The digital archivist: Automating legacy macroseismic data processing using large language models

Aarnav Agrawal[1†]; Susan E. Hough[2]; S. Mostafa Mousavi[3]; Khant N. Hlaing[4]; Clara E. Yoon[2]; Salvador Blanco[3]

[1]Monta Vista High School , [2]United States Geological Survey, [3]Harvard University, [4]University of California, Los Angeles

[†]corresponding author: aarnavagrawal@gmail.com

Abstract

Macroseismic data are a key resource to investigate shaking and damage from pre-instrumental and pre-digital earthquakes. However, data are often stored as unstructured reports describing observed shaking and damage, making manually parsing and interpreting accounts labor-intensive. We present a workflow using Google's Gemini 2.5 Pro LLM and the Google Geocoding API to extract and interpret intensity reports automatically. Applied to the 1957 M5.3 Daly City earthquake, Gemini created a dataset of over 2,500 reports, inferring missing MMI values and geocoding addresses. MMI comparisons with instrumental data from the San Francisco Bay Area show a mean absolute error of ~0.5, and ~0.35 with 0.5 MMI increments. Preliminary results for the 1971 Sylmar earthquake further confirm the reliability of LLMs for seismological tasks. This workflow offers a scalable method to digitize macroseismic archives, enabling large-scale analysis for seismic hazard assessment and urban site-effect studies.

Introduction

- Macroseismic data (historical "felt reports") provide critical insight into earthquakes prior to digital and instrumental eras (Mallet, 1862; Ambraseys, 1971; Bakun and Wentworth, 1997)
- Legacy datasets (e.g., postcard questionnaires, Abstracts Reports) remain largely unexploited due to their unstructured, labor-intensive format (Hough et al., 2025)
- New technologies like artificial intelligence (AI), especially large language models (LLMs), demonstrate strong performance in being applied in seismological contexts (Mousavi et al, 2024; Dagdelen et al., 2024)
- We propose The Digital Archivist, a novel workflow that applies LLMs and geocoding to unlock historical macroseismic archives, enabling modern use in ShakeMaps and hazard models

Methodology

Workflow Architecture and Implementation

- Used the Gemini 2.5 Pro LLM via its application programming interface (API) to create a scalable scientific pipeline that can be customized with system prompts and fine-tuned parameters to be applied to other documents
- Governed by a single prompt that employs a zero-shot (untrained) prompting technique with the following sequence of tasks for each "felt report" entry:
  - Only process reports corresponding to the mainshock
  - Extract the full postal address associated with the report
  - Extract the exact qualitative description of shaking and damage
  - If an MMI value is explicitly stated, extract it. Otherwise, infer an MMI value given a definition of the MMI scale and the ability to use 0.5-step increments for precision
  - Consolidate all information into a comma-separated value (CSV) string
- Due to Gemini's unreliability with geocoding, we used the Google Geocoding API to find latitudes and longitudes given standardized postal addresses

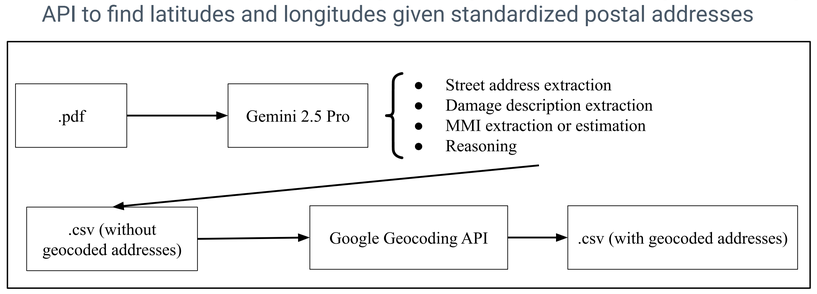

Figure 1: A visual representation of the workflow to process "felt reports" in PDF format.

Case Study

- Applied to the 22 March 1957 M5.3 earthquake in Daly City, California using data from the Abstracts Report (USCGS, 1957)
- Used data from five strong-motion accelerometers in the San Francisco Bay Area for quantitative validation against instrumental ground truth

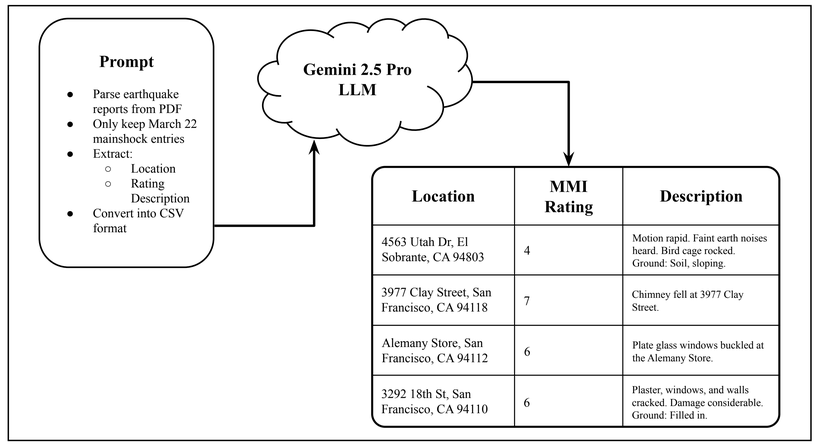

Figure 2: Example of zero-shot prompting applied to the 1957 Daly City earthquake. A detailed instruction prompt is given to the LLM without any pre-solved examples, and the model outputs structured data in CSV format parsed from raw earthquake reports.

Results

- Processed ~2,300 shaking and damage accounts
- Converted PGA and PGV from strong-motion stations to estimated MMI values using GMICE by Worden et al. (2012)
- Created an intensity field and compared instrumental MMI values with interpolated MMI values at the locations of the strong-motion stations
- Calculated a mean absolute error (MAE) of ~0.5; decreased to ~0.35 when Gemini was allowed to assign MMI values in increments of 0.5

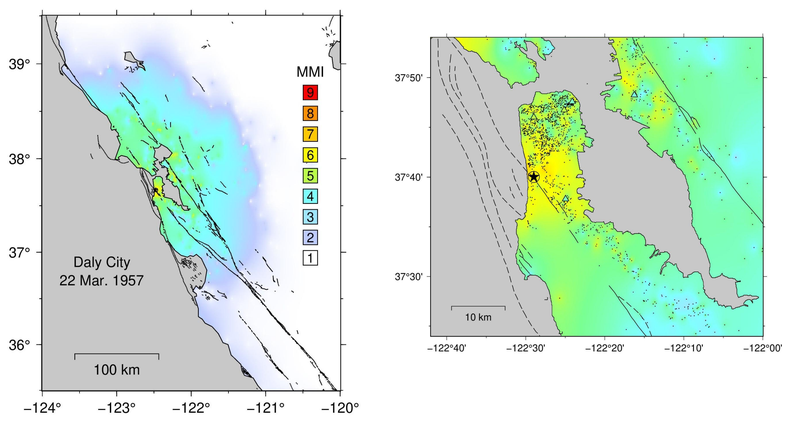

Figure 3: (left) MMI intensity distribution generated with a Laplacian operator to interpolate between control points. (right) Zoomed-in map. Filled triangles show instrumental intensities estimated from strong-motion data. Small dots indicate locations for which intensities are estimated.

Results

- On average, the intensity distribution fits the predicted intensities well (calculated using the Atkinson et al. (2014) Intensity Prediction Equation)

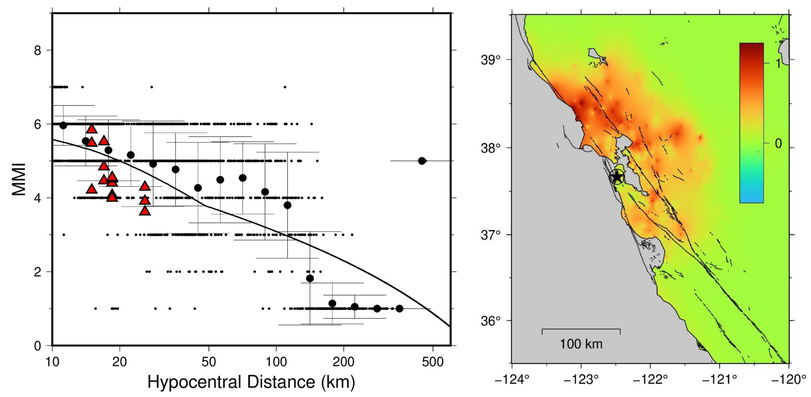

Figure 4: (left) Intensity values (small dots) and bin-averaged values (larger black dots) for the 1957 Daly City earthquake. Red triangles indicate instrumental intensities. Solid line shows predicted intensities. (right) Residuals to IPE predictions (observed - predicted MMIs) in map view.

- We consider the 1971 Sylmar, California earthquake to test the workflow's ability to analyze severe damage, creating a dataset with 1,100+ accounts

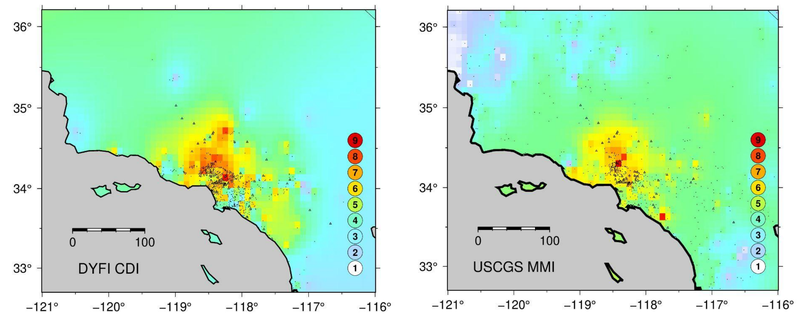

Figure 5: (left) Did You Feel It? intensity distribution for the 1971 Sylmar earthquake. Filled triangles indicate instrumental intensities. Small dots indicate MMI locations. (right) Similar plot showing USCGS intensities.

Conclusion/Discussion

- We developed a workflow using LLMs and APIs to transform unstructured legacy macroseismic data into a format amenable to quantitative analysis
- Accuracy matches traditional human intensity assignments while providing scalability across decades of U.S. earthquakes (e.g., the Abstract Report series published by the USCGS from 1928–1973)
- Can assist with the construction of modern data products (e.g., USGS ShakeMaps) for historic events
- Generalizable approach for mining other unstructured seismic archives (e.g., 1933 Long Beach, 1971 San Fernando, 1994 Northridge)
- Future work can refine refine LLM-based inference at higher shaking levels

References

Ambraseys, N. (1971). Value of historical records of earthquakes. *Nature*, 232, 375–379.

Atkinson, G. M., D. M. Boore, and D. J. Wald (2014). Intensity prediction equations for California and eastern North America. *Bull. Seismol. Soc. Am.*, 104(1), 308–319. doi:10.1785/0120140178

Bakun, W. H., and C. M. Wentworth (1997). Estimating earthquake location and magnitude from seismic intensity data. *Bull. Seismol. Soc. Am.*, 87, 1502–1521. doi:10.1785/BSSA0870061502

Dagdelen, J., A. Dunn, S. Lee, N. Walker, A.S. Rosen, G. Ceder, K.A. Persson, and A. Jain (2024). Structured information extraction from scientific text with large language models, Nat. Comm. 15, 1418. doi:10.1038/s41467-024-45563-x

Hough, S. E., L. Dengler, R. McPherson, L. Hajos, and M. Hellweg (2025). Did they feel it? Legacy macroseismic data illuminates an enigmatic 20th century earthquake, in press, Earth and Space Science.

Mallet, R. (1862). Great Neapolitan Earthquake of 1857: The First Principles of Observational Seismology. Chapman and Hall, London.

Mousavi, S. M., M. Stogaitis, T. Gadh, R. M. Allen, A. Barski, R. Bosch, P. Robertson, Y. Cho, N. Thiruverahan, and A. Raj (2024). Gemini and physical world: large language models can estimate the intensity of earthquake shaking from multimodal social media posts. *Geophys. J. Int.*, 240(2), 1281–1294. doi:10.1093/gji/ggae436

U.S. Coast and Geodetic Survey (USCGS) (1957). Abstracts of Earthquake Reports for the Western United States, March 1957. U.S. Government Printing Office, Washington, D.C.

Worden, C. B., M. C. Gerstenberger, D. A. Rhoades, and D. J. Wald (2012). Probabilistic relationships between ground-motion parameters and Modified Mercalli Intensity in California. *Bull. Seismol. Soc. Am.*, 102(1), 204–221. doi:10.1785/0120110156

Acknowledgements

This research was supported by the Statewide California Earthquake Center (Contribution No.14630). SCEC is funded by NSF Cooperative Agreement EAR-2225216 and USGS Cooperative Agreement G24AC00072.

We thank Jim Dewey for helpful guidance with legacy macroseismic data, and Ryan Gold for his stewardship of the postcard questionnaire data.