LLMs as World Models: Data-Driven and Human-Centered Pre-Event Simulation for Disaster Impact Assessment

Lingyao Li[1*†], **Dawei Li**[2*†], **Zhenhui Ou**[2], **Xiaoran Xu**[1],
Jingxiao Liu[3], **Zihui Ma**[4], **Runlong Yu**[5], **Min Deng**[6]
[1] University of South Florida [2] Arizona State University
[3] Massachusetts Institute of Technology [4] New York University
[5] University of Alabama [6] Texas Tech University
`lingyao@usf.edu`, `daweili5@asu.edu`, `zhenhuio@asu.edu`, `xiaoranxu@usf.edu`,
`jingxiao@mit.edu`, `zihuima@nyu.edu`, `ryu5@ua.edu`, `mindeng@ttu.edu`

Abstract

Efficient simulation is essential for enhancing proactive preparedness for sudden-onset disasters such as earthquakes. Recent advancements in large language models (LLMs) as world models show promise in simulating complex scenarios. This study examines multiple LLMs to proactively estimate perceived earthquake impacts. Leveraging multimodal datasets including geospatial, socioeconomic, building, and street-level imagery data, our framework generates Modified Mercalli Intensity (MMI) predictions at zip code and county scales. Evaluations on the 2014 Napa and 2019 Ridgecrest earthquakes using USGS "Did You Feel It? (DYFI)" reports demonstrate significant alignment, as evidenced by a high correlation of 0.88 and a low RMSE of 0.77 as compared to real reports at the zip code level. Techniques such as retrieval-augmented generation (RAG) and in-context learning (ICL) can improve simulation performance, while visual inputs notably enhance accuracy compared to structured numerical data alone. These findings show the promise of LLMs in simulating disaster impacts that can help strengthen pre-event planning. **Data Access**: `https://doi.org/10.5281/zenodo.17148713`; **Code Access**: `https://github.com/Lingyao1219/llm-disaster-simulation`.

1 Introduction

Natural disasters often disrupt infrastructure, causing significant human and economic losses (Jones et al., 2022). Efficient impact assessment is critical for emergency response and evaluating community resilience (Ma et al., 2024). However, most existing methods are designed for post-event assessment, including expert inspections, ground sensors, and remote sensing (Li et al., 2021; Kucharczyk and Hugenholtz, 2021; Sarkar et al., 2023). While effective for characterizing observed damage, these approaches are reactive by nature and offer limited utility in pre-event planning, especially for sudden-onset events like earthquakes, where early awareness is crucial (Li et al., 2023). Traditional methods for pre-event simulation like scenario-based planning are useful (Ma et al., 2024; Deierlein et al., 2020), but they need extensive domain expertise for region-specific models and often lack empirical validation by addressing human-centered factors.

Advances in large language models (LLMs) have shown promise in contextual simulation and complex reasoning across various domains (Li et al., 2024a; Wang et al., 2024b; Li et al., 2025b,a). Beyond text-based reasoning, recent multimodal LLMs have also demonstrated strong visual reasoning capabilities, allowing them to interpret and reason about physical environments (Xiang et al., 2023; Li et al., 2025c). These abilities have led researchers to increasingly view LLMs as potential world models—systems capable of learning to simulate and predict real-world scenarios (Wong et al., 2023; Hao et al., 2023). Through training on large-scale datasets that encode spatial, temporal, and causal relationships, LLMs have shown potential in learning representations of how the world works. For example, current research has demonstrated their ability to understand environment status (Hao et al., 2023), plan household activities (Xiang et al., 2023), and predict time-series events (Lee et al., 2025).

In disaster management, while LLMs are not yet widely applied as "world models," researchers have explored their utility in tasks such as damage detection from satellite imagery (Zhang and Wang, 2024) or social media (Wang et al., 2024a), and emergency identification (Otal et al., 2024). However, key research gaps still remain. ❶ First, existing studies mainly use LLMs to analyze available textual or visual data for post-event assessment, not to simulate pre-event situations. ❷ Second, while LLMs have well-demonstrated reasoning abilities,

---

[*] Equal contribution as first authors.
[†] Corresponding authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 3078–3096
November 4-9, 2025 ©2025 Association for Computational Linguistics

effective disaster assessment requires integrating domain-specific knowledge and data fusion so that these models can accurately reason about potential disaster scenarios. To address these limitations, this study poses the fundamental question in the context of sudden-onset disasters: **Can LLMs simulate how humans perceive seismic risks before an event occurs?**

To answer this, we develop an LLM-based framework to simulate how humans perceive seismic risks, as illustrated in Figure 1. By integrating rich pre-event contextual information, the LLMs are tasked with "reasoning" the likely severity of damage across spatial scales. Importantly, our study moves beyond theoretical simulations, grounding model evaluations in real-world events by testing on two actual earthquakes and comparing outputs against authoritative USGS reports. Our simulation demonstrates strong alignment with real reports at the zip code level, highlighting the potential of leveraging LLMs to improve pre-event planning. Our key contributions include:

- **Benchmark the performance of LLM-as-world-model reasoning**. We establish a dataset that merges multimodal data resources, and propose paired RMSE and rank-correlation metrics to estimate performance. This is among the early open resources for evaluating how well LLMs reason about disaster scenarios.
- **A multimodal LLM-based framework for pre-event simulation**. We introduce a framework that leverages LLMs as virtual sensors by inputting multimodal features and outputting the risk estimates. By combining structured numerical data with vision-language prompts, the pipeline moves LLMs from post-event assessment to pre-event simulation.
- **Comprehensive cross-model evaluation and reasoning analysis**. We benchmark nine open- and closed-source LLMs across multiple prompting strategies (vanilla, ICL, and RAG). Our ablation studies provide a baseline and actionable insights to guide future improvements in LLM reasoning for real-world disaster applications.

2 Related Work

2.1 Seismic Hazard Simulation

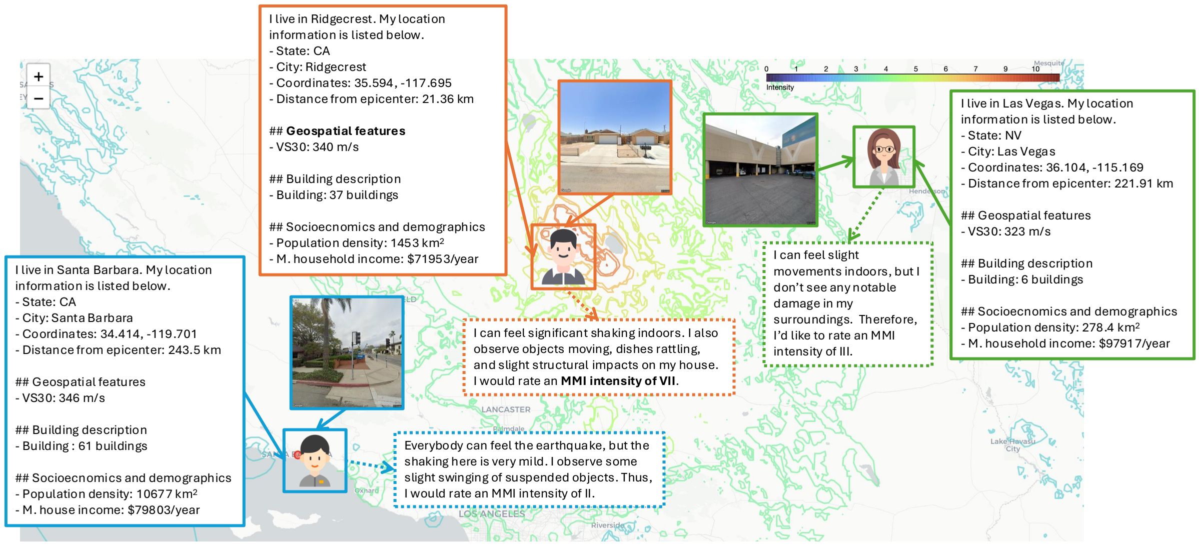
Researchers have employed empirical, physics-based, and data-driven methods for simulating seismic disasters. Traditional approaches characterize earthquakes, such as the moment magnitude M_w (Moschetti et al., 2024), and then utilize empirical ground-motion prediction equations (GMPEs) (Moschetti et al., 2024; Iervolino, 2023) or physics-based simulations (Deierlein et al., 2020) to estimate site-specific shaking. While these simulations can capture complex local effects and rupture dynamics, they demand extensive data and computational resources. Recent data-driven methods have used machine learning to learn damage patterns from historical events, simulations, and remote sensing data (Cardellicchio et al., 2023; Yu et al., 2020)). These AI-driven techniques offer advantages in scalability and flexible feature integration but are dependent on high-quality labeled data, may struggle with generalization, and often present challenges in interpretability.

A major gap across these models, whether physics-based or data-driven, is the limited integration of human-perceived shaking into predictive frameworks. The USGS has developed the "Did You Feel It" (DYFI) system (Atkinson and Wald, 2007), a crowdsourced platform where individuals report the shaking intensity they experience after an earthquake. These reports are aggregated and converted into Modified Mercalli Intensity (MMI) (U.S. Geological Survey, 1989), which provide a human-centric, ground-truth reference for how seismic shaking is felt. While DYFI has been widely used to validate models or interpolate intensities after earthquakes, simulating human-centric perceived risks for pre-event could be important for developing predictive frameworks that anticipate not only physical shaking but also tangible impacts on communities.

2.2 LLM as World Models

Recent advance in LLMs such as GPT-4 (Achiam et al., 2023) and deepseek-R1 (Guo et al., 2025) has motivated researchers to leverage LLMs to solve complex tasks, such as reasoning (Li et al., 2025d; Yu et al., 2025) and domain-specific exploration (Yang et al., 2024; Zhang et al., 2024). In the context of disaster, researchers have leveraged LLMs to process multimodal information for vulnerability evaluation (Martelo et al., 2024), impact assessment (Taghian Dinani et al., 2023; Akinboyewa et al., 2024), information coordination (Yu and Wang, 2024), and recovery planning (White and Liptak, 2025).

Beyond conventional dialogue and analytical



```
I live in Ridgecrest. My location
information is listed below.
- State: CA
- City: Ridgecrest
- Coordinates: 35.594, -117.695
- Distance from epicenter: 21.36 km

## Geospatial features
- VS30: 340 m/s

## Building description
- Building: 37 buildings

## Socioecnomics and demographics
- Population density: 1453 km²
- M. household income: $71953/year
```

I can feel significant shaking indoors. I also observe objects moving, dishes rattling, and slight structural impacts on my house. I would rate an **MMI intensity of VII**.

```
I live in Las Vegas. My location
information is listed below.
- State: NV
- City: Las Vegas
- Coordinates: 36.104, -115.169
- Distance from epicenter: 221.91 km

## Geospatial features
- VS30: 323 m/s

## Building description
- Building: 6 buildings

## Socioecnomics and demographics
- Population density: 278.4 km²
- M. household income: $97917/year
```

I can feel slight movements indoors, but I don't see any notable damage in my surroundings. Therefore, I'd like to rate an MMI intensity of III.

```
I live in Santa Barbara. My location
information is listed below.
- State: CA
- City: Santa Barbara
- Coordinates: 34.414, -119.701
- Distance from epicenter: 243.5 km

## Geospatial features
- VS30: 346 m/s

## Building description
- Building : 61 buildings

## Socioecnomics and demographics
- Population density: 10677 km²
- M. house income: $79803/year
```

Everybody can feel the earthquake, but the shaking here is very mild. I observe some slight swinging of suspended objects. Thus, I would rate an MMI intensity of II.

Figure 1: An illustration of LLM-simulated human-centric sensors.

tasks (Yang et al., 2025), LLMs are being used as world models (Hao et al., 2023; Zhao et al., 2023; Gu et al., 2024) for complex simulation. While definitions of world models vary, their central concept involves leveraging forward reasoning to predict future states and dynamics in real-world scenarios. For example, they can produce large-scale simulations—an LLM-driven city simulation platform, OpenCity, runs tens of thousands of agents to emulate daily urban activities, successfully reproducing emergent patterns like mobility flows and residential segregation (Yan et al., 2024). In disaster scenarios, LLM-based world models have been explored in flooding forecasting (Wang et al., 2025) and evacuation planning (Hostetter et al., 2024).

Building on these advances, we propose leveraging LLMs as simulation tools to estimate how humans might perceive and report seismic risks before an event occurs. Our work addresses two critical gaps: first, the limited pre-event simulation of seismic risk; and second, the underexplored application of LLMs as world models for deriving human-centric insights.

3 Data and Methods

3.1 Framework Design

To simulate disaster impacts before events, we propose a novel framework that treats LLM as synthetic observers. This framework conceptualizes LLMs as "virtual sensors" capable of "perceiving" multimodal inputs and "reasoning" about disaster risks based on the MMI levels (see appendix A) that approximate human perception of shaking. As illustrated in Figure 1, we associate each sampled spatial location with a bundle of features, including seismic distance, site conditions, local building characteristics, and socioeconomic factors. In addition, Google Street View provides a first-person view of the built environment. Collectively, this feature set closely replicates the perceptual input available to disaster responders during an event.

Formally, let each sample i be associated with a fused feature representation. We specifically select the following features that prior research has illustrated as critical predictors of seismic risk. (Frigerio et al., 2016; Kassem et al., 2020; Riedel et al., 2015; Mori et al., 2020).

$$\mathcal{X}_i = \{E_i, G_i, L_i, B_i, S_i, V_i\}$$

- E_i: Earthquake parameters (e.g., magnitude, epicenter distance, depth),
- G_i: Geospatial features (e.g., VS30),
- L_i: Location metadata (e.g., state, city, zip code),
- B_i: Building attributes (e.g., number, type, height, material),
- S_i: Socioeconomic indicators (e.g., population density, income),
- V_i: Street-level view (Google Street image).

The LLM acts as a reasoning function f_θ with parameters θ, generating both a reasoning trace and an MMI rating:

$$\hat{y}_i, e_i = f_\theta(\mathcal{X}_i), \quad \hat{y}_i \in \{\text{I}, \text{II}, \ldots, \text{XII}\}$$

The full pipeline consists of five components, as shown in Figure 2: (1) spatial sampling, (2) data fusion, (3) prompt engineering, (4) experiment design, and (5) result analysis, which we specifically explain in the following sections.

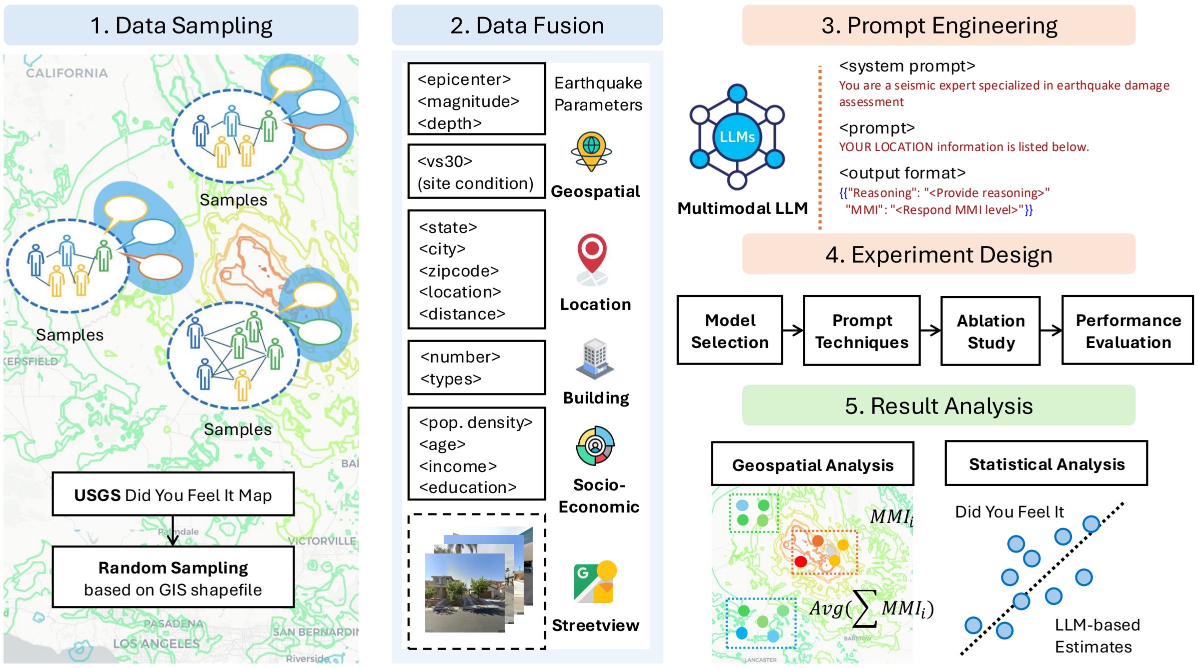


Figure 2: An illustration of the framework design.

3.2 Data Sampling

In Step 1 (Figure 2), we use a polygon-based GIS shapefile to define administrative zones (e.g., zip codes) and apply within-polygon stratified random sampling to ensure spatial representativeness across the study area. Let: $\mathcal{Z} = \{z_1, z_2, \ldots, z_M\}$ represent the set of all zip code polygons. For each polygon z_j, we sample:

$$\mathcal{P}_j = \{p_{j1}, p_{j2}, \ldots, p_{jn_j}\} \sim \text{Uniform}(z_j)$$

ensuring that points are uniformly drawn from within each polygon's spatial boundary. We then sample 50 data points per zip code. This stratified random sampling strategy can help ensure spatial representativeness and mitigate biases associated with population density or urban–rural areas.

3.3 Data Fusion

In Step 2 (Figure 2), for each sampled point p_{ji}, we collect and assemble the feature set $\mathcal{X}_{ji}$ from open-source and public datasets including USGS earthquake data, OpenStreetMap building data, American Community Survey (ACS) data, and Google Maps street views.

Earthquake data, site conditions, and location information. We source earthquake parameters, including moment magnitude, epicentral coordinates, and focal depth, from the USGS ShakeMap product (Wald et al., 2006) as E. To account for local site amplification and deamplification, we incorporate the USGS VS30 dataset (McPhillips et al., 2020) as the geospatial features G, a widely used proxy for near-surface geological effects.

We also incorporate location information, including latitude, longitude, state, county, zip code, and the distance from the epicenter for each sampled point. Additionally, we utilize MMI measurements from the USGS DYFI program as ground-truth labels, represented by y_j for each zip code.

Building description. We use OpenStreetMap (Ramm and Topf, 2010) (OSM) data to extract building features B, including the total number of buildings, type distribution, height range, and average height within a 100-meter radius of each sampled point. Meanwhile, we summarize the prevalence of major construction materials (e.g., concrete, masonry, timber, steel). These features can help characterize the distribution, physical attributes, and seismic-design status of buildings within the sensor's surroundings.

Socioeconomic and demographic factors. We collect socioeconomic and demographic factors S at the Census Block Group (CBG) level from the American Community Survey (ACS) (USC, 2022). We spatially join the coordinates of each randomly selected sample to their corresponding CBG polygon and extract relevant ACS key indicators, including population, population density, urbanization ratio, 65- and over-age proportion, median household income, and higher education attainment rate. These variables imply a quantitative evaluation of population vulnerability in a disaster event.

Google Maps street imagery. We further incorporate Google Street View imagery V to enrich the environmental context at each sampled location. These street-level images capture fine-grained visual cues of the surrounding built environment—such as building, vegetation, curb conditions, and street density—that are difficult to numerically encode but essential for human-like visual reasoning. Using the Google Maps API (Google, 2025), we query the available image for each point in our sampling grid. This visual context allows the LLM to "see" the landscape as if it conducts a field visit.

3.4 Prompt Design

To guide the reasoning process of the LLM and ensure interpretability and consistency, we design a prompt template that mirrors the workflow of a seismic domain expert. The prompt follows a role-based instruction format in which the model is assigned the role of a seismic specialist responsible for evaluating earthquake damage using the MMI scale. The MMI scale provides a human-centric interpretation that consists of twelve levels describing the severity of earthquake shaking. The detailed descriptors of MMI are attached in Appendix A.

The prompt includes six distinct sections corresponding to the multimodal feature set $\mathcal{X}_i$ introduced earlier: earthquake parameters E_i, geospatial features G_i, location metadata L_i, building descriptions B_i, socioeconomic context S_i, and street-level view V_i. The model's response is constrained to a JSON format with two fields: a free-text reasoning explanation and the predicted MMI category (see prompt template in Appendix B). We implement Chain-of-Thought (CoT) to instruct LLM output a detailed reasoning justifying its assessment before final MMI estimate.

3.5 Experiment Design

LLM Selection: We select both open- and closed-source LLMs to conduct the simulation. For open-source LLMs, we use models from Llama-3.2 and Qwen-2.5 family with various sizes, as listed in Table 1.

Prompting Techniques: We incorporate the following prompting techniques in our experiment:

- In-Context Learning (ICL) (Brown et al., 2020; Dong et al., 2024): This helps LLM adapt to tasks by incorporating examples directly within the input prompt. ICL is applied by embedding a detailed MMI reference guide directly within the prompt.
- Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Tan et al., 2024): It combines information retrieval with text generation that allows LLMs to access external data sources at inference. We provide LLM with a set of multimodal features and the reported MMI within the prompt as the retrieved context to ground their MMI estimates.

Evaluation: The LLM generates a MMI prediction, $\hat{y}_{ji} = f_\theta(\mathcal{X}_{ji})$, for each sampled location i within a predefined administrative area j (e.g., zip code, county). These point-level predictions are then aggregated to compute an average predicted MMI for each area j. If area j contains n_j samples, its average predicted MMI $\bar{\hat{y}}_j$ is calculated as:

$$\bar{\hat{y}}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \hat{y}_{ji}$$

Similarly, a corresponding ground-truth MMI value for area j, denoted as $\bar{y}_j$, is derived from USGS DYFI reports. To quantify the model's predictive accuracy using these aggregated area-level values, we compute two metrics. First, the Root Mean Square Error (RMSE) is used:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (\bar{\hat{y}}_j - \bar{y}_j)^2}$$

where N is the total number of administrative areas being evaluated (indexed by j). Second, we calculate Pearson's correlation coefficients r to assess the strength and direction of the association between the LLM-predicted MMI ($\bar{\hat{y}}_j$) and the ground-truth MMI ($\bar{y}_j$). These evaluations are performed at both zip code and county levels.

4 Experimental Results

We select two cases to demonstrate our proposed framework: (1) the 2014 Napa earthquake (magnitude 6.0) and (2) the 2019 Ridgecrest earthquake (magnitude 7.1), both of which occurred in California, U.S. (see details in Appendix C). For each case, we use the USGS DYFI reports as the ground-truth dataset (U.S. Geological Survey, 2014a, 2019a). Additionally, for each case, We compile feature sets for 50 sample points from each of the top 100

Model	Open Source	2014 Napa				2019 Ridgecrest			
		$\mathrm{RMSE_Z}\downarrow$	$\mathrm{Corr_Z}\uparrow$	$\mathrm{RMSE_C}\downarrow$	$\mathrm{Corr_C}\uparrow$	$\mathrm{RMSE_Z}\downarrow$	$\mathrm{Corr_Z}\uparrow$	$\mathrm{RMSE_C}\downarrow$	$\mathrm{Corr_C}\uparrow$
Closed-Source Models									
GPT-4o-2024-08-06	✗	2.43	**0.77**	2.37	**0.88**	1.97	**0.75**	1.91	**0.77**
GPT-4.1-mini	✗	2.56	0.61	2.48	0.67	**0.92**	0.64	**0.77**	0.76
Claude-3.5-haiku	✗	2.11	0.58	2.05	0.70	1.35	0.59	1.38	0.71
Open-Source Models									
Llama-3.2-11B-VI	✓	3.19	0.44	3.05	0.86	3.22	0.33	3.22	0.27
Llama-3.2-90B-VI	✓	2.62	0.57	2.55	0.66	2.06	0.62	2.19	0.59
Qwen2.5-VL-3B	✓	3.63	0.29	3.59	0.15	3.88	0.01	4.08	-0.20
Qwen2.5-VL-7B	✓	1.79	0.43	1.68	0.70	1.53	0.05	1.59	-0.18
Qwen2.5-VL-32B	✓	**1.59**	0.70	**1.56**	0.79	0.99	0.71	0.96	0.80
Qwen2.5-VL-72B	✓	2.17	0.46	2.12	0.44	1.39	0.64	1.28	0.86

Table 1: Main experiment results on two earthquake datasets. Best per-column values are highlighted in blue and bold. Alternating gray rows improve readability.

zip codes with the highest number of responses, resulting in 5,000 samples per event. Due to limitations in Google image availability for the Napa case, only 4,920 samples are retrieved. Comparisons with DYFI data are first conducted at the zip code level, where each of the 100 aggregated values represents the average of 50 simulated samples. A more fine-grained city-level analysis is provided in Appendix G.

Figure 7 presents the spatial distribution of predicted MMI at the zip code level for the Napa and Ridgecrest earthquakes. This visualization highlights variations in simulated seismic impacts across geographical areas and among different LLMs. Based on the best-performing models (lowest RMSE: GPT-4.1-mini for the 2019 Ridgecrest earthquake and Qwen-2.5-32B for the 2014 Napa earthquake), we observe consistent geospatial patterns in both cases. Specifically, the simulations indicate elevated perceived risk near the epicenter (a red star mark in Figure 7), with diminishing simulated impact as distance increases. Moreover, the LLM-based predictions align well with DYFI reports from these two events: it is important to note that the Napa earthquake, despite its lower magnitude, led to more significant impacts.

The following sections are organized below. First, we evaluate the performance of the selected LLMs, comparing their accuracy using quantitative metrics, and examining the influence of model scaling and prompting strategies. Next, we conduct an input feature analysis to evaluate how different data modalities can impact predictive performance. Lastly, the output reasoning analysis explores the internal decision-making processes of the models, which identifies linguistic nuances that illustrate how LLMs interpret the inputs.

4.1 Model Performance

Before the main experiment, we perform a data leakage test using two close-source models—Claude-3.5-haiku and GPT-4.1-mini—to show that our simulation is free of data leakage issues (see Appendix D). The main experiment results are shown in Table 1, from which we draw the following findings. In addition, further comparisons with ShakeMap and traditional machine learning methods are provided in Appendix H.

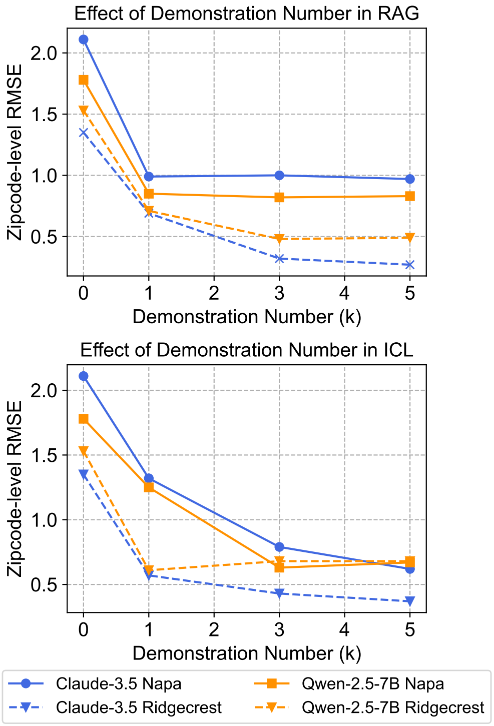


Figure 3: Demonstration analysis on RAG and ICL.

LLMs deliver promising simulation results. As shown in Table 1, all of the best correlation scores across datasets and area levels exceed 0.7, indicating a strong alignment between the predicted outcomes and the ground truth impact labels. These high correlation values suggest that LLMs hold significant potential for effectively simulating human-perceived risks in disaster scenarios.

Closed-source LLMs generally outperform open-source models. Table 1 shows a clear trend that closed-source LLMs consistently outperform open-source counterparts, achieving the best results in 6 out of 8 cases. This suggests that commercial closed-source models possess stronger geospatial reasoning capabilities and align more closely with human judgment in disaster sensing tasks (Li et al., 2024b). Among the open-source models, Qwen-2.5-32B obtains the top results with lowest RMSE for the Napa case.

RMSE and correlation can be inconsistent. Another noteworthy observation from Table 1 is that the two evaluation metrics—correlation and RMSE—do not always align. For instance, Llama-3.2-11B shows a high correlation but a poor RMSE at the county level for the Napa case. This discrepancy arises because correlation captures the model's ability to predict relative ordering of seismic impacts, whereas RMSE reflects the absolute prediction errors. Thus, even when models effectively estimate the relevant severity (lower RMSE), they may not correctly distinguish between higher- or lower-impact areas (lower correlation). This suggests the model can correctly rank the relative severity of events but struggles to accurately predict the absolute MMI values.

We then evaluate the influence of incorporating demonstration methods with RAG and ICL on the simulation, as shown in Figure 3. These demonstration techniques can enhance the capacity of LLMs to generalize from provided information. In particular, the enhanced models consistently show higher performance, illustrated by decreasing $RMSE_Z$. It is also well-noted that even limited demonstrations can greatly improve model predictions.

4.2 Input Feature Analysis

In this section, we examine how input features beyond earthquake and location information affect LLMs' simulation performance. We conduct experiments using Claude-3.5-haiku and Qwen-2.5-7B, with the results presented in Figure 4. Interestingly, we find that only street view information contributes to improved simulation performance. In contrast, removing any of the other three features alone—geospatial, building, or socioeconomic data—can decrease the zip code-level RMSE.

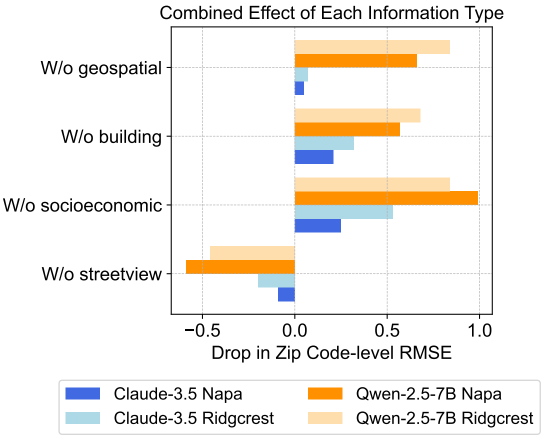


Figure 4: Input information feature analysis results.

We attribute this performance degradation to several potential factors: (1) limitations of LLMs in processing numerical information as they could complicate LLM's reasoning process (Yin et al., 2024; Bodensohn et al., 2025); (2) the lack of domain-specific knowledge required to interpret geospatial, building, and community-related data (Gao et al., 2024); (3) the inherent limitations of the self-attention mechanism in capturing spatial adjacency and performing geometric reasoning (Requeima et al., 2024).

4.3 Output Reasoning Analysis

To complement our quantitative evaluations, we examine how GPT-4.1-mini and Qwen-2.5-32B (given their lowest RMSE in Table 1) reason when predicting MMI values. Figure 5 illustrates how they model the relationship between epicentral distance, VS30 values, and MMI predictions. We summarize our findings as below:

LLMs capture seismic attenuation but underutilize local site conditions. As shown in Figure 5, both models display a clear negative correlation between epicentral distance and predicted MMI, most notably in Qwen2.5-32B's Napa earthquake predictions, which indicates that LLMs have internalized the concept of seismic attenuation. However, the relationship between VS30 values (a proxy for local ground conditions) and MMI is weak across both models. High MMI values occur almost exclusively near the epicenter, suggesting limited sensitivity to local site effects.

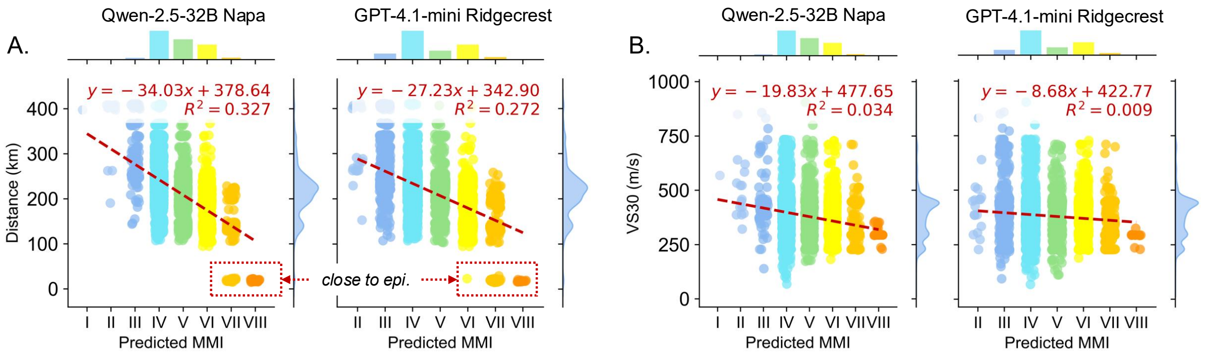


Figure 5: Output reasoning analysis in terms of (A) distance (where the x-axis is the predicted MMI, and y-axis is the distance from the epicenter (km)) and (B) VS30 (where the x-axis is the predicted MMI, and the y-axis the local site condition represented by VS30 (m/s)).

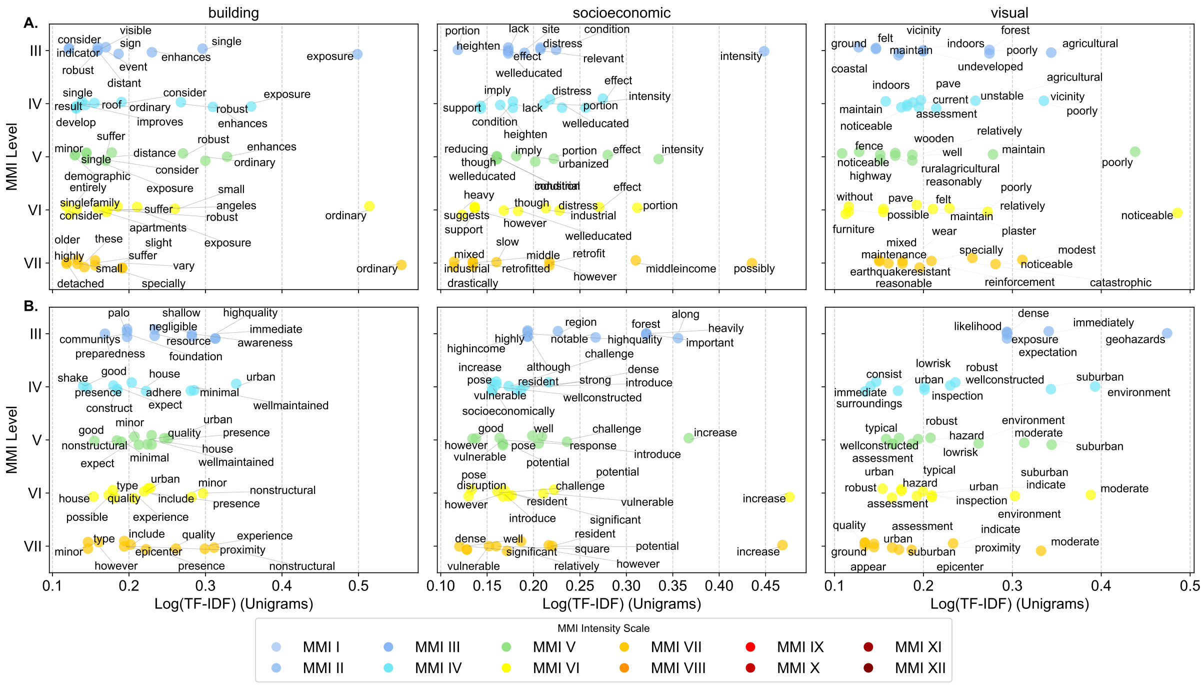


Figure 6: Output reasoning analysis in terms of different input features with (A) Qwen-2.5-32B for the Napa case, and (B) GPT-4.1-mini for the Ridgecrest case. The x-axis is the log(TF-IDF), while the y-axis is the predicted MMI.

LLMs use distinct lexical cues for MMI reasoning across multimodal inputs. Figure 6 presents a taxonomy of language used by both models across three perspectives: buildings, socioeconomic context, and street-level imagery. Unigram analysis reveals that GPT-4.1-mini and Qwen2.5-32B employ different vocabularies and reasoning styles. For the building assessment, GPT-4.1-mini adopts a descriptive and hedged reasoning style. At low MMI levels, it uses terms like "enhance," "robust," and "improve," while shifting to "suffer" and "detached" at higher levels. Qwen2.5-32B relies on more technical terms, such as "compliance" and "stringent" at low levels, and "crack" and "susceptibility" at higher ones.

Socioeconomic reasoning diverges in focus and tone. GPT-4.1-mini links lower MMI levels to terms like "urbanized" and "welleducated," and higher levels to "industrial" and "heavy," occasionally incorporating cautious language such as "possibly" or "suggest." Qwen2.5-32B emphasizes systemic vulnerability, shifting from "limited" and "stable" to "cascade" and "amplification" as predicted MMI increases.

Visual reasoning contrasts environmental vs. structural emphasis. GPT-4.1-mini references broader environmental cues, from "forest" and "agricultural" at lower levels to "catastrophic" at higher ones. In contrast, Qwen2.5-32B focuses on structural compliance, mentioning "firm," "code," and “reinforced” in a progression of seismic risk.

5 Discussion

Based on our experimental result analysis, we conclude the following insights and opportunities for future study:

- LLMs can effectively simulate human-centric seismic risks, showing strong alignment with real-world USGS DYFI reports. This can help develop pre-event impact assessment. One promising direction is to explore broader application in LLM-based simulations for disaster responses.
- Simulation performance depends on model type, size, and input information: closed-source and larger models generally perform better; techniques like RAG and ICL help; street-level imagery boosts accuracy, while structured data may hinder it due to modality alignment limitations. Future works can explore more effective prompting strategies and reasoning structures to further improve the simulation performance.
- LLMs show diverse reasoning styles and strong practical value, as model-specific rhetorical patterns reflect architectural and data differences. These differences suggest the potential impact of training data and model architecture on LLMs’ reasoning styles. It would be meaningful to further explore the reasoning mechanisms of LLMs when used as world models.

Our study has significant practical implications. Leveraging LLMs and open-source data to simulate seismic risk enables authorities to proactively assess potential disaster impacts. Integrating our framework into early-warning systems can also help identify vulnerable communities and enhance disaster response.

6 Conclusions

Our study demonstrates the potential of using LLMs to simulate seismic risk before an earthquake occurs. The alignment between model predictions and real-world reports highlights the importance of multimodal inputs and advanced LLM techniques like RAG and ICL in simulation settings. Moreover, our findings reveal that simulation performance can vary across LLMs and input features. These results make a meaningful contribution to advancing data-driven, human-centric simulation with LLMs for real-world scenarios.

7 Limitations

Several limitations warrant further consideration. First, our experiments focus only on two earthquake cases, including the 2014 Napa and 2019 Ridgecrest events, which may not fully represent global variations in seismic hazards, urban densities, and construction practices. Therefore, generalizability requires additional validation. In particular, applying this framework to other regions may be limited by data-sharing restrictions and the reliance on USGS DYFI reports, requiring alternative human-perception datasets for international validation. Practical consideration for deployment can be found in Appendix I.

Second, despite employing stratified random sampling within zip codes, potential biases remain due to gaps in data availability, such as incomplete Google Street View coverage or inconsistencies in socioeconomic and building datasets. This may lead to underrepresentation of certain neighborhoods or misalignment with high-risk zones.

Third, although our framework integrates diverse heterogeneous data, we do not conduct detailed feature selection or examine individual parameters (e.g., housing age, infrastructure proximity). Consequently, interpreting model behavior at a granular level requires further investigation.

8 Ethics

Responsible data use. Our research relies exclusively on publicly available and anonymized datasets, including geospatial, demographic, and imagery data from USGS, OpenStreetMap, and Google Street View. All analyses are conducted at aggregated administrative levels (zip code and county), minimizing risks to individual privacy.

Deployment guidance. It is important to recognize that LLM-based simulations cannot fully capture the complexity and diversity of individual experiences in real disaster scenarios. Therefore, model predictions should be viewed as supportive tools rather than replacements for expert judgment, and used responsibly with validated empirical assessments in practical decision-making contexts.

References

2022. American community survey (acs).

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

Temitope Akinboyewa, Huan Ning, M Naser Lessani, and Zhenlong Li. 2024. Automated floodwater depth estimation using large multimodal model for rapid flood mapping. *Computational Urban Science*, 4(1):12.

Gail M. Atkinson and David J. Wald. 2007. "did you feel it?" intensity data: A surprisingly good measure of earthquake ground motion. *Seismological Research Letters*, 78(3):362–368.

Jan-Micha Bodensohn, Ulf Brackmann, Liane Vogel, Anupam Sanghi, and Carsten Binnig. 2025. Unveiling challenges for llms in enterprise data engineering. *arXiv preprint arXiv:2504.10950*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Angelo Cardellicchio, Sergio Ruggieri, Valeria Leggieri, and Giuseppina Uva. 2023. A machine learning framework to estimate a simple seismic vulnerability index from a photograph: the vulma project. *Procedia Structural Integrity*, 44:1956–1963. XIX ANIDIS Conference, Seismic Engineering in Italy.

Gregory G Deierlein, Frank McKenna, Adam Zsarnóczay, Tracy Kijewski-Correa, Ahsan Kareem, Wael Elhaddad, Laura Lowes, Matthew J Schoettler, and Sanjay Govindjee. 2020. A cloud-enabled application framework for simulating regional-scale impacts of natural hazards on the built environment. *Frontiers in Built Environment*, 6:558706.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Ivan Frigerio, Stefania Ventura, Daniele Strigaro, Matteo Mattavelli, Mattia De Amicis, Silvia Mugnano, and Mario Boffi. 2016. A gis-based approach to identify the spatial variability of social vulnerability to seismic hazard in italy. *Applied geography*, 74:12–22.

Yanjun Gao, Skatje Myers, Shan Chen, Dmitriy Dligach, Timothy A Miller, Danielle Bitterman, Matthew Churpek, and Majid Afshar. 2024. When raw data prevails: Are large language model embeddings effective in numerical data representation for medical machine learning applications? *arXiv preprint arXiv:2408.11854*.

Google. 2025. Google Maps Platform. `https://developers.google.com/maps`. Accessed: 2025-05-19.

Yu Gu, Kai Zhang, Yuting Ning, Boyuan Zheng, Boyu Gou, Tianci Xue, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, and 1 others. 2024. Is your llm secretly a world model of the internet? model-based planning for web agents. *arXiv preprint arXiv:2411.06559*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Haley Hostetter, MZ Naser, Xinyan Huang, and John Gales. 2024. Large language models in fire engineering: An examination of technical questions against domain knowledge. *arXiv preprint arXiv:2403.04795*.

Iunio Iervolino. 2023. Implications of gmpe's structure for multi-site seismic hazard. *Soil Dynamics and Earthquake Engineering*, 172:108022.

Rebecca Louise Jones, Debarati Guha-Sapir, and Sandy Tubeuf. 2022. Human and economic impacts of natural disasters: can we trust the global data? *Scientific data*, 9(1):572.

Moustafa Moufid Kassem, Fadzli Mohamed Nazri, and Ehsan Noroozinejad Farsangi. 2020. The seismic vulnerability assessment methodologies: A state-of-the-art review. *Ain Shams Engineering Journal*, 11(4):849–864.

Maja Kucharczyk and Chris H Hugenholtz. 2021. Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities. *Remote Sensing of Environment*, 264:112577.

Geon Lee, Wenchao Yu, Kijung Shin, Wei Cheng, and Haifeng Chen. 2025. Timecap: Learning to contextualize, augment, and predict time series events with large language model agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18082–18090.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Bowen Li, Zhaoyu Li, Qiwei Du, Jinqi Luo, Wenshan Wang, Yaqi Xie, Simon Stepputtis, Chen Wang, Katia Sycara, Pradeep Ravikumar, and 1 others. 2024a. Logicity: Advancing neuro-symbolic ai with abstract urban simulation. *Advances in Neural Information Processing Systems*, 37:69840–69864.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2024b. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Dawei Li, Zhen Tan, and Huan Liu. 2025a. Exploring large language models for feature selection: A data-centric perspective. *ACM SIGKDD Explorations Newsletter*, 26(2):44–53.

Dawei Li, Zhen Tan, Peijia Qian, Yifan Li, Kumar Chaudhary, Lijie Hu, and Jiayi Shen. 2025b. Smoa: Improving multi-agent large language models with s parse m ixture-o f-a gents. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 54–65. Springer.

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*, 98:104062.

Lingyao Li, Michelle Bensi, Qingbin Cui, Gregory B Baecher, and You Huang. 2021. Social media crowdsourcing for rapid damage assessment following a sudden-onset natural hazard event. *International Journal of Information Management*, 60:102378.

Lingyao Li, Runlong Yu, Qikai Hu, Bowei Li, Min Deng, Yang Zhou, and Xiaowei Jia. 2025c. From pixels to places: A systematic benchmark for evaluating image geolocalization ability in large language models. *arXiv preprint arXiv:2508.01608*.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025d. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

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):03124003.

Rafaela Martelo, Kimia Ahmadiyehyazdi, and Ruo-Qian Wang. 2024. Towards democratized flood risk management: An advanced ai assistant enabled by gpt-4 for enhanced interpretability and public engagement. *arXiv preprint arXiv:2403.03188*.

Devin F McPhillips, Julie A Herrick, Sean Ahdi, Alan K Yong, and Scott Haefner. 2020. Updated compilation of vs30 data for the united states. *(No Title)*.

Federico Mori, Amerigo Mendicelli, Massimiliano Moscatelli, Gino Romagnoli, Edoardo Peronace, and Giuseppe Naso. 2020. A new vs30 map for italy based on the seismic microzonation dataset. *Engineering Geology*, 275:105745.

Morgan P Moschetti, Brad T Aagaard, Sean K Ahdi, Jason Altekruse, Oliver S Boyd, Arthur D Frankel, Julie Herrick, Mark D Petersen, Peter M Powers, Sanaz Rezaeian, and 1 others. 2024. The 2023 us national seismic hazard model: Ground-motion characterization for the conterminous united states. *Earthquake Spectra*, 40(2):1158–1190.

Hakan T Otal, Eric Stern, and M Abdullah Canbaz. 2024. Llm-assisted crisis management: Building advanced llm platforms for effective emergency response and public collaboration. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 851–859. IEEE.

Frederik Ramm and Jochen Topf. 2010. *OpenStreetMap: Die freie Weltkarte nutzen und mitgestalten*. Lehmanns Media.

James Requeima, John Bronskill, Dami Choi, Richard Turner, and David K Duvenaud. 2024. Llm processes: Numerical predictive distributions conditioned on natural language. *Advances in Neural Information Processing Systems*, 37:109609–109671.

Ismaël Riedel, Philippe Guéguen, Mauro Dalla Mura, Erwan Pathier, Thomas Leduc, and Jocelyn Chanussot. 2015. Seismic vulnerability assessment of urban environments in moderate-to-low seismic hazard regions using association rule learning and support vector machine methods. *Natural hazards*, 76:1111–1141.

Argho Sarkar, Tashnim Chowdhury, Robin Roberson Murphy, Aryya Gangopadhyay, and Maryam Rahnemoonfar. 2023. Sam-vqa: Supervised attention-based visual question answering model for post-disaster damage assessment on remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16.

Soudabeh Taghian Dinani, Doina Caragea, and Nikesh Gyawali. 2023. Disaster tweet classification using fine-tuned deep learning models versus zero and few-shot large language models. In *International Conference on Data Management Technologies and Applications*, pages 73–94. Springer.

Zhen Tan, Chengshuai Zhao, Raha Moraffah, Yifan Li, Song Wang, Jundong Li, Tianlong Chen, and Huan Liu. 2024. Glue pizza and eat rocks-exploiting vulnerabilities in retrieval-augmented generative models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1610–1626.

U.S. Geological Survey. 1989. The modified mercalli intensity (mmi) scale. `https://pubs.usgs.gov/gip/earthq4/severitygip.html`. Accessed: 2024-05-18.

U.S. Geological Survey. 2014a. Did you feel it? – community internet intensity map for 2014 napa earthquake. `https://earthquake.usgs.gov/earthquakes/eventpage/nc72282711/dyfi/responses`. Accessed: 2024-05-18.

U.S. Geological Survey. 2014b. M 6.0 - south napa earthquake, california. https://earthquake.usgs.gov/earthquakes/eventpage/nc72282711/executive. Accessed: 2024-05-18.

U.S. Geological Survey. 2019a. Did you feel it? – community internet intensity map for 2019 ridgecrest earthquake. https://earthquake.usgs.gov/earthquakes/eventpage/ci38457511/dyfi/intensity. Accessed: 2024-05-18.

U.S. Geological Survey. 2019b. M 7.1 - 2019 ridgecrest earthquake sequence. https://earthquake.usgs.gov/earthquakes/eventpage/ci38457511/executive. Accessed: 2024-05-18.

David J Wald, Bruce C Worden, Vincent Quitoriano, and Kris L Pankow. 2006. Shakemap® manual. *Technical Manual, users guide, and software guide Version*.

Chenguang Wang, Davis Engler, Xuechun Li, James Hou, David J Wald, Kishor Jaiswal, and Susu Xu. 2024a. Near-real-time earthquake-induced fatality estimation using crowdsourced data and large-language models. *International Journal of Disaster Risk Reduction*, 111:104680.

Gelan Wang, Yu Liu, Shukai Liu, Ling Zhang, and Liqun Yang. 2025. Remflow: Rag-enhanced multi-factor rainfall flooding warning in sponge airports via large language model. *International Journal of Machine Learning and Cybernetics*, pages 1–21.

Yue Wang, Tianfan Fu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Yingzhou Lu, Honghao Gao, Jian Wu, and Jintai Chen. 2024b. Twin-gpt: digital twins for clinical trials via large language model. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Gwen White and Sadie Liptak. 2025. Small business continuity and disaster recovery plans using ai and chatgpt.

Lionel Wong, Gabriel Grand, Alexander K Lew, Noah D Goodman, Vikash K Mansinghka, Jacob Andreas, and Joshua B Tenenbaum. 2023. From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*.

Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. 2023. Language models meet world models: Embodied experiences enhance language models. *Advances in neural information processing systems*, 36:75392–75412.

Yuwei Yan, Qingbin Zeng, Zhiheng Zheng, Jingzhe Yuan, Jie Feng, Jun Zhang, Fengli Xu, and Yong Li. 2024. Opencity: A scalable platform to simulate urban activities with massive llm agents. *arXiv preprint arXiv:2410.21286*.

Hang Yang, Hao Chen, Hui Guo, Yineng Chen, Ching-Sheng Lin, Shu Hu, Jinrong Hu, Xi Wu, and Xin Wang. 2024. Llm-medqa: Enhancing medical question answering through case studies in large language models. *arXiv preprint arXiv:2501.05464*.

Shiping Yang, Jie Wu, Wenbiao Ding, Ning Wu, Shining Liang, Ming Gong, Hengyuan Zhang, and Dongmei Zhang. 2025. Quantifying the robustness of retrieval-augmented language models against spurious features in grounding data. *arXiv preprint arXiv:2503.05587*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Yunhang Shen, Chunjiang Ge, Yan Yang, Zuwei Long, Yuhan Dai, Yongdong Luo, Haoyu Cao, and 1 others. 2024. Sparrow: Data-efficient video-llm with text-to-image augmentation. *arXiv preprint arXiv:2411.19951*.

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

Qian Yu, Chaofeng Wang, Frank McKenna, Stella X Yu, Ertugrul Taciroglu, Barbaros Cetiner, and Kincho H Law. 2020. Rapid visual screening of soft-story buildings from street view images using deep learning classification. *Earthquake Engineering and Engineering Vibration*, 19:827–838.

Yiyao Yu, Yuxiang Zhang, Dongdong Zhang, Xiao Liang, Hengyuan Zhang, Xingxing Zhang, Ziyi Yang, Mahmoud Khademi, Hany Awadalla, Junjie Wang, and 1 others. 2025. Chain-of-reasoning: Towards unified mathematical reasoning in large language models via a multi-paradigm perspective. *arXiv preprint arXiv:2501.11110*.

Chenhui Zhang and Sherrie Wang. 2024. Good at captioning bad at counting: Benchmarking gpt-4v on earth observation data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7839–7849.

Hengyuan Zhang, Yanru Wu, Dawei Li, Sak Yang, Rui Zhao, Yong Jiang, and Fei Tan. 2024. Balancing speciality and versatility: a coarse to fine framework for supervised fine-tuning large language model. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7467–7509.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2023. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36:31967–31987.

A Modified Mercalli Intensity (MMI) Scale

Table 2 shows the MMI scale used to support the classification of seismic risks in our study.

B Prompt Template Design

MMI Level	Description
I	Not felt except by a very few under especially favorable conditions.
II	Felt only by a few persons at rest, especially on upper floors of buildings. Delicately suspended objects may swing.
III	Felt quite noticeably by persons indoors, especially on upper floors of buildings. Many people do not recognize it as an earthquake. Standing motor cars may rock slightly. Vibration similar to the passing of a truck. Duration estimated.
IV	Felt indoors by many, outdoors by few during the day. At night, some awakened. Dishes, windows, doors disturbed; walls make cracking sound. Sensation like heavy truck striking building. Standing motor cars rocked noticeably.
V	Felt by nearly everyone; many awakened. Some dishes, windows broken. Unstable objects overturned. Pendulum clocks may stop.
VI	Felt by all, many frightened. Some heavy furniture moved; a few instances of fallen plaster. Damage slight.
VII	Damage negligible in buildings of good design and construction; slight to moderate in well-built ordinary structures; considerable damage in poorly built or badly designed structures; some chimneys broken.
VIII	Damage slight in specially designed structures; considerable damage in ordinary substantial buildings with partial collapse. Damage great in poorly built structures. Fall of chimneys, factory stacks, columns, monuments, walls. Heavy furniture overturned.
IX	Damage considerable in specially designed structures; well-designed frame structures thrown out of plumb. Damage great in substantial buildings, with partial collapse. Buildings shifted off foundations.
X	Some well-built wooden structures destroyed; most masonry and frame structures destroyed with foundations. Rail bent.
XI	Few, if any (masonry) structures remain standing. Bridges destroyed. Rails bent greatly.
XII	Damage total. Lines of sight and level are distorted. Objects thrown into the air.

Table 2: Description of the Modified Mercalli Intensity (MMI) scale (U.S. Geological Survey, 1989).

```
SYSTEM_PROMPT = """
You are a seismic expert specialized in earthquake damage assessment and disaster response. You analyze earthquake data, local conditions, and building characteristics to provide damage assessments using the Modified Mercalli Intensity (MMI) scale.
"""
```

```
EARTHQUAKE_PROMPT = """
The earthquake happened date is 2025-06-01.

Here is the EARTHQUAKE information.
- Epicenter: {eq_place}
- Coordinates: {eq_lat}, {eq_lng}
- Magnitude: {eq_magnitude} mw
- Depth: {eq_depth} km

YOUR LOCATION information is listed below.
- State: {state}
- City: {city}
- Zip code: {zip code}
- Coordinates: {lat}, {lng}
- Distance from epicenter: {distance} km

## Geospatial features in YOUR LOCATION
- VS30 at your location: {vs30} m/s
(VS30 represents the time-averaged shear-wave velocity (VS) to a depth of 30 meters, which is a key index to account for seismic site conditions)

## Building Description in YOUR LOCATION (within a 100-meter radius)
- Building description: {building}

## Community Socioecnomics and Demographics in YOUR LOCATION (at Cencus Block Group level)
- Population density: {population_density} people per square km
- Urban population percentage: {urban_population_pct}%
- Over 65 percentage: {over_65_rate}%
- Median household income: ${median_household_income}/year
- Education (bachelor's or higher): {education}%

## Visual Context in YOUR LOCATION
The image provided shows your surrounding environment and infrastructure.

Based on the information provided, ASSESS the potential earthquake damage level using the Modified Mercalli Intensity (MMI) scale.
1. Identify the damage level.
2. Explain your reasoning by addressing the following factors and considering the visual context.
   - Distance to the epicenter and earthquake magnitude
   - Geospatial features
   - Infrastructure quality and building characteristics
```

```
    - Population density and
        socioeconomic vulnerabilities
    - Visual image of surroundings

The following is an abbreviated
    description of the 12 levels of
    Modified Mercalli intensity. {MMI
     Scale}

Output the result in JSON format:
{{
    "Reasoning": "<Provide reasoning
        >"
    "MMI": "<Respond MMI level>",
}}
"""
```

C Earthquake Scenarios

2014 Napa Earthquake (U.S. Geological Survey, 2014b). On August 24, 2014, a magnitude 6.0 earthquake struck near Napa, California, causing significant structural damage despite its moderate magnitude. Approximately 613 buildings were tagged for various degrees of structural integrity concerns, including fractures, road cracks, and damage to wine storage facilities. The earthquake resulted in one death and nearly 200 injuries.

2019 Ridgecrest Earthquake (U.S. Geological Survey, 2019b). The Ridgecrest earthquake occurred on July 6, 2019, with a magnitude of 7.1, significantly larger than the Napa event but with fewer human casualties. The quake damaged around 50 homes, caused gas leaks and road cracks, and triggered fires in residential properties. Significant infrastructural damage occurred at the Naval Air Station, and widespread power outages were reported.

D Data Leakage Test

To further assess the potential for data leakage in the LLMs used in our experiments, we conduct a leakage test on Claude-3.5-haiku and GPT-4.1-mini, the two best-performing models in our main study. Specifically, we remove city and state names from the prompt—two elements most likely to serve as shortcuts for the models to associate with MMI levels and potentially memorize. As shown in Table 3, the removal of location information does not significantly affect the models' simulation performance. It is fair to rule out the possibility of data leakage in our main results.

E Spatial Distribution of Predicted MMI by LLMs

Figure 7 shows the spatial distribution of predicted MMI for the 2014 Napa earthquake and the 2019 Ridgecrest earthquake at zip code level from different LLMs.

F Scaling Law Analysis

To investigate the impact of LLM scaling laws on simulation performance, we conduct a scaling analysis using the Llama-3.2 and Qwen-2.5 model families. As shown in Table 4, we observe that performance in simulation generally improves with model size, excepting Qwen-2.5-72B. This indicates that the simulation performance tend to be strengthened with a larger size of LLMs.

G Fine-grained City-level Analysis

The following tables present a comparative analysis of MMI predictions from LLMs for the two investigated seismic events. The predictions from each model are juxtaposed with the actual average MMI values recorded for cities to evaluate their performance and accuracy. This city-level comparison provides a more fine-grained analysis of each model's predictive performance across different locations and intensities.

Based on the results, we have listed some key observations. First, LLMs show variation in their MMI predictions at the city level, with most models consistently overestimating damage compared to actual MMI values, especially for the 2014 Napa earthquake. GPT-4.1-mini shows the largest prediction errors, consistently overestimating MMI values by 2-3 levels (e.g., predicting 6.43 vs actual 3.80 for San Francisco). The Qwen models show intermediate performance, with Qwen2.5-VL-32B being more conservative than its 72B counterpart. The case varies significantly between earthquake events, with all models performing better on the 2021 Ridgecrest earthquake compared to the 2014 Napa earthquake. For Ridgecrest, model predictions show better alignment with actual MMI values, particularly at the epicenter, where models predicted 6.6-7.7 versus the actual 7.0.

Most LLMs show consistency in their relative city rankings within each earthquake event, with most models generally maintaining similar ordering of cities as compared to the ground truth. This suggests that while the models struggle with precise MMI calibration, they can capture relation-

Model	Earthquake prompt				w/o location			
	$\text{RMSE}_Z \downarrow$	$\text{Corr}_Z \uparrow$	$\text{RMSE}_C \downarrow$	$\text{Corr}_C \uparrow$	$\text{RMSE}_Z \downarrow$	$\text{Corr}_Z \uparrow$	$\text{RMSE}_C \downarrow$	$\text{Corr}_C \uparrow$
`claude-3-5-haiku`	2.11	0.58	2.05	0.70	2.35	0.38	2.26	0.62
`gpt-4.1-mini`	2.56	0.61	2.48	0.67	2.67	0.62	2.58	0.73

Table 3: Experiment results on data leakage test.

Model Family	Model Size	Napa RMSE	Ridgecrest RMSE
Llama-3.2	11B	3.19	3.22
	90B	2.62	2.06
Qwen-2.5	3B	3.63	3.88
	7B	1.79	1.53
	32B	1.59	0.99
	72B	2.17	1.39

Table 4: Scaling law analysis: Zip code-level RMSE across model sizes (in billions of parameters).

City	Count	claude-3.5 -haiku	gpt-4.1 -mini	gpt-4o	llama-3.2 -90b	qwen2.5 -vl-32b	qwen2.5 -vl-72b	Actual average MMI
San Francisco	750	5.12	6.43	6.00	6.27	5.16	5.79	3.80
Berkeley	360	5.39	6.44	5.81	6.63	5.45	5.71	3.86
Oakland	350	5.26	6.57	5.90	6.58	5.27	5.53	3.57
Santa Rosa	300	5.18	6.52	6.04	6.56	5.29	5.71	4.00
San Mateo	200	4.50	6.02	6.00	6.11	4.22	5.61	3.75
Mountain View	150	4.62	5.65	5.96	5.61	4.09	4.98	3.00
Walnut Creek	100	5.28	6.25	6.04	6.40	5.33	5.84	4.00
San Rafael	100	5.39	6.49	5.92	6.80	5.83	5.59	3.00
Palo Alto	100	4.78	5.66	5.89	5.91	4.12	5.32	3.00
Redwood City	100	4.71	5.76	5.72	5.80	4.07	5.05	3.50
San Jose	100	5.41	5.32	6.00	5.96	4.08	4.98	3.00
Fremont	100	4.57	6.05	6.00	5.92	3.97	5.65	3.00
Fairfield	100	5.70	6.63	6.63	6.83	6.11	5.36	4.50
Vallejo	100	5.67	6.86	7.00	6.96	6.86	5.26	6.00
Vacaville	100	5.37	6.49	6.12	6.64	5.46	5.73	4.00
Sunnyvale	100	4.88	5.55	5.93	5.78	3.99	5.11	2.50
Davis	100	5.13	6.12	5.96	6.09	4.34	5.83	3.00
Concord	100	5.54	6.36	5.99	6.57	5.05	5.82	4.00
Petaluma	76	5.17	6.53	6.14	6.61	5.46	5.36	4.53
Napa	73	5.73	6.78	6.84	6.83	6.67	5.66	7.56

Table 5: City-level MMI predictions vs. actual MMI for the 2014 Napa Earthquake.

ships about earthquake impact distribution, making them potentially valuable for comparative damage assessment and emergency response prioritization.

H Comparison with Traditional Methods

ShakeMap is a highly effective method that quantifies physical parameters like ground shaking intensity and infrastructure vulnerability (Wald et al., 2006). We provide quantitative comparison against the USGS ShakeMap product and standard machine learning models to ground our performance evaluation.

H.1 Comparison with USGS ShakeMap

Table 7 compares our pre-event LLM simulations against the post-event ShakeMap results, both evaluated against ground-truth “Did You Feel It?” (DYFI) reports.

The post-event ShakeMap product aligns more closely with DYFI reports, as reflected in its lower RMSE and higher correlation. This advantage stems from ShakeMap’s reactive nature: it is generated minutes after an earthquake using real-time ground-motion recordings from seismic stations, interpolated across a regional grid. By contrast, our LLM-based method operates in a pre-event simulation and emphasizes the human-centered dimension

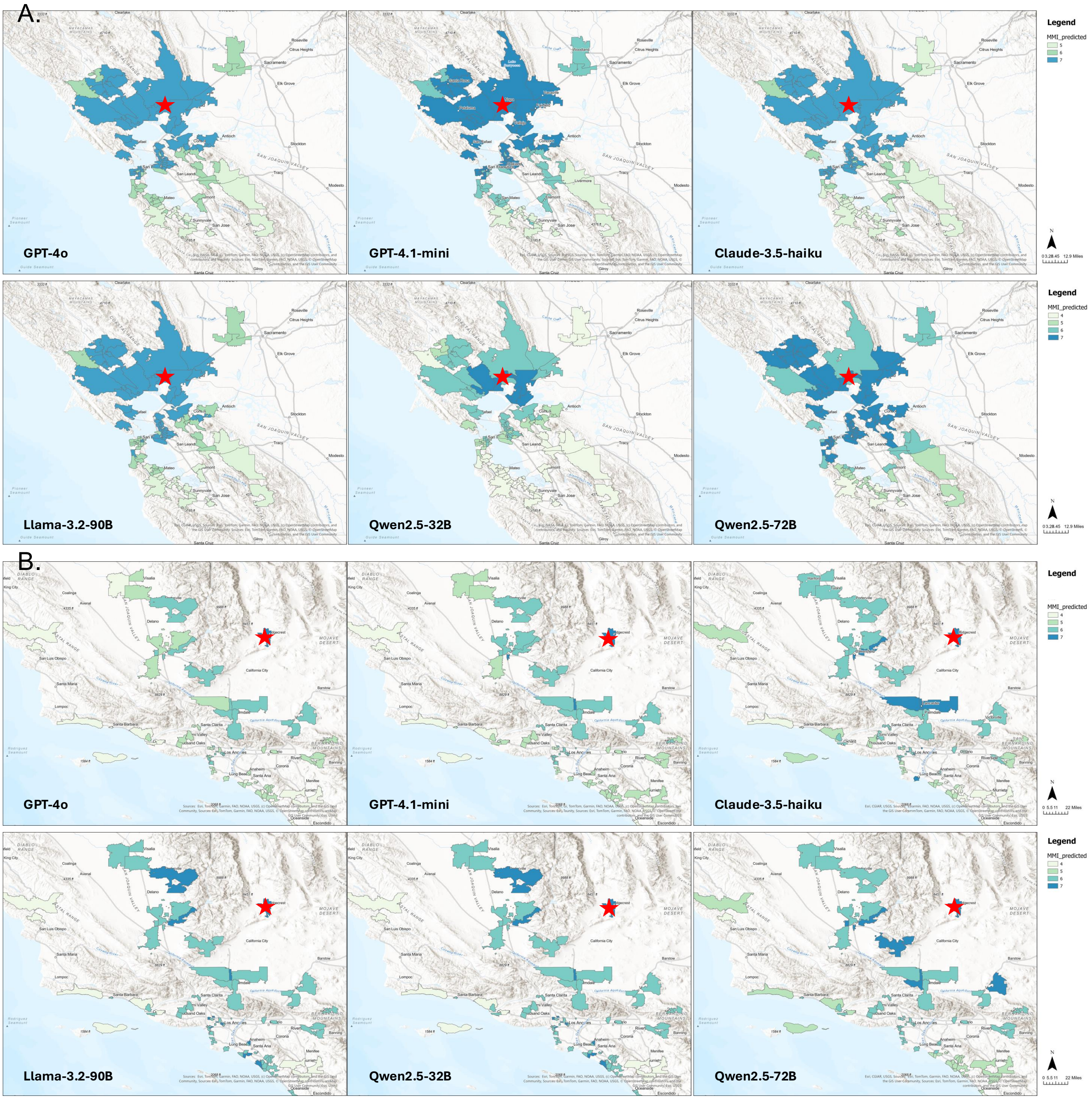


Figure 7: Spatial distribution of predicted MMI at zip code level: (A) The 2014 Napa earthquake, and (B) the 2019 Ridgecrest earthquake based on GPT4o, GPT-4.1-mini, Claude-3.5-haiku, Llama-3.2-90B, Qwen-2.5-32B, and Qwen-2.5-72B. The red star sign shows the epicenter of the earthquake. These maps compare outputs from different LLMs, showing consistent geospatial patterns with high-intensity predictions concentrated near epicenters

of how individuals perceive seismic impacts.

H.2 Comparison with Machine Learning Baselines

We further benchmark our approach against standard machine learning models trained for the same pre-event simulation task. Using the 2014 Napa earthquake data for training and the 2019 Ridgecrest event for testing, we train five models on the same multimodal data provided to our LLM pipeline.

Our evaluation reveals two advantages of using LLMs for this task. First, the best-performing LLM achieves a zip code-level Pearson correlation of 0.75 with DYFI reports (Table 7). In contrast, the strongest traditional ML model reaches a correlation of only 0.62 (Table 8). This demonstrates the LLM's superior ability to synthesize complex, multimodal inputs for pre-event simulation. Second, unlike traditional models that output a single numeric value, the LLM generates chain-of-thought explanations for its MMI predictions. These narratives reference factors such as distance from the epicenter, local site conditions, building materials,

City	**Count**	claude-3.5 -haiku	gpt-4.1 -mini	gpt-4o	llama-3.2 -90b	qwen2.5 -vl-32b	qwen2.5 -vl-72b	Actual average MMI
Los Angeles	500	4.80	5.49	5.61	6.00	4.58	5.43	4.00
Las Vegas	300	4.25	4.36	4.93	5.96	4.62	5.09	4.00
Bakersfield	250	4.93	5.47	5.96	6.02	5.33	5.15	4.00
San Diego	200	4.26	3.87	4.11	5.87	3.58	3.78	3.00
Lancaster	150	4.73	5.49	6.22	6.11	5.00	5.28	4.67
Huntington Beach	100	4.57	4.36	5.62	6.03	3.96	4.56	4.00
Redondo Beach	100	4.46	4.77	5.95	5.98	4.06	4.81	4.00
Simi Valley	100	4.68	4.70	5.50	6.01	4.31	5.51	4.00
Henderson	100	3.90	4.32	4.82	5.85	4.25	4.69	3.50
Mission Viejo	100	3.96	4.31	5.88	5.82	4.05	4.50	4.00
Anaheim	100	4.17	4.42	5.88	5.99	4.21	5.11	4.00
Ventura	100	3.96	3.99	5.29	5.78	3.90	4.77	4.00
Palmdale	100	5.03	5.61	6.10	5.95	4.80	5.41	4.50
Paso Robles	50	4.54	3.76	5.00	5.67	3.16	3.80	3.00
Pasadena	50	5.08	5.60	5.62	6.00	4.76	5.58	4.00
Porterville	50	4.90	5.28	6.16	6.10	4.52	5.24	4.00
Rancho Cucamonga	50	4.84	4.62	4.70	6.04	4.06	5.28	4.00
Rancho Santa Margarita	50	4.12	4.36	5.16	5.84	3.59	4.56	4.00
Redlands	50	4.48	4.70	5.30	5.98	4.48	5.12	4.00
Palm Springs	50	4.48	4.18	5.54	5.92	4.10	5.28	4.00
Aliso Viejo	50	4.22	4.12	5.78	5.92	4.02	4.34	4.00
San Clemente	50	3.92	4.04	5.02	5.88	4.10	4.08	4.00
Ridgecrest	50	6.64	7.24	7.66	7.30	7.50	6.74	7.00

Table 6: City-level MMI predictions vs. actual MMI for the 2019 Ridgecrest Earthquake.

Model	**2014 Napa**		**2019 Ridgecrest**	
	$\mathrm{RMSE}_Z \downarrow$	$\mathrm{Corr}_Z \uparrow$	$\mathrm{RMSE}_Z \downarrow$	$\mathrm{Corr}_Z \uparrow$
GPT-4o	2.43	0.77	1.97	0.75
GPT-4.1-mini	2.56	0.61	0.92	0.64
Claude-3.5-haiku	2.11	0.58	1.35	0.59
Llama-3.2-90B	2.62	0.57	2.06	0.62
Qwen-2.5-32B	1.59	0.70	0.79	0.71
Qwen-2.5-72B	2.17	0.46	0.44	0.64
ShakeMap (Post-event)	**1.08**	**0.81**	**0.72**	**0.79**

Table 7: Performance comparison of LLM simulations vs. post-event ShakeMap.

Table 8: Performance of pre-event machine learning baselines on the 2019 Ridgecrest earthquake.

Model	$\mathrm{RMSE}_Z \downarrow$	$\mathrm{RMSE}_C \downarrow$	$\mathrm{Corr}_Z \uparrow$	$\mathrm{Corr}_C \uparrow$
Logistic Regression (Lasso)	1.93	1.76	0.55	0.30
Logistic Regression (Ridge)	1.93	1.76	0.55	0.30
Multilayer Perceptron (MLP)	1.74	1.57	0.62	0.50
Random Forest	0.99	0.81	0.12	0.17
Support-Vector Machine (SVM)	1.83	1.71	0.50	0.29
XGBoost	1.52	1.49	0.57	0.32

and social vulnerability indicators, offering emergency planners a clearer understanding of why a location is judged to be at high risk.

I Practical Consideration for Deployment

In this section, we discuss the practical considerations of our framework, including computational performance and data dependencies. For real-time latency, our full inference requires approximately 156 seconds to process 50 samples for a single zip code profile using the gpt-4.1-mini model. The associated costs and performance details are summarized in Table 9.

The availability of regional data is a key factor for the successful application of the simulation. Our approach depends on a rich set of open geospatial data, building inventories, socioeconomic factors, and street-level imagery, which allows the

Metric	**Value**
Model	gpt-4.1-mini-2025-04-14
Event	2019 Ridgecrest
Samples per zip code	50 samples
Averaged cost per zip code	$0.02 USD (50 samples)
Averaged time per zip code	156 seconds (50 samples)

Table 9: Inference metrics using the gpt-4.1-mini model.

LLM to "see" what residents see. Gaps in data sources, such as incomplete Google Street View coverage or inconsistent census data, can introduce bias, as noted in our limitations. Furthermore, applying this framework to other regions, particularly less-developed countries, may be challenging if such multimodal datasets are restricted by data sharing policies. The validation process also relies on the USGS DYFI program, which provides public response data at ZIP code resolution, primarily for U.S. earthquakes. When the model is applied internationally, the scarcity or unavailability of DYFI reports would necessitate the use of alternative, openly accessible human-perception datasets for validation.

J Samples of LLM Output

Sample output by LLMs are presented in Table 10, including their predicted MMI and specific reasoning.

Samples of LLM Reasoning Outputs with Different MMI Predictions
GPT-4.1-mini | 2019 Ridgecrest | Predicted MMI: III
The earthquake has a high magnitude of 7.1 Mw, which is capable of causing significant damage near the epicenter. However, the location in question, San Diego, CA, is approximately 339 km away from the epicenter in Ridgecrest, CA. At this distance, seismic waves would have attenuated considerably, reducing the intensity of shaking experienced. The VS30 value of 417 m/s indicates a site with relatively soft soil conditions, which can amplify shaking somewhat, but given the large distance, this effect is limited. The building stock consists mostly of general buildings and schools, likely built to modern California seismic codes, which enhances resilience. The population density is moderate, with a well-educated and relatively affluent community, reducing socioeconomic vulnerabilities. The visual context shows a well-maintained urban street with no visible signs of damage or distress, supporting the assessment of low shaking impact. Considering all these factors, the expected Modified Mercalli Intensity (MMI) level in San Diego would be around III, where shaking is noticeable indoors but causes no damage.
GPT-4.1-mini | 2019 Ridgecrest | Predicted MMI: IV
The earthquake has a high magnitude of 7.1 Mw, which is capable of causing significant damage near the epicenter. However, the location in Ventura, CA is approximately 221 km away from the epicenter in Ridgecrest, CA, which greatly reduces the intensity of shaking experienced. The VS30 value of 391 m/s indicates a site with relatively soft soil conditions, which can amplify shaking somewhat, but the distance still plays a dominant role in reducing impact. The building information is not available, but the visual context shows modern infrastructure such as a highway bridge and vehicles, suggesting reasonably good construction standards typical of California. The population density is low (65 people/km^2) with a moderate urban population percentage (25.43 %), and socioeconomic indicators such as median income and education level are moderate, implying some resilience but also potential vulnerabilities in older or poorly maintained structures. Given the distance, moderate site conditions, and infrastructure, the expected shaking intensity would be moderate with little to no structural damage. This corresponds to an MMI level around IV to V, where shaking is felt by many but damage is minimal or limited to minor non-structural effects.
GPT-4.1-mini | 2019 Ridgecrest | Predicted MMI: V
The earthquake had a high magnitude of 7.1 Mw, which is capable of causing severe damage near the epicenter. However, the location in question, Apple Valley, CA, is approximately 152.25 km away from the epicenter, which significantly reduces the intensity of shaking experienced. The VS30 value of 607 m/s indicates relatively firm ground conditions, which typically reduce amplification of seismic waves. The visual context shows a rural, sparsely populated area with no visible dense infrastructure or tall buildings, suggesting limited exposure to structural damage. The population density is low (18.73 people per square km), and the urban population is only about 32.65 %, further reducing potential damage and casualties. Socioeconomic factors such as median income and education level are moderate, but given the lack of detailed building information and the rural setting, the risk of significant structural damage is low. Considering these factors, the shaking would likely be felt by many but cause only slight damage, consistent with MMI level VI.
Qwen-2.5-32B | 2014 Napa | Predicted MMI: VI
The location is 19.27 km from the epicenter of a 6.0 Mw earthquake, indicating strong shaking. The VS30 value of 344 m/s suggests stiff soil, reducing amplification but not eliminating damage due to proximity. The presence of general buildings and houses, along with a rural setting and a high percentage of older residents, points to moderate to strong damage. The visual context shows a well-maintained residential area, which may indicate better-quality buildings, but this is not definitive.
Qwen-2.5-32B | 2014 Napa | Predicted MMI: VII
The earthquake has a magnitude of 6.0 Mw, and the location is 14.59 km from the epicenter, indicating strong shaking. The VS30 value of 354 m/s suggests moderate amplification. The area consists of general buildings, and the visual context shows industrial/commercial structures that appear relatively robust. High population density (9,318.61 people per sq km) increases risk, but affluence and education suggest better infrastructure. Considering these factors, the damage is expected to be noticeable but not severe.
Qwen-2.5-32B | 2014 Napa | Predicted MMI: VIII
The earthquake has a magnitude of 6.0 Mw and occurred 10.24 km from the epicenter. The firm ground conditions (VS30 = 508 m/s) reduce shaking intensity. The area is sparsely populated (23.04 people per sq km), with low urbanization (2.97 %) and high affluence ($93,750 median income). The visual context shows a rural landscape with minimal infrastructure, indicating low vulnerability to structural damage. Based on these factors, the expected damage level is moderate.

Table 10: Samples of LLM outputs with predicted MMI reasoning for selected earthquake scenarios.