilient communities through social media data and the Susceptible-Infected-Recovered (SIR) model: A case study of 2020 Western US wildfire season. *Sustainable Cities and Society* 106 (2024), 105362.

[28] Zihui Ma, Lingyao Li, Yujie Mao, Yu Wang, Olivia Grace Patsy, Michelle T Bensi, Libby Hemphill, and Gregory B Baecher. 2024. Surveying the Use of Social Media Data and Natural Language Processing Techniques to Investigate Natural Disasters. *Natural Hazards Review* 25, 4 (2024), 03124003.

[29] Petros Maragos, Patrick Gros, Athanassios Katsamanis, and George Papandreou. 2008. Cross-modal integration for performance improving in multimedia: A review. *Multimodal Processing and Interaction: Audio, Video, Text* (2008), 1–46.

[30] Ayaz Mehmood, Muhammad Tayyab Zamir, Muhammad Asif Ayub, Nasir Ahmad, and Kashif Ahmad. 2024. A named entity recognition and topic modeling-based solution for locating and better assessment of natural disasters in social media. *arXiv preprint arXiv:2405.00903* (2024).

[31] James L Merlo, Aaron R Duley, and Peter A Hancock. 2010. Cross-modal congruency benefits for combined tactile and visual signaling. *The American journal of psychology* 123, 4 (2010), 413–424.

[32] Volodymyr V Mihunov, Navid H Jafari, Kejin Wang, Nina SN Lam, and Dylan Govender. 2022. Disaster impacts surveillance from social media with topic modeling and feature extraction: Case of Hurricane Harvey. *International Journal of Disaster Risk Science* 13, 5 (2022), 729–742.

[33] Siddharth Mishra-Sharma, Yiding Song, and Jesse Thaler. 2024. Paperclip: Associating astronomical observations and natural language with multi-modal models. *arXiv preprint arXiv:2403.08851* (2024).

[34] Yuki Miura, Huda Qureshi, Chanyang Ryoo, Philip C Dinenis, Jiao Li, Kyle T Mandli, George Deodatis, Daniel Bienstock, Heather Lazrus, and Rebecca Morss. 2021. A methodological framework for determining an optimal coastal protection strategy against storm surges and sea level rise. *Natural Hazards* 107 (2021), 1821–1843.

[35] Phyo Yi Win Myint, Siaw Ling Lo, and Yuhao Zhang. 2024. Unveiling the dynamics of crisis events: Sentiment and emotion analysis via multi-task learning with attention mechanism and subject-based intent prediction. *Information Processing & Management* 61, 4 (2024), 103695.

[36] New York University. 2024. NYU High Performance Computing. https://hpc.nyu.edu Accessed: 2024-05-18.

[37] O Odubola, TS Adeyemi, OO Olajuwon, et al. 2025. AI in Social Good: LLM powered Interventions in Crisis Management and Disaster Response. *J Artif Intell Mach Learn & Data Sci 2025* 3, 1 (2025), 2353–2360.

[38] Niall O'Mahony, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan, and Joseph Walsh. 2020. Deep learning vs. traditional computer vision. In *Advances in computer*

vision: proceedings of the 2019 computer vision conference (CVC), volume 1 1. Springer, 128–144.

[39] Jiangao Peng, Zhongyuan Han, Huan Zhang, Jingyan Ye, Chang Liu, Biao Liu, Mingcan Guo, Haoyang Chen, Zijie Lin, and Yujiao Tang. 2024. A multilingual text detoxification method based on few-shot learning and co-star framework. *Working Notes of CLEF* (2024).

[40] Pew Research Center. 2022. 10 facts about Americans and Twitter. https://www.pewresearch.org/fact-tank/2022/05/05/10-facts-about-americans-and-twitter/ Accessed: 2025-05-18.

[41] Qwen Team, Alibaba Cloud. 2024. Qwen2.5: The Latest Vision-Language Model from Alibaba. https://qwenlm.github.io/blog/qwen2.5/ Accessed: 2024-05-18.

[42] Mariia Rizhko and Joshua S Bloom. 2024. Self-supervised Multimodal Model for Astronomy. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*.

[43] Philip J Schneider and Barbara A Schauer. 2006. HAZUS—Its development and its future. *Natural Hazards Review* 7, 2 (2006), 40–44.

[44] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324* (2023).

[45] Siqing Shan, Feng Zhao, Yigang Wei, and Mengni Liu. 2019. Disaster management 2.0: A real-time disaster damage assessment model based on mobile social media data—A case study of Weibo (Chinese Twitter). *Safety science* 115 (2019), 393–413.

[46] Tanusree Sharma, Yujin Potter, Zachary Kilhoffer, Yun Huang, Dawn Song, and Yang Wang. 2024. From Experts to the Public: Governing Multimodal Language Models in Politically Sensitive Video Analysis. *arXiv preprint arXiv:2410.01817* (2024).

[47] Kristin Stock. 2018. Mining location from social media: A systematic review. *Computers, Environment and Urban Systems* 71 (2018), 209–240.

[48] Bairavel Subbaiah, Kanipriya Murugesan, Prabakeran Saravanan, and Krishnamurthy Marudhamuthu. 2024. An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network. *Artificial Intelligence Review* 57, 2 (2024), 34.

[49] Eric Tate, Cristina Munoz, and Jared Suchan. 2015. Uncertainty and sensitivity analysis of the HAZUS-MH flood model. *Natural Hazards Review* 16, 3 (2015), 04014030.

[50] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[51] Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (LLM) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining* 15, 1 (2025), 1–30.

[52] Matthew M Torok, Mani Golparvar-Fard, and Kevin B Kochersberger. 2014. Image-based automated 3D crack detection for post-disaster building assessment. *Journal of Computing in Civil Engineering* 28, 5 (2014), A4014004.

[53] David Jay Wald, Vincent Quitoriano, Charles Bruce Worden, Margaret Hopper, and James W Dewey. 2011. USGS "Did You Feel It?" internet-based macroseismic intensity maps. *Annals of geophysics* 54, 6 (2011).

[54] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. 2024. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems* 37 (2024), 75392–75421.

[55] Ming Wang, Yuanzhong Liu, Xiaoyu Liang, Songlian Li, Yijie Huang, Xiaoming Zhang, Sijia Shen, Chaofeng Guan, Daling Wang, Shi Feng, et al. 2024. Langgpt: Rethinking structured reusable prompt design framework for llms from the programming language. *arXiv preprint arXiv:2402.16929* (2024).

[56] Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2025. Cross-modal retrieval: a systematic review of methods and future directions. *Proc. IEEE* (2025).

[57] Zhenyu Wang, Zikang Wang, Jiyue Jiang, Pengan Chen, Xiangyu Shi, and Yu Li. 2025. Large Language Models in Bioinformatics: A Survey. *arXiv preprint arXiv:2503.04490* (2025).

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[59] Yang Wu, Shilong Wang, Hao Yang, Tian Zheng, Hongbo Zhang, Yanyan Zhao, and Bing Qin. 2023. An early evaluation of gpt-4v (ision). *arXiv preprint arXiv:2310.16534* (2023).

[60] Shaorong Xie, Chunning Hou, Hang Yu, Zhenyu Zhang, Xiangfeng Luo, and Nengjun Zhu. 2022. Multi-label disaster text classification via supervised contrastive learning for social media data. *Computers and Electrical Engineering* 104 (2022), 108401.

[61] Kai Yin, Chengkai Liu, Ali Mostafavi, and Xia Hu. 2024. Crisissense-llm: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics. *arXiv preprint arXiv:2406.15477* (2024).

[62] Chen Yu and Zhiguo Wang. 2024. Multimodal social sensing for the spatio-temporal evolution and assessment of nature disasters. *Sensors* 24, 18 (2024), 5889.

[63] Rui Zhong, Yuefeng Xu, Chao Zhang, and Jun Yu. 2024. Leveraging large language model to generate a novel metaheuristic algorithm with CRISPE framework. *Cluster Computing* 27, 10 (2024), 13835–13869.

[64] Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. *arXiv preprint arXiv:2410.12405* (2024).