Geophysical Journal International

Geophys. J. Int. (2024) **240,** 1281–1294 https://doi.org/10.1093/gji/ggae436
Advance Access publication 2024 December 11
GJI Seismology

Gemini and physical world: large language models can estimate the intensity of earthquake shaking from multimodal social media posts

S. Mostafa Mousavi,[1,2] Marc Stogaitis,[1] Tajinder Gadh,[1] Richard M. Allen,[1,3] Alexei Barski,[1] Robert Bosch,[1] Patrick Robertson,[4] Youngmin Cho,[1] Nivetha Thiruverahan[1] and Aman Raj[1]

[1]*Google LLC, Mountain View, CA* 94043*, USA. E-mail: mousavim@google.com,smousavi@fas.harvard.edu*
[2]*Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA* 02138*, USA*
[3]*Seismological Laboratory, University of California, Berkeley; Berkeley, CA* 94720*, USA*
[4]*Google Germany GmbH, Munich* 80636*, Germany*

Accepted 2024 November 30. Received 2024 October 6; in original form 2024 August 7

SUMMARY

This paper presents a novel approach to extract scientifically valuable information about Earth's physical phenomena from unconventional sources, such as multimodal social media posts. Employing a state-of-the-art large language model (LLM), Gemini 1.5 Pro's, we estimate earthquake ground shaking intensity from these unstructured posts. The model's output, estimated intensity values, aligns well with independent observational data. Furthermore, our results suggest that LLMs, trained on vast internet data, may have developed a unique understanding of physical phenomena. Specifically, Google's Gemini models demonstrate a simplified understanding of the general relationship between earthquake magnitude, distance and intensity, accurately describing observational data even though it is not identical to established models. These findings raise intriguing questions about the extent to which Gemini's training has led to a broader understanding of the physical world and its phenomena. The ability of Generative AI models like Gemini to generate results consistent with established scientific knowledge highlights their potential to augment our understanding of complex physical phenomena like earthquakes. The flexible and effective approach proposed in this study holds immense potential for enriching our understanding of the impact of physical phenomena and improving resilience during natural disasters. This research is a significant step toward harnessing the power of social media and AI for natural disaster mitigation, opening new avenues for understanding the emerging capabilities of Generative AI and LLMs for scientific applications.

Key words: Machine learning; Earthquake early warning; Earthquake ground motions; Earthquake hazards.

1 INTRODUCTION

Earthquakes pose significant risks to life and property, with the potential to cause widespread devastation. Due to their unpredictable nature and potential for catastrophic consequences, mitigating the devastating impact of earthquakes requires a comprehensive approach. This encompasses preparedness measures (e.g. mitigation in the form of better buildings and infrastructure), early warning systems, and effective post-earthquake response strategies (Böse *et al.* 2023). Earthquake early warning (EEW) systems have potential to play a crucial role by providing timely alerts before the arrival of strong ground shaking, enabling individuals to take protective actions such as 'drop, cover, and hold on' (Allen *et al.* 2019). Furthermore, effective post-earthquake responses include search and rescue operations, damage assessment and infrastructure restoration, which are critical to saving lives. By combining preparedness measures, early warning systems and rapid response strategies, communities can significantly reduce the impact of earthquakes and promote resilience in earthquake-prone regions.

Earthquake intensity is a qualitative measure of the effects of an earthquake at a specific location, distinct from magnitude, which quantifies the energy released at the source (Davison 1900). Intensity scales have evolved over time to characterize the severity of ground shaking based on observations of human reactions, structural damage and environmental impacts (Musson, Grünthal & Stucchi 2010). Intensity data are crucial for understanding historical earthquakes that predate instrumental recordings, allowing for magnitude estimations and insights into long-term seismicity patterns (Nuttli 1973; Johnston 1996; Ambraseys & Bilham 2003). While intensity can be correlated with ground motion parameters, such as peak ground acceleration, these relationships are complex and subject to significant scatter (Musson, Grünthal & Stucchi 2010). The intensity provides valuable insights into the actual experiences observed at a given site, making it particularly useful for assessing seismic hazards, understanding site response, ascertaining the shaking level and communicating earthquake impacts to the public (Hough *et al.* 2000; Musson 2000). Recent advancements, including online questionnaires and automatic algorithms, have improved the accuracy and reliability of intensity data collection and assessment (Wald *et al.* 1999; Musson 2009).

In the realm of EEW systems, the intensity is used to estimate the appropriate level of alert and guide immediate safety measures (Allen *et al.* 2012). In post-earthquake responses, intensity guides emergency responders in prioritizing areas for assistance and assessing the extent of damage (Wald *et al.* 2005, 2022). It also finds utility in other domains, including insurance, earthquake engineering, loss estimation and disaster planning. Despite its wide-ranging applications, there exist several practical challenges associated with measuring earthquake intensity. One primary challenge stems from the subjective nature of intensity, as it is contingent upon observed effects rather than instrumental measurements. Another challenge lies in inconsistencies in reporting and/or the significant variability of intensity even at the same distance from earthquakes of the same magnitude (Hanks & Johnston 1992; Minson *et al.* 2021) due to factors like proximity to the fault, earth structure and shallow geological properties. Lastly, measuring intensity in real time poses difficulties, as it necessitates the involvement of trained observers for data collection for properly assigning higher (damaging) intensities (Wald *et al.* 2024). These challenges can introduce uncertainties into intensity assignment.

The advent of social media platforms has fundamentally transformed the way information is disseminated during natural disasters and crises. Social media users are now recognized as valuable providers of timely information, enabling the characterization of physical-world events (Hughes & Palen 2009). They often share real-time updates, eyewitness accounts and multimedia content, offering a rich source of data for disaster response and research purposes (Earle *et al.* 2010). In the context of earthquake disaster management, several studies have shown the application of social media data for earthquake detection and damage area identification (e.g. Earle *et al.* 2011; Sakaki *et al.* 2012; Flores *et al.* 2017; Bossu *et al.* 2018), as well as for intensity estimation (e.g. Burks *et al.* 2014; Cresci *et al.* 2014; D'Auria & Convertito 2016; Mendoza *et al.* 2019). Furthermore, the increasing prevalence of CCTV cameras in urban and rural areas provides an unprecedented amount of real-world CCTV footage and video postings from social media platforms that can be harnessed for scientific research. This footage offers a unique opportunity to study natural hazards such as earthquakes, floods and wildfires in unprecedented detail. By analysing CCTV footage, scientists can track the spatiotemporal evolution of these events, identify previously hidden patterns and relationships, and gain deeper insights into the underlying physical processes as well as human behaviour during such events (e.g. Vinnell *et al.* 2022). Additionally, it can be used to validate and calibrate numerical models, which are essential tools for predicting the behaviour of natural hazards and assessing their potential impacts. However, extracting useful knowledge from social media data presents significant challenges, including issues of misinformation, irrelevant content and language variations. Sophisticated method developments are required to overcome these challenges and extract valuable insights from the often noisy data (Imran *et al.* 2015).

Generative artificial intelligence (GenAI) and large language models (LLMs) have experienced remarkable advancements in recent years, demonstrating successful applications in various domains. These powerful tools hold immense potential for further enhancing the capabilities of crowdsourcing in earthquake studies and risk mitigation efforts. GenAI and LLMs possess the ability to extract pertinent information from unstructured social media posts. In this paper, we show that they can analyse this information to estimate the intensity of local ground shaking with unprecedented ease and flexibility. We use Modified Mercalli Intensity scale (Wood & Neumann 1931; Richter 1958) for intensity estimation. This capability paves the way for novel approaches to understand collective behaviour patterns (Zhou *et al.* 2012), gain insight into the dynamics of information propagation during crisis situations (Palen & Anderson 2016, Bagrow *et al.* 2011), rapidly evaluate the impact of earthquakes, guiding emergency response efforts, provide situational awareness (Yin *et al.* 2012), and automatically collect valuable macroseismic data from unconventional sources (e.g. Bossu *et al.* 2024). By combining the collective wisdom of social media users with AI-driven analytics, stakeholders such as emergency responders, policymakers and researchers can gain a comprehensive understanding of disaster events. This includes real-time insights into the affected areas, and the severity of shaking. Consequently, EEW systems can be improved, and resilience strategies can be enhanced based on this real-time, user-generated information.

2 DATA

YouTube, X and TikTok (each ∼33 per cent) serve as the primary sources of data collected and analysed in experiments of this study. Although official application programming interfaces (APIs) are available for automated data retrieval, at this stage, we opted to collect data in the form of screenshots and screen recordings of relevant posts containing text, images, audio and/or videos. This approach allows us to directly evaluate the ability of LLMs to extract and analyse relevant information from unstructured data in its final form as presented to humans. By doing so, we avoid the additional challenges and complexities associated with working with each platform's specific API, such as rate limits, access restrictions and data formatting inconsistencies for our explorative work. We collected our data set by searching for earthquake-specific keywords, including terms like 'earthquake,' 'tremor,' 'shaking,' and specific earthquake event names. We then removed posts curated by professionals or those whose authenticity or associated earthquake could not be confidently identified. Recognizing the crucial role of location in studying spatially variable ground shaking intensity, we limited our data to posts that included at least one location identifier. These identifiers could be mentions of the post/observation location in the text or speech, or visual cues in the provided image or video, such as recognizable landmarks or prominent distinctive buildings.

This procedure resulted in a total of approximately 84 posts documenting observed or experienced ground shaking caused by 7 earthquakes occurring in the United States (Northern, Southern California, Oklahoma and New Jersey), Japan (Noto and Tohoku), and Taiwan (Hualien) between 2011 March and 2024 mid-April, spanning a magnitude range of 4.6 to 9.1 Mw (Table 1). The collected

Table 1. List of earthquakes and the number of posts collected and analysed for each event in this study.

Earthquake location	Earthquake date time (UTC)	Magnitude MW	Number of posts
Tohoku, Japan	2011–03-11 05:46:24	9.1	7
Lake Almanor, California USA	2023–05-11 23:19:41	5.5	3
Noto Peninsula, Japan	2024–01-01 07:10:09	7.5	8
Prague, Oklahoma, USA	2024–02-03 05:24:28	5.1	10
Malibu, California, USA	2024–02-09 21:47:27	4.6	2
Hualien City, Taiwan	2024–04-02 23:58:12	7.4	7
Tewksbury, New Jersey, USA	2024–04-05 14:23:20	4.8	45

data can be categorized into two main groups: (a) CCTV footage capturing the moment of shaking, and (b) social media posts or news interviews where individuals share their personal experiences of feeling the earthquake. It is important to acknowledge the potential limitations of this data collection methodology. The focus on specific social media platforms and keyword-based search may introduce bias into the data set. To address this, we employed multiple verification methods, including cross-referencing with official earthquake reports and news articles, to ensure the authenticity of the collected data and the associated earthquake events. This process was performed manually in this study but it can be automated as well.

To assess the model's ability to attend to and extract relevant information (textual, visual and/or auditory) from diverse social media posts, we intentionally included screenshots and recordings with varying characteristics. These variations encompassed different video lengths, cropping sizes, languages and background colours, simulating the wide range of formats encountered in real-world social media content. The captured views ranged from those containing the opinion or experience of a single individual, such as a short tweet, to those encompassing a tweet and a selection of its replies or comments. These replies and comments could be relevant, confirming the experience of the same earthquake or providing supplementary information about the earthquake source, or they could be irrelevant, such as jokes, unrelated discussions, or even misinformation. Additionally, both video and image views might contain unrelated information, such as advertisements or background visuals, intentionally introduced to assess the model's ability to focus on the earthquake-related content (Fig. 1).

The content within the posts provides valuable insights into the subjective experiences of individuals, including descriptions of shaking intensity, emotional responses and reports of damage to personal belongings or surrounding structures. The inclusion of irrelevant information is crucial for evaluating the model's performance in real-world scenarios, where social media posts often contain extraneous content that is not pertinent to the earthquake event. By testing the model's ability to discern relevant information from noise, we can gain a better understanding of its effectiveness and robustness in extracting valuable insights from the complex and dynamic landscape of social media data.

3 METHOD

GenAI and LLMs have revolutionized the field of natural language processing (Wei *et al.* 2022). Models such as (Bidirectional Encoder Representations from Transformers) BERT (Devlin *et al.* 2018), (Bidirectional Auto-Regressive Transformers) BART (Lewis *et al.* 2019) and (Generative Pre-trained Transformer) GPT-4 (Achiam *et al.* 2023) have demonstrated remarkable capabilities in generating coherent and contextually relevant text, translating languages, summarizing documents, and performing various other language-related tasks, often surpassing human performance in benchmarks. In this study, we utilize the Gemini 1.5 Pro (Reid *et al.* 2024) model, from Gemini (Gemini Team 2023) family, to process and analyse the collected data.

Gemini (Gemini Team 2023), a state-of-the-art large language model developed by Google, is a suite of generative AI models designed to interpret and respond to user inputs using natural language processing. What sets Gemini apart is its multimodality, meaning it can reason across different input data types, including text, audio, images and video, making it particularly well suited for analysing the diverse social media content in our data set. We chose Gemini 1.5 Pro (Reid *et al.* 2024) for this study due to its unique combination of multimodality processing and long-context understanding. Gemini 1.5 Pro delivers a breakthrough in long-context understanding, with the ability to process up to two million tokens consistently, achieving the longest context window of any large-scale foundation model to date. This expanded context window allows for more comprehensive processing of information, leading to more consistent, relevant and useful results. These features are particularly valuable for our research, as they allow us to effectively extract insights from the full range of information present in the data and for developing a more nuanced understanding of earthquake impacts.

Gemini, like many large language models, operates through prompting. This involves providing the pre-trained model with a natural language instruction or 'prompt' that guides its response generation without requiring further training or parameter updates. The effectiveness of a prompt is crucial in steering the model towards the desired output. A well-crafted prompt should exhibit clarity, conciseness and sufficient context, enabling the model to accurately comprehend and execute the task. Key elements include explicit task specification, relevant background information, natural language phrasing, illustrative examples and consistent coherence throughout the prompt. By meticulously designing prompts, users can fine-tune the model's behaviour and leverage its capabilities to produce precise and pertinent outputs tailored to specific needs and objectives.

Several prompting techniques exist, including zero-shot, one-shot, few-shot and multishot prompting (Brown *et al.* 2020; Kojima *et al.* 2022). Zero-shot prompting, also known as direct prompting, provides the model with only instructions and no examples. This approach is well suited for creative tasks and benchmarking. One-shot prompting offers the model a single concise and descriptive example to guide its output. Few-shot and multishot prompting, on the other hand, provide multiple examples, proving more effective for complex tasks requiring pattern replication or specific output structures that are difficult to describe explicitly. Our experiments demonstrate that Gemini can estimate ground shaking intensity based on the content of a social media post even through a simple zero-shot prompt such as: '`Use the video, audio and text in this social media post shared by a person who`

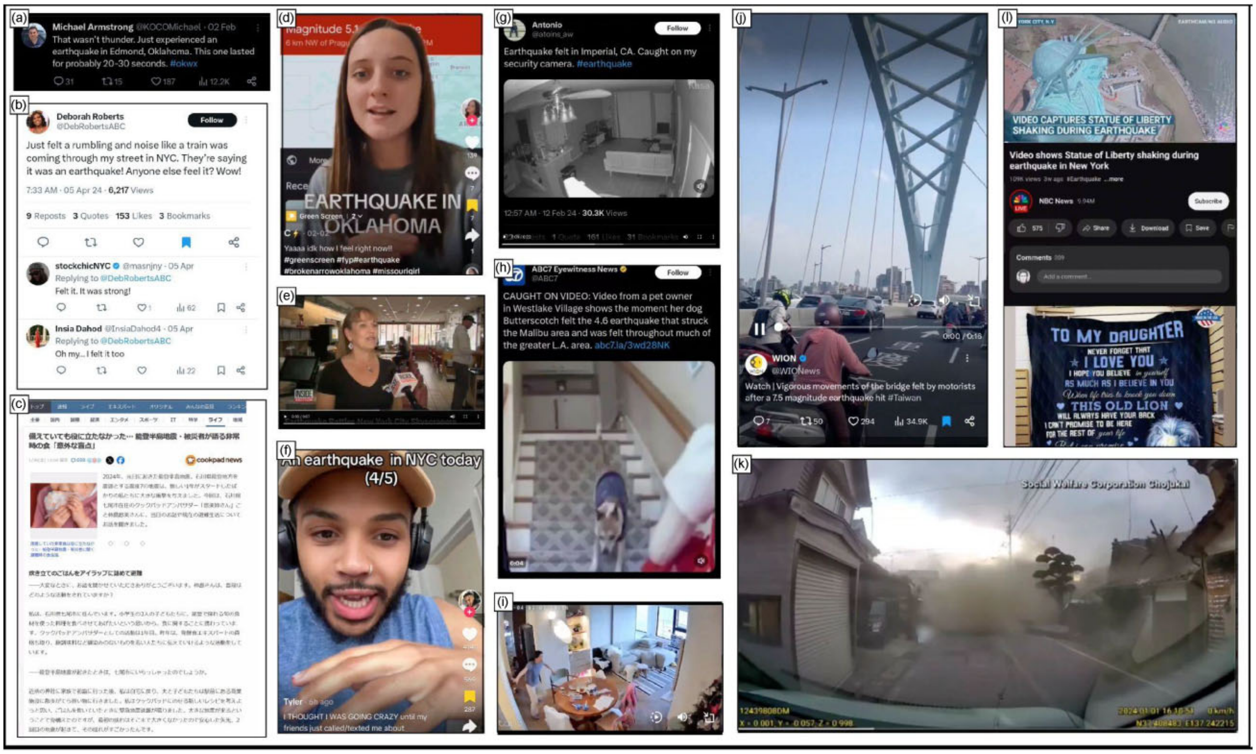

Figure 1. A composite of screenshots that show the diversity of the data set used in this study, which comprises screenshots and screen recordings of social media posts documenting individual (e.g. a, d, e and f) or group (e.g. b) experiences of earthquake shaking, as well as the responses of animals (e.g. h). These posts encompass a range of formats, including images (b) and videos (d to k) containing textual information, presented in various languages (e.g. c), sizes/durations, and background settings. The video content spans both indoor (g, h and i) and outdoor (j, k and l) environments. Indoor videos primarily consist of CCTV footage capturing the moment of earthquake shaking, while outdoor videos include similar CCTV footage as well as recordings of infrastructure damage. Additionally, the data set incorporates post-earthquake narrative videos where individuals describe their personal experiences and observations during the earthquake (e.g. d to f).

`felt an earthquake to estimate the intensity of ground shaking at its location in the MMI Scale.'` This capability stems from the inclusion of MMI scale definitions and information within its training data. However, for improved accuracy and consistency in these estimations, more sophisticated prompt engineering techniques are necessary (Wang *et al.* 2023).

The few-shot prompt employed in this study, detailed in Table S1 (Supporting Information), comprises four key components: background information (persona and context), instruction or query, desired output format and exemplar. The exemplar serves as a template, outlining the structure of the expected output. This structure includes: (1) summarizing pertinent information extracted from the input data (e.g. post time and location, potential earthquake association, shaking duration, building type, people's reactions; (2) organizing collected visual, auditory and textual observations and evidence; (3) providing an estimated MMI value; (4) explaining the reasoning behind the estimation process; and (5) acknowledging limitations that may impact the model's estimate. To facilitate complex reasoning, the prompt utilizes the chain-of-thought (CoT) prompting technique (Cobbe *et al.* 2021; Suzgun *et al.* 2022). CoT prompting encourages LLMs to break down problems into intermediate steps, mimicking human thought processes. By presenting the LLM with examples that explicitly demonstrate reasoning steps, it is encouraged to follow suit, leading to more accurate and transparent results.

To enhance the reliability and stability of responses, our pipeline incorporates contextual information such as the distance to the earthquake epicentre and the earthquake's magnitude (Fig. 2). Initially, we extract the post's location (city, state and country) from the input video/image using a one-shot prompt: `'Analyse the provided image/video and extract any information that indicates the location of the user or post. Identify the specific location name, including city, state (if applicable) and country. Limit your response only to the extracted location. Example output: Imperial, CA'.` Following this, we automatically calculate the distance from the earthquake epicentre. Both the earthquake magnitude and the computed epicentral distance are then integrated into the main prompt to query the full intensity analysis.

We conducted our analysis using Vertex AI, a Google Cloud machine learning (ML) platform that facilitates the training, deployment and customization of AI applications and ML models. Our pipeline leveraged the Gemini 1.5 Pro model with specific parameters to ensure consistency and control over the generated outputs. We set the temperature value to 0.5, which promotes more deterministic and probability-driven predictions, favouring the most likely tokens and reducing randomness. Additionally, we maintained the default output token limit of 8192, equivalent to approximately 32 768 characters, to accommodate the desired output length. Finally, we employed a top-p value of 0.95, which dictates the selection of tokens based on their cumulative probability distribution, ensuring diversity while maintaining focus on the most probable options.

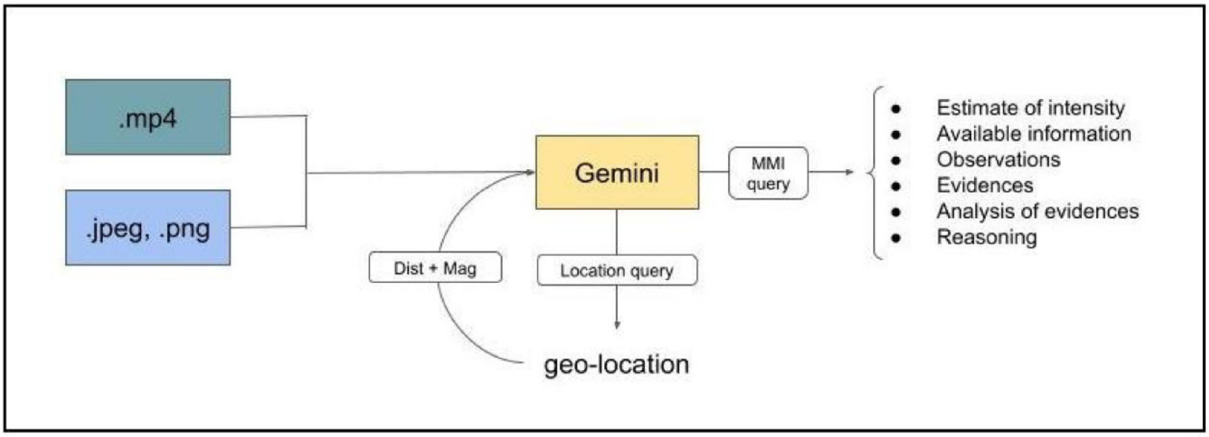

Figure 2. A schematic illustration of the processing workflow employed in this study. Initially, images and videos are input into the model. A basic query is then executed to extract available evidence pertaining to the location of observation. This evidence can manifest in textual, visual or auditory formats. The extracted location information is subsequently utilized to calculate the distance (great circle in kilometer) to the earthquake's epicentre. This distance, along with the earthquake's magnitude, serves as input for the next prompting stage in which a comprehensive intensity analysis is conducted.

4 RESULTS

We used Gemini 1.5 Pro (Reid *et al.* 2024) model and few-shot prompting (see the method section for more details), for processing and analysing the collected data. Table 2 showcases an example of Gemini's output, processing and analysing the contents of a social media post featuring a CCTV video during the 2024 M4.8 earthquake in New Jersey. The model effectively extracts key information such as the post's location, earthquake magnitude, shaking duration, building type and reactions of humans and animals. It then summarizes and analyses visual, auditory and textual cues (e.g. the sound of objects rattling) to estimate the shaking intensity on the MMI scale, providing a rationale for its estimation. The ability to perform cross-modal analysis and acknowledge limitations like incomplete views or unknown epicentral distances is noteworthy. However, it is crucial to recognize that LLMs like Gemini can generate hallucinations, leading to variations in estimated MMI values across different inferences. To address this and obtain more consistent results with uncertainty quantification, we calculate the mean, median and standard deviation of the estimated MMI values (represented as '`mid_mmi_value_numeric`') across ten inferences on each sample. Gemini's outputs for each post, along with the corresponding input data, are available in the supplementary materials of this paper.

To validate our findings, we compared them with independently determined intensities from the 'Did You Feel It?' (DYFI) data set (Atkinson & Wald 2007; Wald *et al.* 2011; Quitoriano & Wald 2020). This valuable USGS (United States Geological Survey) resource collects post-earthquake reports (through online questionnaires) from individuals who experienced the event, providing insights into ground shaking intensity and its geographic distribution. Participants share observations through the DYFI website or app, contributing to scientific understanding of earthquake effects. Despite inherent variability and uncertainties, DYFI data serves as a crucial 'ground truth' reference for validating Gemini's estimates. Fig. 3 illustrates the mean MMI values estimated by Gemini (with ± 1 standard deviation) for two well-documented earthquakes (New Jersey and Oklahoma) overlaid with USGS DYFI data, seismogram-based MMI estimations (Worden *et al.* 2010; Wald *et al.* 2022), and the region's expected attenuation model. Gemini's MMI estimates align with the expected intensity range derived from ground motion prediction models, felt reports and instrumental measurements at comparable distances, supporting the validity of our approach. The highest computed uncertainties in estimated MMI valused is related to a CCTV footage recorded in High Bridge, NJ, ~11.92 km from the epicentre of M4.8 Tewksbury earthquake. This video, with a limited view to the sky and a few trees in an open area, offers limited evidence of shaking intensity such as the tree swaying and lacks other strong audio and textual supports for a more deterministic intensity estimation. Furthermore, Fig. 4 demonstrates that Gemini's mean MMI values for these earthquakes fall within a similar intensity-distance distribution as the DYFI data. Both events, classified as moderate with reported intensities primarily between III and IV, exhibit clustering around population centres like New York City and Tulsa.

While we used DYFI data as the ground truth for this study due to its widespread use and accessibility, it is important to acknowledge its limitations. The original MMI scale was designed to reliably distinguish between shaking levels that differ by roughly a factor of two. However, DYFI intensities may not accurately reflect MMI values as originally defined.

The low-range MMIs (i.e. III to IV) often form the majority of felt reports collected for the moderate earthquakes. To assess the model's ability to generalize to other regions and perceive higher intensity levels, we analysed data from larger events in diverse locations. However, due to a limited number of available posts, we restricted our analysis to comparing Gemini's estimates with the distributions of DYFI data reported for the city where most of our post data originated (Fig. 5). The results demonstrate that Gemini's estimated MMIs generally fall within the interquartile ranges of DYFI reports. Notably, Gemini's intensity estimates for major events (M7.4, 7.5 and 9.1) are consistently and significantly higher than those for moderate events (M4.6, 4.8 and 5.1). The most significant discrepancy between our estimates and DYFI data is observed for the M7.4 Taiwan event. This could potentially be attributed to high uncertainties in estimating content source locations and the sparsity of DYFI data for the region.

In practice, it is unusual to assign intensities to single observations (Grunthal *et al.* 1998). Some systems, like DYFI, calculate

Table 2. An example of Gemini's output for a social media post. This post documents the ground shaking experienced in Boonton, NJ, located approximately 37.6 km from the epicentre of the M4.8 earthquake that occurred on 2024 April 5, in Tewksbury, New Jersey, USA.

{ 'felt_by_user': 'Yes',
'class_source': 'Surveillance Footage (YouTube)',
'post_location': 'Boonton, NJ (based on the news ticker)',
'post_date_time': 'April 5, 2024 (exact time not specified)',
'earthquake_location': 'North-east U.S. (N18V)',
'earthquake_magnitude': '4.8 (provided in video description)',
'distance_to_earthquake_epicentre': '37.59 km (provided by user)',
'shaking_duration': 'Several seconds, as observed in the video footage',
'building_type': 'Coffee shop (commercial building)',
'building_materials': 'Unknown',
'human_reaction': 'Panic and fear, evident from their hurried movements, seeking cover under tables, and expressions of distress.',
'animals_reaction': 'Not visible in the video',
'furnishing': 'Tables, chairs, light fixtures, and other typical coffee shop items',
'language': 'English',
'natural_environment': 'Not visible in the video',
'video_evidence': {
'Object Movement': 'Visible shaking of light fixtures and furniture, indicating ground movement.',
'People Running and Seeking Cover': 'Individuals exhibiting panicked behaviour, rushing for exits, and taking cover under tables, demonstrating a perception of danger.',
'Limited Object Falling': 'Some small objects may appear to be falling or displaced, but the video quality makes it difficult to ascertain the extent.'},
'audio_evidence': {
'Sounds of Distress': 'People expressing fear and concern, with exclamations like 'Earthquake!' and 'Mommy!'',
'Background Noise': 'General coffee shop ambiance and potential sounds of objects rattling or moving.'},
'textual_evidence': {
'Video Description': 'Provides information about the earthquake's magnitude and location.',
'News Overlay': 'Mentions tremors lasting several seconds and being felt 200 miles away, suggesting a wider impact of the earthquake.'},
'mmi_estimation': 'MMI IV or MMI V',
'mid_mmi_value_numeric': '4.5″',
'estimation_confidence': '0.7″,
'visual_observation': 'The video evidence suggests moderate ground shaking, with notable movement of objects and people reacting with fear and taking cover. The absence of significant damage or widespread falling objects indicates that the intensity is likely not higher than MMI V.',
'auditory_cues': 'The audio evidence supports the visual observations, with people's vocal expressions of fear and distress indicating a level of alarm consistent with moderate shaking.',
'textual_information': 'The video description and news overlay provide valuable context about the earthquake's magnitude and its broader impact, reinforcing the assessment of moderate shaking intensity.',
'analysis_of_evidences': 'The combined evidence from visual, auditory and textual sources points to an earthquake intensity in the range of MMI IV to MMI V at the coffee shop location. The observed effects align with the characteristics of these MMI levels, where objects move notably, people feel frightened and react, but significant damage is not widespread.',
'reasoning': 'The observed effects, such as the shaking of light fixtures and furniture, people's reactions, and the lack of major damage, are consistent with the descriptions of MMI IV and V. The distance from the epicentre and the earthquake's magnitude also support this estimation.'}

a Community Decimal Intensity by spatially averaging answers to questions about earthquake effects (Wald *et al.* 2024). Others average the intensity values themselves or use fuzzy logic to determine representative answers. A simple way to aggregate intensity estimations in our proposed approach is to analyse colocated posts from multiple users simultaneously. Gemini's ability to handle large context ranges allows us to do this. We can analyse multiple posts at once and estimate intensity based on the combined evidence. To test this, we processed up to 15 files (images and videos) or 66 000 tokens, simultaneously for each location. The results, shown as diamonds in Fig. 5(a), demonstrate that processing multiple posts leads to reasonable results. While this significantly increased the inference time (up to a few minutes), we observed that the intensity estimates are more consistent across different inferences. This suggests that multiple user inputs help to better constrain the MMI estimates. An added benefit of this aggregation method is that the results can be shared publicly without revealing personally identifiable information or raising other privacy concerns.

Fig. 6 presents a comparison of estimation uncertainty (through bootstrapping) in Gemini's results for CCTV footage versus social media posts and news interviews. Our estimated model uncertainty indicates a higher variability in model's estimates derived from CCTV footage relative to those derived from social-media posts (Fig. 6a). In contrast, Fig. 6(b) suggests that the model often rates its estimates from CCTV footage with relatively higher confidence values.

One possible explanation for the model's higher confidence in CCTV footage could be the richer set of information typically available in videos, compared to the often limited evidence found in social media posts, such as short tweets. However, this could result in a higher variability in the model's estimates as well, potentially, due to a wider range of observations/evidence available for the analysis and reasoning. We did not observe a strong correlation between the estimated uncertainties and the confidence values provided by Gemini. Interestingly, the narrative descriptions of shaking experiences shared by users in social media posts and interviews, although less direct than the captured shaking moments in CCTV footage and reducing the confidence of the model on its estimate, appear to constrain the model's estimates, resulting in lower overall uncertainties compared to estimates from CCTV footage. However, the less deterministic nature of descriptive evidence leads to a wider range of uncertainties compared to the more visually

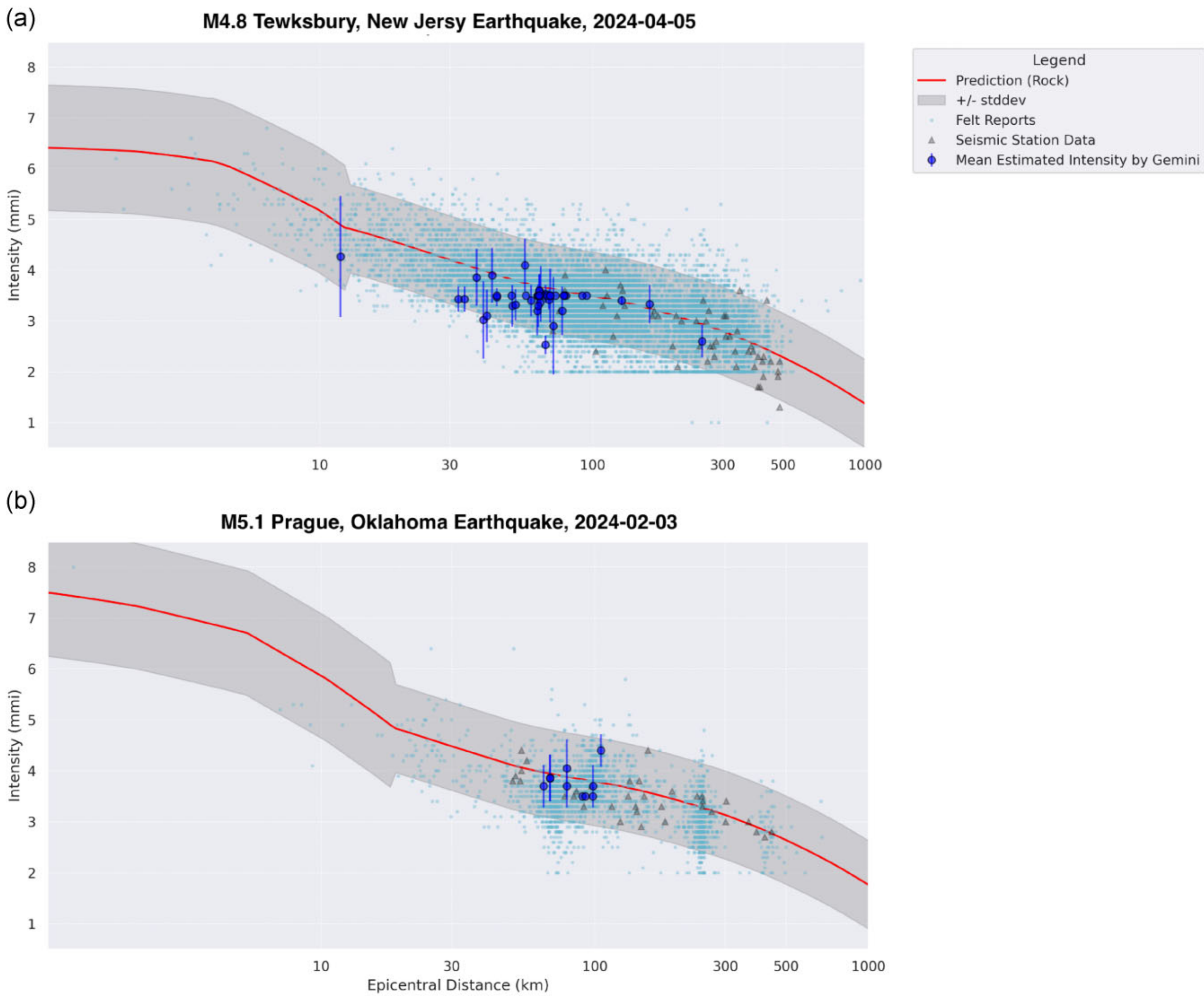

Figure 3. Scatter plots presenting a comparison of estimated MMIs from the Gemini model (circles with error bars) against several sources of observed data. These include: (1) instrumentally derived MMIs computed from peak ground acceleration recorded by seismic stations; (2) DYFI macroseismic data collected by the USGS; and (3) the expected ground motion attenuation model for rock sites on the East Coast, along with its ± 1 standard deviation range. Panels (a) and (b) display the results for the New Jersey and Oklahoma earthquakes, respectively. Note that the distance scale on the *x*-axis is logarithmic.

explicit evidence in videos. Factors such as limited views, low-quality audio, and the lack of informative contextual information in some CCTV footage could contribute to the higher uncertainties observed in MMI estimations for this data type. Further investigation is warranted to better understand the observed differences in uncertainty between data types and to explore methods for improving the model's performance and confidence across all sources of information. This could involve analysing the specific types of descriptive evidence that contribute to lower uncertainties, as well as developing techniques to extract more contextual information from CCTV footage.

5 DISCUSSION

5.1 How does gemini estimate MMI?

In addition to reading and listening, is Gemini capable of understanding and applying concepts in physics? Our results suggest that Gemini can estimate ground motion intensity from social media posts and CCTV videos with a comparable variability to the estimates based on conventional felt reports and instrumental measurements. This capability stems from Gemini's advanced understanding of language, images, video and audio. While the model effectively retains detailed and relevant information from the input data in most cases, its capabilities, similar to humans, are not without limitations and can be prone to errors due to factors such as low-quality inputs, limited evidence, and the presence of noise. For example, misinterpretations of visual cues or ambiguous language in social media posts can lead to inaccurate estimations. Beyond its information extraction and understanding abilities, Gemini appears to utilize additional sources of knowledge about general magnitude–distance–intensity relationships of earthquakes, presumably acquired during its training, in its reasoning and decision-making processes. Visual inspection of the model's outputs reveals frequent references to epicentral distance, earthquake magnitude, and post comments in the reasoning section, suggesting that Gemini actively incorporates this contextual information,

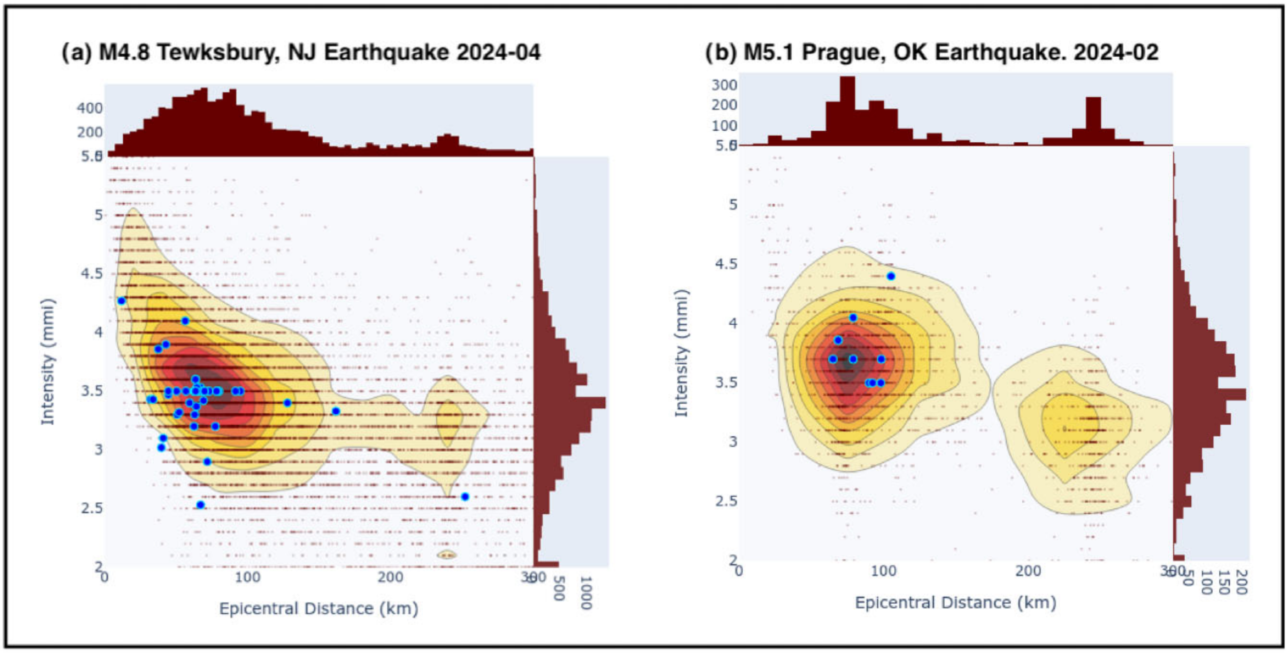

Figure 4. Histograms illustrating the distribution of earthquake intensity with respect to epicentral distance for the New Jersey (a) and Oklahoma (b) earthquakes, utilizing USGS DYFI data. Each panel presents a 2-D histogram (represented by colour coded contours), while its margins display 1-D histograms with counts on the axes. The circle markers represent the estimated mean MMI values for individual social media posts, as determined by the Gemini model.

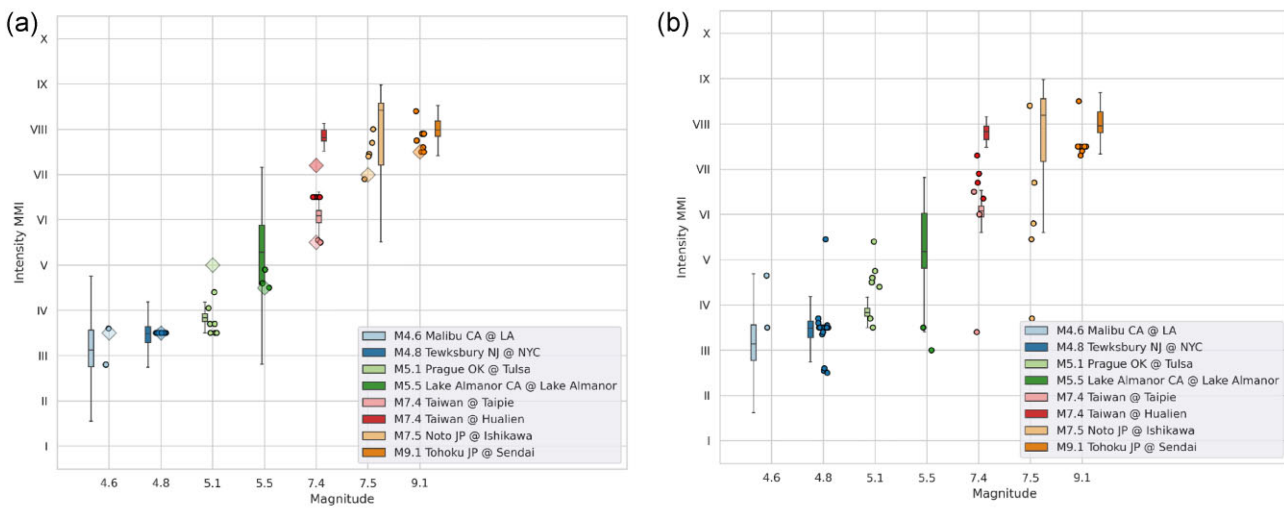

Figure 5. Comparison of Gemini's estimate with (a) and without (b) epicentral distance and earthquake magnitude in the prompt. Each boxplot with whiskers illustrates the distribution, quartiles and outliers of reported DYFI data for individual earthquakes within specific zip codes of a city. Circle markers represent the estimated mean MMI values derived from each social media post, within the same city, analysed by the Gemini model for each event. Diamond markers show the Gemini's estimates based on simultaneous processing of all posts for one location.

whether provided directly in the input data or through the prompt, during its analysis. To further investigate the influence of provided context on the output results, we conducted two experiments. In the first experiment, we selected two sample inputs: a CCTV video recorded approximately 5 km from the M7.5 Noto earthquake and a tweet from about 68 km away from the Prague earthquake. In the first experiment, multiple queries were performed on each input sample, varying only the value of the epicentral distance in the prompt while keeping the magnitude fixed at its true value. For both cases, Gemini's estimated MMI value systematically decreased as larger epicentral distances were used in the prompt, aligning with established principles of earthquake physics and empirical relationships in seismology (Fig. 7, top). Similar observations were made when varying the magnitude, the second experiment, while keeping the distance fixed (Fig. 7, bottom). These results indicate that Gemini might have a knowledge of relationships between earthquakes magnitude, distance to the epicentre, and ground shaking intensity and actively use it in its predictions. However, our observations, from performing a similar test to multiple other examples, suggest that Gemini uses this knowledge in conjunction with the evidence it extracts from the provided data. This is why its upper and lower estimation bounds, that is MMI 8.5 and 2 for the Video example

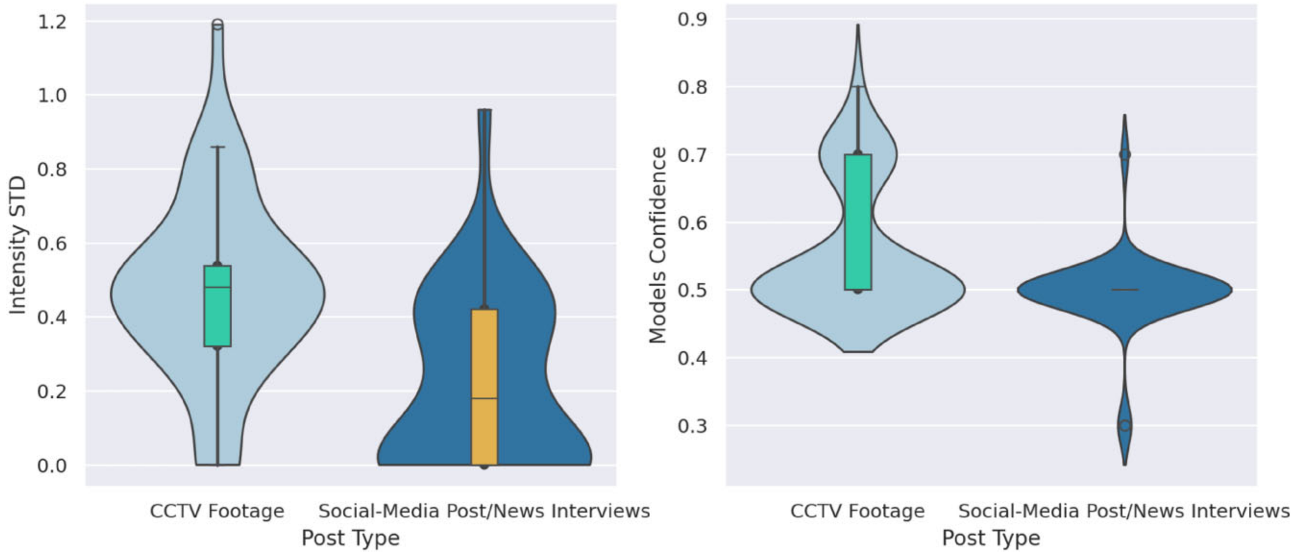

Figure 6. Comparison of model uncertainties (a) (variations in Gemini's estimates across multiple inferences) and the model confidence (provided by Gemini directly) (b) in intensity estimation from various types of data sources (like CCTV footage and social-media posts/news interviews). The box plots show the median, quartiles, and outliers of the confidence scores for each category, while the violin plots reveal the shape and density of the distributions, highlighting where most scores are concentrated and how they spread out.

and MMI 5.5 and 2 for the Tweet example, respectively, remain within the range of viable intensity levels based on the observations and evidence within the input data. As an example, even when the prompt states that provided Tweet was posted by a person 1000 km away from a M 0.5 earthquake, a case where all the empirical relationships predict MMI I or not felt intensity, Gemini still outputs the minimum felt intensity level (MMI II) as it can not ignore the provided evidence by the user who clearly mentioned feeling of an earthquake.

These findings raise intriguing questions about the extent of Gemini's knowledge and reasoning capabilities. To further investigate this, we simply asked Gemini for its reference and if it uses any particular ground motion prediction equation. Here is its response: `'I did not use a specific ground motion prediction equation (GMPE) to arrive at the MMI estimates. My response was based on a simplified understanding of the general relationship between magnitude, distance and MMI intensity. This approach does not involve the complexity of GMPEs, which consider various factors like fault type, regional geology and site conditions to predict ground motion parameters.'` To get a feeling for this 'simplified understanding of the general relationship' that Gemini has learned about earthquakes, we performed a large number of queries asking for a general estimate of MMI level at different epicentral distance, earthquake magnitude and earthquake depth ranges without providing any input data, exemplar, and additional contextual information (Fig. 8). The results clearly demonstrate an understanding of the relationship between earthquake characteristics and shaking intensity. It is important to consider that this knowledge likely stems from the vast amount of data it was trained on, rather than an inherent understanding of physics principles. However, more studies, perhaps on more well-defined problems with less uncertain ground truths and more available data, are needed to further investigate the extent of Gemini's general understanding of the physical world and its phenomena. Nevertheless, the ability to incorporate contextual information and generate results consistent with established scientific knowledge highlights the potential of LLMs like Gemini in augmenting our understanding of complex physical phenomena such as earthquakes.

While the MMI scale is widely used, it does not uniquely define earthquake intensity. Though primarily trained on data referencing the MMI scale, the Gemini model may be capable of estimating intensity on other scales as well. To investigate this, we conducted a preliminary experiment where Gemini estimated intensities on the European Macroseismic Scale (EMS) and the Japan Meteorological Agency (JMA) scale for three videos, each depicting an earthquake of a different magnitude (Table 3). With the exception of the JMA estimate for the M 7.5 event, the estimated intensity values across the different scales largely align with the known relationships between these scales (Musson *et al.* 2010). However, further testing with a larger data set is necessary to thoroughly evaluate Gemini's ability to estimate earthquake intensity across various scales.

5.2 Potential applications

The results of our study indicate LLMs hold immense potential for deriving macroseismic intensity values from crowdsourced eyewitness accounts. Recently, macroseismic data collection relied on voluntary participation in structured surveys like 'Did You Feel It?'. This presents a fantastic opportunity for public science participation. DYFI already facilitates community involvement, fosters a sense of connection between users and their government, and provides a degree of catharsis (Wald *et al.* 2011). Multimodal LLMs like Gemini offer the potential for even broader engagement by providing more flexible ways to contribute. For example, this system could be integrated with current online questionnaire systems to automatically extract relevant information from supplementary materials like images or videos provided by volunteers. Furthermore, it can overcome language barriers and enable intensity estimation from free-form descriptive reports (instead of relying solely on structured questionnaires), which may offer richer and more detailed information. Such a system could also be used to validate the credibility of reports.

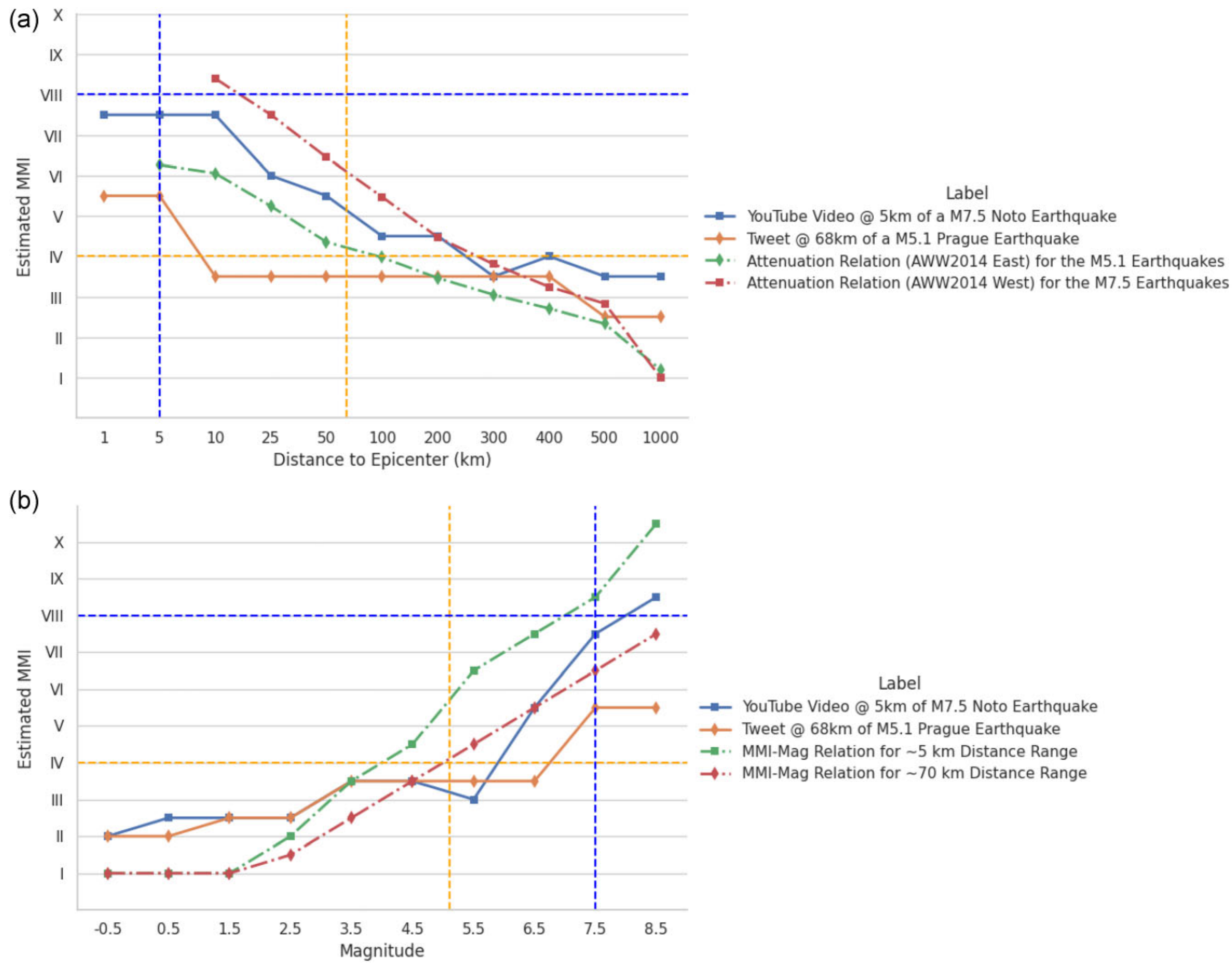

Figure 7. Estimated MMI values by Gemini for two sample inputs (a CCTV video recorded at 5 km of M7.5 Noto earthquake and a tweet ∼68 km of M5.1 Prague earthquake) but varying provided information on epicentral distance (top) and earthquake magnitude (bottom) in the prompt. The true distance, magnitude and MMI for each event is depicted by vertical and horizontal lines, respectively. The dotted dashed lines are the expected MMI values as a function of distance and magnitude from known empirical relationships.

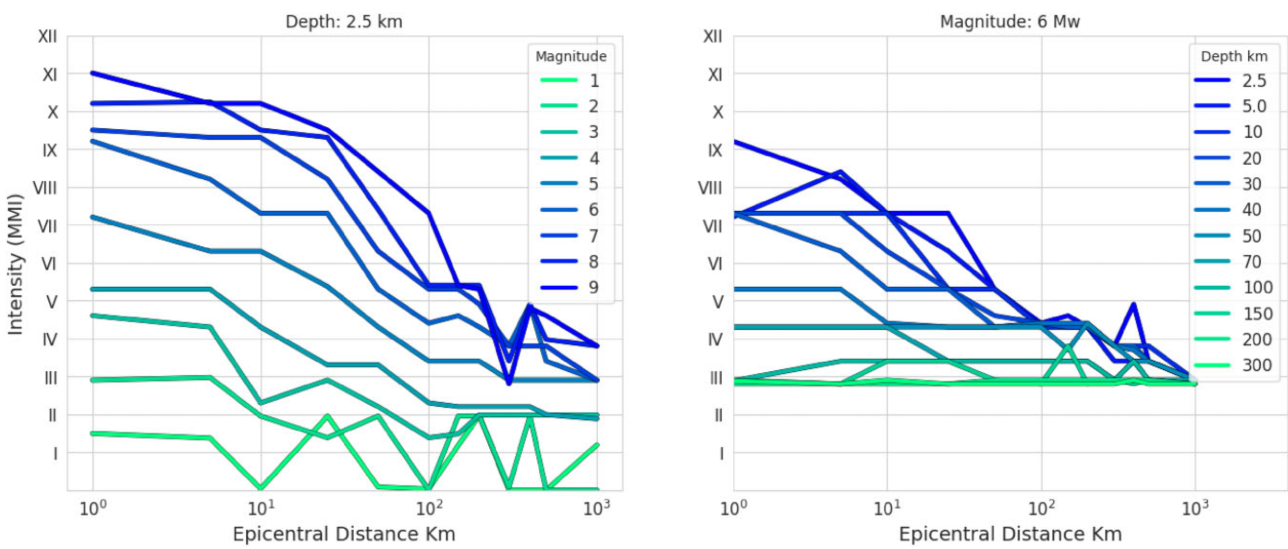

Figure 8. Gemini's estimates of MMI level at different epicentral distance ranges for varying earthquake magnitudes (left) and depths (right).

Table 3. Comparison of estimated intensity by Gemini for three sample videos associated with three different earthquakes. MMI, EMS (Gru¨nthal, 1998) and JMA (1996).

Event magnitude	MMI	EMS	JMA
M 9.1	7	7	5+
M 7.5	6	6	4
M 4.8	4	4	2

The ability to generate scientific measures like earthquake shaking intensity directly from unstructured data such as CCTV footage and social media posts opens exciting new avenues in citizen seismology (Bossu *et al.* 2011) and deep-learning seismology (Mousavi & Beroza 2022). This makes collecting valuable scientific data from unconventional resources such as social media posts, open comments and geo-located pictures and videos more feasible, enriching our understanding of seismic events and their impact. This can be a valuable source of information to constrain and improve these estimates of often uncertain damage scenario estimates solely based on earthquake magnitude and location, particularly in sparsely instrumented regions. LLMs offer efficient and effective methods for real-time processing and analysis of crowdsourced data during natural disasters. This can contribute to a better understanding of their impact on communities by providing real-time crisis mapping and rapid situation awareness (Middleton 2013). The extracted intensities (along with improved geolocation data) can be used for various purposes such as: improving the alert models in EEW systems, rapid impact assessment (Vieweg *et al.* 2010; Kryvasheyeu *et al.* 2016; Bossu *et al.* 2024), rapid determination of fault geometry (Böse *et al.* 2021), reducing uncertainties in ShakeMaps (Quitoriano & Wald 2022), improving human loss estimations (e.g. USGS PAGER system), distinguishing high-impact from low-impact events (e.g. Lilienkamp *et al.* 2023), providing situational awareness (Yin *et al.* 2012), protective action decision-making (Zhang *et al.* 2024) and assessing information credibility (Castillo *et al.* 2013).

5.3 Current challenges and potential solutions

Currently, a major challenge/limitation lies in the lack of reliable (and precise enough) location information which could enable a more rigorous testing and evaluation against ground truths. Often, CCTV footage of earthquake shaking goes viral quickly, making it difficult to identify the original source to verify the location of the observation. To fully leverage such systems for crowdsourcing earthquake information, supplementary geoparsing or geotagging techniques can be employed (Middleton *et al.* 2018; Huang & Carley 2019). Similarly, the use of official API might offer a more precise location estimation in some cases. Beside location estimation, to improve the intensity estimations, more sophisticated prompt engineering techniques and a variety of exemplars can be employed. Furthermore, the LLM models can be fine tuned using labelled data sets like DYFI. Additionally, incorporating supplementary seismological data, such as recorded ground acceleration, earthquake depths and historical earthquake evidence at that location into the analysis process might lead to further improvements (e.g. Burks *et al.* 2014). This is where the multimodality of LLMs like Gemini becomes particularly useful. Finally, multi-LLM based intelligent agents could be developed to automate the entire process (e.g. Boiko *et al.* 2023).

6 LIMITATIONS AND POTENTIAL RISKS

While LLMs show promise for earthquake and natural disaster research, offering new avenues for autonomous scientific inquiry, it is crucial to acknowledge their limitations and potential risks (Bommasani *et al.* 2021). One challenge with using crowdsourced data is that it is often insufficient for accurately assigning higher intensities (above VI, or damaging levels). This is because people are less likely to submit detailed reports during stronger earthquakes due to factors like safety concerns and disruptions to communication networks. Wald *et al.* (2024) highlight this challenge and emphasize the need to explore alternative approaches for estimating higher intensities. They also discuss the broader limitations of using MMI for higher intensities, pointing out that EMS-98 and international macroseismic scale offer more quantitative measures for assessing earthquake impacts. This suggests that incorporating data from these more quantitative scales could improve the accuracy of intensity estimations, especially for stronger earthquakes.

6.1 Prompt engineering and its effects

Like all LLMs, Gemini's output is highly sensitive to the input prompt. Even minor changes can lead to different MMI estimations. For example, adopting different personas in the prompt, such as '`earthquake engineer`' or '`disaster management professional`,' can yield to different sets of outputs each tailored to specific professional needs and interests. The high variability of inference outputs and inconsistency of estimates in some cases pose the other main issue. Several techniques can improve the reliability and consistency of LLM outputs. These include calibrating output probabilities (Zhao *et al.* 2021), using a noisy channel (Min *et al.* 2021), augmenting few-shot examples with intermediate steps (Reynolds & McDonell 2021), or employing the Graph of Thoughts (GoT) technique (Besta *et al.* 2024) for multistep reasoning. Furthermore, mining and paraphrasing methods can automatically augment prompt sets (Jiang *et al.* 2020). Deeper exploration of successful prompting strategies (Xie *et al.* 2021) may reveal how to elicit emergent abilities of the models. However, understanding why models work often lags behind the development and popularization of techniques like few-shot prompting. Best practices for prompting are also likely to evolve as more powerful models emerge.

6.2 Summation effect

Gemini 1.5 Pro boasts a remarkable ability to understand information within a long context, allowing it to process lengthy videos containing multiple reports. We investigated the impact of this capability on model output by analysing a news report featuring interviews with three individuals who experienced the 2024 April M4.8 earthquake in New Jersey. The report was processed in two ways: first, as a single video containing all three interviews, and second, as three separate videos, each trimmed to include only one interview. The estimated MMI for the long video encompassing all three interviews was 4.5, which matched the highest MMI estimate obtained from the individual interview segments (MMI 3.5, 3.5 and 4.5, respectively). When the original video was shortened to include only the first two interviews, the estimated MMI dropped to 3.5. This suggests that the content of the last interview influenced the overall MMI estimation.

6.3 Multilingual capabilities and limitations

A significant advantage of Gemini is its ability to process content in multiple languages directly. Our data set incorporates posts and dialogues written in Chinese and Japanese, which the model successfully comprehended and analysed. However, it is important to note that encountering unsupported languages can lead to inaccurate location estimations (for both the post and the earthquake event) as the model is still able to distinguish the language and wrongly associated with the earthquake vehicle not fully comprehending its content. An interesting observation was that even in such a case, the estimated MMI value would still remain unaffected as it relies on evidence the model can extract and confidently relate to the earthquake shaking.

6.4 Irrelevant information and estimation accuracy

Across several experiments, we examined the potential impact of extraneous information, such as advertisements or user comments, on the accuracy of intensity estimations. Our findings revealed no notable difference in model performance between posts containing such distractions and those without. This suggests that the models effectively filter irrelevant content, focusing primarily on the pertinent information for estimation. Consequently, the presence of extraneous elements appears to have negligible influence on the models' accuracy.

7 CONCLUSION

This research demonstrates the promising capabilities of large language models like Google's Gemini in estimating earthquake shaking intensity from multimodal social media data. Gemini's ability to analyse text, images and videos, coupled with its long-context understanding and reasoning capabilities, allows for a nuanced interpretation of earthquake experiences and accurate MMI estimations. This innovative approach offers several advantages over traditional methods, including:

(i) Rapid and flexible data collection and processing: LLMs can efficiently process unstructured data from diverse sources, bypassing the limitations of structured surveys and questionnaires.

(ii) Overcoming language barriers: Gemini's multilingual capabilities enable analysis of posts in various languages, expanding the reach and inclusivity of data collection.

(iii) Real-time insights: LLMs facilitate real-time analysis of social media data, providing valuable situational awareness and supporting rapid response efforts during earthquakes and other natural disasters.

(iv) Improved impact assessment: MMI estimations derived from social media data can contribute to more accurate and comprehensive assessments of earthquake impacts, aiding in disaster response, recovery and mitigation efforts.

While LLMs show immense potential for citizen seismology, further research is needed to address challenges such as location accuracy, prompt engineering and potential biases. Integrating supplementary data sources and exploring advanced prompting techniques can further enhance the reliability and consistency of LLM-based MMI estimations. As LLM technology continues to evolve, its application in earthquake research and disaster response holds significant promise for building more resilient communities in earthquake-prone regions.

However, it is important to acknowledge the potential limitations of AI models and the need for human oversight to ensure accuracy and mitigate biases in training data. Additionally, ethical considerations regarding privacy and data ownership must be carefully addressed when utilizing social media information. Despite these challenges, this innovative approach holds great promise for leveraging the vast potential of social media in disaster response and advancing our understanding of the societal impacts of natural hazards.

ACKNOWLEDGMENTS

We would like to thank Dr David Wald, Prof. Duncan Agnew and Dr Susan Hough for insightful remarks during the review process. Some of the python codes for plotting the results of this study have been generated by Gemini. Figures in this paper were generated using the Seaborn (Waskom 2021) and Matplotlib (Hunter 2007). SMM was supported by the Harvard University William F. Milton Fund.

SUPPORTING INFORMATION

Supplementary data are available at *GJIRAS* online.

Table S1. The structure of the designed and used prompt in this study.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.

DATA AVAILABILITY

The data used in this study were collected from public accounts on YouTube, Twitter and TikTok. The data set and Gemini's results are available in the Dryad Digital Repository, at https://dx.doi.org/10.5061/dryad.rfj6q57kz. Maximum intensity 'Did You Feel It?' data for each event were collected from USGS website, https://earthquake.usgs.gov (last accessed May 2024).

REFERENCES

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L. & McGrew, B., 2023. Gpt-4 technical report. arXive: (arXiv:2303.08774).

Allen, R.M. & Melgar, D., 2019. Earthquake early warning: Advances, scientific challenges, and societal needs, *Annual Review of Earth and Planet. Sci.,* **47**(1), 361–388.

Allen, T.I., Wald, D.J. & Worden, C.B., 2012. Intensity attenuation for active crustal regions, *J. Seismol.,* **16**(3), 409–433.

Ambraseys, N.N. & Bilham, R., 2003. Re-evaluated intensities for the great Assam earthquake of 12 June 1897, Shillong, India, *Bull. seism. Soc. Am.,* **93**(2), 655–673.

Atkinson, G.M. & Wald, D.J., 2007. "Did you feel it?" intensity data: a surprisingly good measure of earthquake ground motion, *Seismol. Res. Lett.,* **78**(3), 362–368.

Bagrow, J.P., Wang, D. & Barabasi, A.L., 2011. Collective response of human populations to large-scale emergencies, *PLoS One,* **6**(3), e17680. doi:10.1371/journal.pone.0017680.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L. & Hoefler, T., 2024. Graph of thoughts: solving elaborate problems with large language models, *Proc. AAAI Conf. Artif. Intell.,* **38,** (16), 17682–17690.

Boiko, D.A., MacKnight, R. & Gomes, G., 2023. Emergent autonomous scientific research capabilities of large language models, preprint (arXiv:2304.05332).

Bommasani, R. *et al.*, 2021. On the opportunities and risks of foundation models, preprint (arXiv:2108.07258).

Böse, M. *et al.*, 2023. The European seismic hazard model: current State and outlook, *Bull. Earthq. Eng.,* **21**(1), 1–36.

Bo¨se, M., Julien-Laferriere, S., Bossu, R. & Massin, F., 2021. Near real-time earthquake line-source models derived from felt reports, *Seismol. Res. Lett.,* **92**(3), 1961–1978.

Bossu, R., Böse, M., Steed, R. & Wald, D.J., 2024. The potential of crowd-sourced data for the rapid impact assessment of large earthquakes: the 2023 M 7.8 Kahramanmaraş-Pazarcık, Türkiye, earthquake, *Seismol. Res. Lett.,* **95**(4), 2058–2070.

Bossu, R., Gilles, S., Mazet-Roux, G., Roussel, F., Miller, D.S. & Rivera, J.D., 2011. Citizen seismology: how to involve the public in earthquake response, *Comparative Emergency Management: Examining Global and Regional Responses to Disasters,* Routledge, Milton Park, pp. 237–260.

Bossu, R., Roussel, F., Fallou, L., Landès, M., Steed, R., Mazet-Roux, G. & Petersen, L., 2018. LastQuake: from rapid information to global seismic risk reduction, *Int. J. Disaster Risk Reduct.,* **28,** 32–42.

Brown, T. *et al.*, 2020. Language models are few-shot learners, *Adv. Neural Inf. Process. Syst.,* **33,** 1877–1901.

Burks, L., Miller, M. & Zadeh, R., 2014. Rapid estimate of ground shaking intensity by combining simple earthquake characteristics with tweets, In *10th US National Conf. Earthquake Engineering, Frontiers of Earthquake Engineering,* 10NCEE, Anchorage, AK, USA.

Castillo, C., Mendoza, M. & Poblete, B., 2013. Predicting information credibility in time-sensitive social media, *Internet Res.,* **23**(5), 560–588.

Cobbe, K. *et al.*, 2021. Training verifiers to solve math word problems, preprint (arXiv:2110.14168).

Cresci, S., La Polla, M., Marchetti, A., Meletti, C. & Tesconi, M., 2014. *Towards a Timely Prediction of Earthquake Intensity with Social media*. IIT TR-12/2014. Technical report. IIT: Istituto di Informatica e Telematica, CNR.

D'Auria, L. & Convertito, V., 2016. Real-time mapping of earthquake perception areas in the Italian region from Twitter streams analysis, *Earthquakes and Their Impact on Society,* Springer-Verlag, pp. 619–630.

Davison, C., 1900. Scales of seismic intensity, *Philos. Mag.,* **50,** 44–53

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K., 2018. Bert: pre-training of deep bidirectional transformers for language understanding, preprint (arXiv:1810.04805).

Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S. & Vaughan, A., 2010. OMG earthquake! can Twitter improve earthquake response?, *Seismol. Res. Lett.,* **81**(2), 246–251.

Earle, P.S., Bowden, D. & Guy, M., 2011. Twitter earthquake detection: earthquake monitoring in a social world, *Ann. Geophys.,* **54**(6), 708–715.

Flores, J.M., Guzman, J. & Poblete, B., 2017. A lightweight and real-time worldwide earthquake detection and monitoring system based on citizen sensors, *Proc. AAAI Conf. Human Comput. Crowdsourcing,* **5,** 137–146

Grunthal, G., Musson, R.M.W., Schwarz, J. & Stucchi, M., 1998. European Macroseismic Scale 1998 EMS-98: luxembourg, Cahiers du Centre Européen de Géodynamique et de Séismologie, Vol. **15,** pp. 99

Hanks, T.C. & Johnston, A.C., 1992. Common features of the excitation and propagation of strong ground motion for North American earthquakes, *Bull. seism. Soc. Am.,* **82**(1), 1–23.

Hough, S.E., Armbruster, J.G., Seeber, L. & Hough, J.F., 2000. On the modified Mercalli intensities and magnitudes of the 1811–1812 New Madrid, central United States earthquakes, *J. Geophys. Res.,* **105**(B10), 23839–23864.

Huang, B. & Carley, K.M., 2019. A hierarchical location prediction neural network for twitter user geolocation (2019), preprint (arXiv:1910.12941).

Hughes, A.L. & Palen, L., 2009. Twitter adoption and use in mass convergence and emergency events, *Int. J. Emergency Manage.,* **6**(3-4), 248–260.

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

Japanese Meteorological Agency, 1996. Explanation table of JMA seismic intensity scale, https://www.jma.go.jp/jma/en/Activities/inttable.html (accessed June 2024).

Jiang, Z., Xu, F.F., Araki, J. & Neubig, G., 2020. How can we know what language models know?, *Trans. Assoc. Comput. Linguist.,* **8,** 423–438.

John, Hunter, 2007. Matplotlib: a 2D graphics environment, *Computing in science & engineering,* **9**(3), 90–95. doi:10.1109/MCSE.2007.55

Johnston, A.C., 1996. Seismic moment assessment of earthquakes in stable continental regions—I. Instrumental seismicity, *Geophys. J. Int.,* **124**(2), 381–414.

Kojima, T., Gu, S.S., Reid, M., Matsuo, Y. & Iwasawa, Y., 2022. Large language models are zero-shot reasoners, *Adv. Neural Inf. Process. Syst.,* **35,** 22199–22213.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J. & Cebrian, M., 2016. Rapid assessment of disaster damage using social media activity, *Sci. Adv.,* **2**(3), e1500779. doi:10.1126/sciadv.1500779

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. & Zettlemoyer, L., 2019. Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint (arXiv:1910.13461).

Lilienkamp, H., Bossu, R., Cotton, F., Finazzi, F., Lande`s, M., Weatherill, G. & von Specht, S., 2023. Utilization of crowdsourced felt reports to distinguish high-impact from low-impact earthquakes globally within minutes of an event, *Seism. Record,* **3**(1), 29–36.

Mendoza, M., Poblete, B. & Valderrama, I., 2019. Nowcasting earthquake damages with Twitter, *EPJ Data Sci.,* **8**(1), 1–23.

Middleton, S.E., Kordopatis-Zilos, G., Papadopoulos, S. & Kompatsiaris, Y., 2018. Location extraction from social media: geoparsing, location disambiguation, and geotagging, *ACM Trans. Inf. Syst. (TOIS),* **36**(4), 1–27.

Middleton, S.E., Middleton, L. & Modafferi, S., 2013. Real-time crisis mapping of natural disasters using social media, *IEEE Intell. Syst.,* **29**(2), 9–17.

Min, S., Lewis, M., Hajishirzi, H. & Zettlemoyer,, L., 2021. Noisy channel language model prompting for few-shot text classification, arXiv preprint (arXiv:2108.04106).

Minson, S.E., Baltay, A.S., Cochran, E.S., McBride, S.K. & Milner, K.R., 2021. Shaking is almost always a surprise: the earthquakes that produce significant ground motion, *Seismol. Res. Lett.,* **92**(1), 460–468.

Mousavi, S.M. & Beroza, G.C.(2022). Deep-learning seismology, *Science,* **377**(6607), eabm4470. doi:10.1126/science.abm4470

Musson, R.M.W., 2000. Intensity-based seismic risk assessment, *Soil Dyn. Earthq. Eng.,* **20**(5–8), 353–360.

Musson, R.M.W., 2009. Intensity and intensity scales, In *Manual of Seismological Observatory Practice 2 (NMSOP-2),* Geo Forschungs Zentrum GFZ, Deutsches, pp. 12–1-12-49.

Musson, R.M.W., Grünthal, G. & Stucchi, M., 2010. The comparison of macroseismic intensity scales, *J. Seismol.,* **14**(2), 413–428.

Nuttli, O.W., 1973. The Mississippi Valley earthquakes of 1811 and 1812: intensities, ground motion and magnitudes, *Bull. seism. Soc. Am.,* **63**(1), 227–248.

Palen, L. & Anderson, K.M., 2016. Crisis informatics—New data for extraordinary times, *Science,* **353**(6296), 224–225.

Quitoriano, V. & Wald, D., 2020. USGS "did you feel it?"—Science and lessons from 20 years of citizen science-based macroseismology, *Front. Earth Sci.,* **8,** 120. doi:10.3389/feart.2020.00120.

Quitoriano, V. & Wald, D.J., 2022. Combining macroseismic intensity felt reports using ShakeMap, *Proc. of the 3rd European Conf. on Earthquake Engineering & Seismology,* pp. 5017–5027, available athttps://www.infp.ro/index.php?i=3ecees (accessed June 2023).

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.B. & Mustafa, B., 2024. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context, preprint (arXiv:2403.05530).

Reynolds, L. & McDonell, K., 2021. Prompt programming for large language models: beyond the few-shot paradigm, in arXiv(arXiv:2102.07350).

Richter, C.F, 1958. *Elementary Seismology,* W. H. Freeman, pp. 10–10.
Sakaki, T., Okazaki, M. & Matsuo, Y., 2012. Tweet analysis for real-time event detection and earthquake reporting system development, *IEEE Trans. Knowl. Data Eng.,* **25**(4), 919–931.
Suzgun, M. *et al.* 2022. Challenging big-bench tasks and whether chain-of-thought can solve them, preprint (arXiv:2210.09261).
Team Gemini *et al*, 2023. Gemini: a family of highly capable multimodal models, preprint (arXiv:2312.11805).
TikTok, 2024. Developer Portal. Retrieved from, https://developers.tiktok.com/doc/display-api-get-started/ (accessed June 2024).
Vieweg, S., Hughes, A.L., Starbird, K. & Palen, L., 2010. *Microblogging during two natural hazards events: what twitter may contribute to situational awareness,* Proceedings of the SIGCHI Conference on human Factors in Computing Systems, Toronto, Canada, pp. 1079–1088.
Vinnell, L.J., Inch, P., Johnston, D.M. & Horspool, N., 2022. Behavioral responses to earthquake shaking: video footage analysis of the 2016 Kaikōura earthquake in Wellington, Aotearoa New Zealand, *Earthq. Spectra,* **38**(3), 1636–1660.
Wald, D.J., Goded, T., Hortacsu, A. & Loos, S.C., 2024. Developing and implementing an International Macroseismic Scale (IMS) for earthquake engineering, earthquake science, and rapid damage assessment. Open-File Report, *U.S. Geol. Surv.,* **55,** 2023–1098.
Wald, D.J., Quitoriano, V., Dengler, L.A. & Dewey, J.W., 1999. Utilization of the internet for rapid community intensity maps, *Seismol. Res. Lett.,* **70**(6), 680–697.
Wald, D.J., Quitoriano, V., Worden, C.B., Hopper, M. & Dewey, J.W., 2011. USGS "did you feel it?" internet-based macroseismic intensity maps, *Ann. Geophys.,* **54**(6), doi: 10.4401/ag-5354.
Wald, D.J., Worden, B.C., Quitoriano, V. & Pankow, K.L., 2005. *ShakeMap Manual: Technical Manual, User's Guide, and Software guide (No. 12-A1)*.
Wald, D.J., Worden, C.B., Thompson, E.M. & Hearne, M., 2022. ShakeMap operations, policies, and procedures, *Earthq. Spectra,* **38**(1), 756–777.
Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R.K.W. & Lim, E.P., 2023. Plan-and-solve prompting: improving zero-shot chain-of-thought reasoning by large language models, preprint (arXiv:2305.04091).
Waskom, M.L., 2021. Seaborn: statistical data visualization, *J. Open Source Softw.,* **6**(60), 3021. doi:10.21105/joss.03021.
Wei, J. *et al.*, 2022. Emergent abilities of large language models, preprint (arXiv:2206.07682).
Wood, H.O. & Neumann, F., 1931. Modified Mercalli Intensity Scale of 1931, *Bull. seism. Soc. Am.,* **21**(4), 277–283,
Worden, C.B., Wald, D.J., Allen, T.I., Lin, K., Garcia, D. & Cua, G., 2010. A revised ground-motion and intensity interpolation scheme for ShakeMap, *Bull. seism. Soc. Am.,* **100**(6), 3083–3096.
Xie, S.M., Raghunathan, A., Liang, P. & Ma, T., 2021. An explanation of in-context learning as implicit bayesian inference, preprint (arXiv:2111.02080).
Yin, J., Lampert, A., Cameron, M., Robinson, B. & Power, R., 2012. Using social media to enhance emergency situation awareness, *IEEE Intell. Syst.,* **27**(06), 52–59.
Zhang, X. *et al.* 2024. Modeling protective action decision-making in earthquakes by using explainable machine learning and video data, *Sci. Rep.,* **14**(1), 5480. doi:10.1038/s41598-024-55584-7.
Zhao, Z., Wallace, E., Feng, S., Klein, D. & Singh, S., 2021. Calibrate before use: improving few-shot performance of language models, In *International Conference on Machine Learning,* PMLR, pp. 12697–12706.
Zhou, A., Qian, W. & Ma, H., 2012. Social media data analysis for revealing collective behaviors, In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* United States, pp. 1402–1402.