

Automatic Knowledge Graph Construction and Query for Earthquake Catalogs

Yuxin Zhou^{1,2} , Huai Zhang^{*1} , and S. Mostafa Mousavi² 

Abstract

In recent years, the number of events in earthquake catalogs has significantly increased due to the utilization of more effective deep-learning-based detectors and phase pickers but answering open-ended questions such as “what characterizes this sequence?” remains constrained by rigid spatiotemporal windowing and subjective expert interpretation. We present the first systematic application of graph-based retrieval-augmented generation (GraphRAG) directly to raw, tabular catalog records across three independently featured catalogs: a reservoir-adjacent swarm (Qiaojia-Dongchuan), the 2019 Ridgecrest tectonic sequence, and the 2021 Maduo Mw 7.4 aftershock sequence. Without the need for manual data structuring, the pipeline builds structurally complete, queryable knowledge graphs for all three. Rigorous evaluation — 1,200 answers individually verified against catalog-derived ground truth and a rule-based reference graph — exposes failure modes, and four seismology-informed prompt fixes eliminate all targeted fabrications while sharply improving mechanism reasoning (up to 2.90/3). A vector-RAG baseline demonstrates the graph layer’s distinctive value: catalog-wide summarization and temporal-stage comparison. In addition, we have identified two main pitfalls that need attention. GraphRAG thus offers a practical, transferable, near-zero-cost query interface for earthquake catalogs, where careful prompting ensures the results are consistently accurate and trustworthy.

Keywords: knowledge graph; retrieval-augmented generation; GraphRAG; earthquake catalog; large language model; benchmark evaluation

Cite this article as Zhou, Y., H. Zhang, and S.M. Mousavi (2026). Automatic Knowledge Graph Construction and Query for Earthquake Catalogs, *The Seismic Record* **0**(0), 1–10, doi: [00.0000/0000000000](https://doi.org/10.0000/0000000000).

[Supplemental Material](#)

Introduction

Dense-array monitoring (Ross et al., 2019; Shelly, 2020) and deep-learning phase detection/picking (Zhu and Beroza, 2019; Mousavi et al., 2020; Ross et al., 2018; Mousavi and Beroza, 2022) have together driven an order-of-magnitude expansion in earthquake-catalog size, for example, the

2021 Maduo M_w 7.4 aftershock sequence contains over 10,000 relocated events (Guan et al., 2024). Beyond scale, a deeper challenge is characterizing what high-resolution catalogs represent physically: mainshock-aftershock decay (Gutenberg and Richter, 1944; Utsu et al., 1995), swarms, foreshock sequences identifiable only in hindsight, and induced seismicity (Gupta, 2002) are still distinguished mainly by coarse qualitative criteria, or point-process models fit after the fact (Ogata, 1988), that does not allow quantitative, reproducible cross-sequence comparison in real time. However, this increase in the typical sizes of earthquake catalogs has not helped yet to this major challenge. Our analyses and interpretations of the high-resolution deep-learning based catalogs remain more or less limited to a few classical engineered features.

1. State Key Laboratory of Earth System Numerical Modeling and Application, College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing 100049, China,  <https://orcid.org/0009-0000-2877-9173> (FA)  <https://orcid.org/0000-0003-0411-4841> (SA) 2. Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA 02138, USA,  <https://orcid.org/0009-0000-2877-9173> (FA)  <https://orcid.org/0000-0001-5091-5370> (TA)

*Corresponding author: H. Zhang, h Zhang@ucas.ac.cn

© 2026. The Authors. This is an open access article distributed under the terms of the CC-BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Knowledge graphs organize such semantically linked information as reasoning-ready (entity, relationship, entity) triples (Hogan et al., 2021), and have been applied to earthquake emergency response (Qiu et al., 2024) and seismic metadata organization (Davis and Hunt, 2024). Retrieval-Augmented Generation (Lewis et al., 2020) is an artificial intelligence architecture that enhances the accuracy and physical reliability of large language models by grounding their responses in dynamically retrieved, domain-specific external data. Microsoft's GraphRAG (Edge et al., 2024), rely on RAG systems to automatically extracting entities/relationships, detects communities via the Leiden algorithm (Traag et al., 2019), and generates natural-language summaries and query access, with no predefined ontology – part of a broader effort to unify LLMs and knowledge graphs (Pan et al., 2024). This raises a natural question: can GraphRAG be applied directly, end-to-end, to earthquake catalogs of varying scale, region, and character to produce a working natural-language query interface, and can its most damaging failure modes be suppressed with modest, seismology-specific engineering effort rather than a full custom rebuild?

RAG has recently begun to be applied within seismology and geohazard research specifically. Yao et al. (2025) combine knowledge-graph construction with a hybrid RAG strategy for earthquake emergency response, extracting entities and relationships from thousands of professional emergency-management documents; Orantes-Jiménez (2025) use LLMs to build knowledge graphs from earthquake news articles. Recent reviews (Yu et al., 2025; Li and Zhou, 2026) note that RAG in the geosciences remains applied mainly to narrative, prose-style sources – news, reports, technical literature – rather than to the raw, tabular observational records a catalog itself consists of, and call for RAG architectures tailored to structured geoscientific data. Our study differs in three respects: it applies RAG directly to structured, numerical catalog records rather than narrative text describing events after the fact; it targets fully automatic, schema-free construction with no per-catalog ontology or extraction-rule engineering, as opposed to the hand-designed ontologies underlying prior earthquake knowledge-graph work; and it evaluates GraphRAG specifically, whose automatic community detection and hierarchical summarization are built to transform thousands of brief catalog records into comprehensive, sequence-

level summaries—a core capability that our benchmark explicitly evaluates.

Section 3 demonstrates that fully automatic indexing and natural-language querying run end-to-end on all three catalogs under default prompts, establishing mechanical transferability while exposing low baseline answer quality (catalog-only averages 0.76–1.22/3). Section 4 quantifies the catalog-dependent effect of narrative-text enrichment. Section 5 describes an iterative, seismology-oriented prompt-engineering scheme that eliminates the targeted fabrication modes across all 600 post-fix answers. Furthermore, this method introduced a baseline vector-RAG comparison (embedding-similarity retrieval over the same text chunks, with no graph layer) to quantify the community-report layer's impact on holistic questions. Section 6 details the specific catalog errors identified during verification—such as the systematic misdating of the Ridgecrest mainshock and the hallucinated inclusion of an unlisted Maduo event. We report these issues to guide safe deployment, rather than as the study's central finding. Extended examples and full per-condition benchmark results as well as prompt details are provided in the Supplementary Material (SM).

Materials and Methods

Case Studies

We use three independent catalogs differing in scale, region, and origin, totaling 20,027 events (Figure 1; overview table in SM Table S4): (1) the Qiaojia-Dongchuan relocation catalog, 5,218 events recorded by a temporary dense array adjacent to the Baihetan Reservoir between 23 Aug 2022 and 17 Mar 2023; (2) the 2019 Ridgecrest sequence, 4,188 events ($M \geq 2.0$) spanning July 2019, a standard catalog (Shelly, 2020) on a purely tectonic strike-slip system in the Mojave Desert; and (3) the 2021 Maduo $M_w 7.4$ sequence, the largest of the three at 10,621 events ($M \geq 0.5$, 1 Jun 2021–8 Jun 2023), relocated by Guan et al. (2024). This aftershocks-only catalog begins after the true 22 May 2021 mainshock and contains no May-2021 data.

GraphRAG pipeline and evaluation design

To effectively translate complex seismicity data into a structured, queryable knowledge base, the system architecture relies on two primary phases: Indexing and Retrieval (Figure 2a). The indexing phase transforms raw, disparate

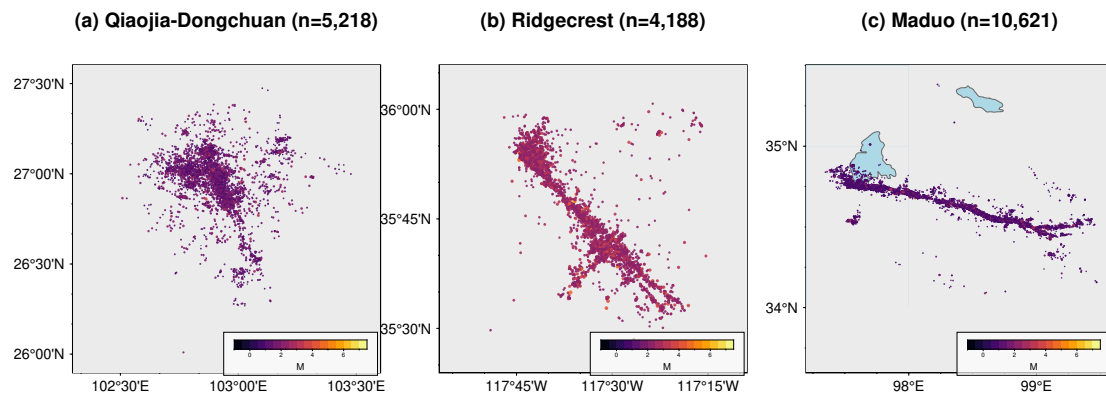


Figure 1. Spatial distribution of seismic events. (a) Qiaojia-Dongchuan; (b) Ridgecrest; (c) Maduo. Epicenters colored by magnitude.

Alt text: Three side-by-side scatter maps with latitude/longitude axes, each titled with catalog name and event count. Dots mark epicenters, colored on a dark-purple-to-yellow magnitude scale (0–7) per a small color-bar legend in each panel's corner. (a) Qiaojia-Dongchuan: a roughly circular, dense blob of points near 103°E, 27°N. (b) Ridgecrest: a narrow, elongated diagonal band of points trending NW-SE. (c) Maduo: a long, thin, mostly east-west line of points spanning about 98–99.5°E, with two small blue lake outlines in the upper left.

earthquake catalog entries into a deeply connected network. It begins with LLM-based entity and relationship extraction (using `gpt-4o-mini`) to identify key seismic features—such as specific earthquakes, fault structures, and their spatiotemporal links. Next, Leiden community detection groups these interconnected events into highly related, tectonic or sequence-based clusters (e.g., distinct swarms or aftershock zones). The system then processes these clusters through automated community-report generation, creating high-level textual summaries that describe the defining characteristics of each localized sequence.

During the retrieval phase, the system supports two complementary search modes for interacting with the indexed seismic data. *Global Search* performs a holistic synthesis across all community reports, making it ideal for answering broad, sequence-level questions regarding overarching migration patterns or aggregate statistics. In contrast, *Local Search* executes precise entity retrieval to isolate exact details concerning specific mainshocks, stations, or localized catalog anomalies. Full parameter configurations for both the indexing and retrieval pipelines are detailed in the Supplementary Material (SM).

Evaluation proceeds in three stages: (i) capability demonstration under default prompts (Section 3); (ii) narrative-text enrichment, evaluated with a 100-question benchmark spanning five categories – (A) overall summarization, (B) precise statistics, (C) local retrieval, (D) physical mechanism, (E) temporal-stage comparison; 20 questions each

(Section 4); and (iii) iterative, seismology-oriented prompt engineering targeting the exposed failure modes (Section 5). Two independent ground truths anchor scoring: statistics computed directly from each catalog's raw records, and, for Qiaojia-Dongchuan, a rule-based knowledge graph (10,766 nodes; 113,620 edges) answering the same questions deterministically. To provide a point of comparison, we identically scored a standard vector-RAG baseline. This baseline used the exact same text chunks, embeddings, and LLM as our GraphRAG pipeline, but relied on top-10 chunk similarity for retrieval rather than a graph layer.

We employed an LLM assistant to score the 1,200 GraphRAG and 300 baseline responses (from 0 to 3) against two ground truths. Following this automated scoring, the authors manually validated all refusals and Category-B/D answers against raw catalog statistics, and the first author resolved all borderline evaluations. To validate the automated scoring, the first author conducted a blind audit of 29 stratified-random answers. When adjudicated against the ground truth, this audit showed 76% exact agreement with the LLM workflow and 100% agreement within one point, with zero unresolved fabrication disputes. In contrast, a purely unaided human review achieved only 52% exact agreement and erroneously accepted two fabricated answers as correct, highlighting the critical need for ground-truth anchoring over subjective human judgment. Responses were graded using the following rubric: a score of 3 indicates a correct, independently verifiable central claim; a score of

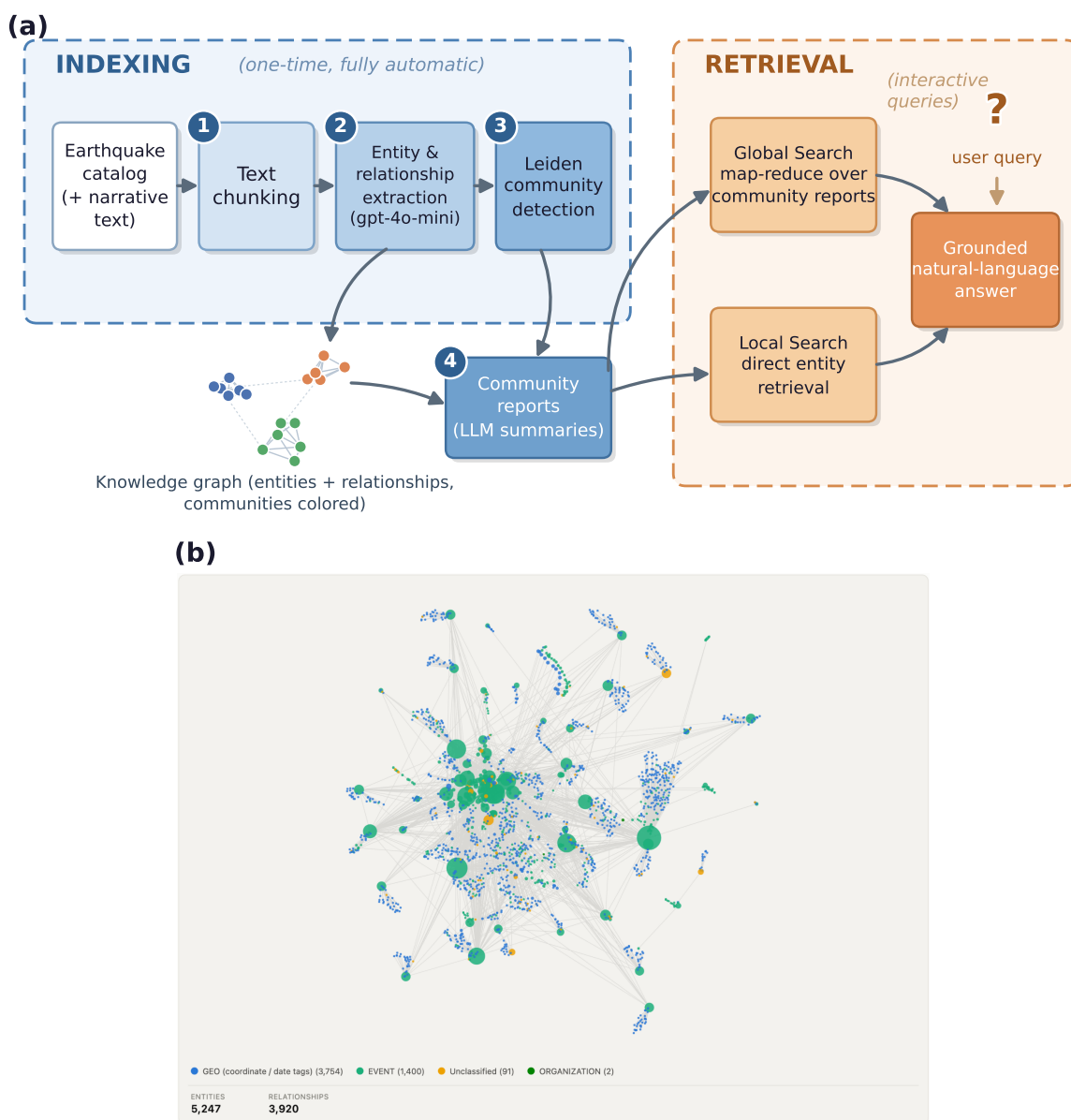


Figure 2. (a) Schematic of the GraphRAG indexing and retrieval pipeline. (b) Entity-relationship graph automatically constructed by GraphRAG from the Qiaojia catalog (5,247 entities; 3,920

relationships), with no hand-designed schema; GEO (coordinate/date tags), EVENT (individual earthquake records), and a small residual of Unclassified/Organization entities from generic extraction.

Alt text: Two-panel figure. (a) Flowchart: a blue dashed box 'INDEXING' contains four linked boxes—catalog input, text chunking, LLM entity/relationship extraction, Leiden community detection—feeding a small node-cluster icon labeled 'Knowledge graph' and a 'Community reports' box; an orange dashed box 'RETRIEVAL' shows 'Global Search' and 'Local Search' boxes both arrowing into a 'Grounded natural-language answer' box, with a user-query arrow above. (b) A dense force-directed node-link network on a gray background: small dots joined by thin gray edges, colored by legend—blue GEO, teal EVENT, orange Unclassified, green Organization—with large teal hub nodes near the center.

2 denotes a correct central claim grounded in the data, but containing secondary errors, omissions, or unverifiable elements; a score of 1 applies when the central claim is wrong or missing (e.g., refusing an answerable question or misrepresenting data subsets) despite using genuine records; and

a score of 0 is reserved for fabricated values, hallucinated records, or false claims presented as factual. Refusals earn a 3 only if the required information is completely absent from the catalog.

To quantify uncertainty, 95% bootstrap confidence intervals were calculated for the per-answer scores. For individual category cells ($n=20$), the interval half-widths reached a maximum of 0.66, meaning score differences below ~ 0.5 are indistinguishable from scoring noise. For broader condition averages ($n=100$), this resolution tightens to ± 0.22 , establishing a ~ 0.3 threshold for significance. Consequently, minor variations—such as Maduo’s +0.04 gain, the narrative-enrichment declines, and GraphRAG’s overall edge over the baseline on Qiaojia and Ridgecrest—fall within this noise margin. However, the sharp improvements in Category D, the overall score gains for Qiaojia (+0.73) and Ridgecrest (+0.45), and the baseline’s outperformance of GraphRAG on Maduo remain statistically robust (Table 2)

Results: Capability Demonstration

With no hand-designed schema and no catalog-specific configuration, GraphRAG automatically built structurally complete, queryable knowledge graphs for all three catalogs from raw tabular records alone: Qiaojia-Dongchuan (5,247 entities, 3,920 relationships, 146 communities; Figure 2b), Ridgecrest (1,985 entities, 2,449 relationships, 83 communities), and Maduo (15,402 entities, 11,784 relationships, 600 communities). Entity types are consistently dominated by coordinate/date (GEO) tags ($\sim 72\text{--}83\%$) across all three, confirming that indexing capability is general and does not depend on catalog region or scale. Per-catalog entity-type and community-size distributions are given in the SM (Figures S1–S2). The resulting graphs support natural-language queries that a fixed-field database cannot: Local Search queries retrieve precise, traceable individual records (e.g., correctly naming Ridgecrest’s *M7.1* mainshock and its 6 July date), and Global Search synthesizes holistic, catalog-wide descriptions of sequence character. However, this out-of-the-box functionality has reliability issues. When reviewing 80 Ridgecrest benchmark answers (excluding local retrieval), the system frequently misidentified smaller earthquakes (*M3.5–M5.5*) as the sequence’s largest event. Additionally, default prompts caused the model to hallucinate fluent but baseless physical mechanisms, such as falsely attributing Ridgecrest’s seismicity to reservoir or fluid processes. Under the default, catalog-only condition, the average scores were generally poor: 0.96 for Qiaojia, 1.22 for Ridgecrest, and 0.76 for Maduo out of a possi-

ble 3. Qiaojia struggled most with mechanism questions (Category D at 0.55/3), while Maduo suffered from uniformly low scores across all categories (0.45–0.95/3). These baseline deficits, detailed in SM Table S5, motivated the targeted prompt-engineering strategy in Section 5, which successfully doubled several of these scores. Table 1 gives six representative question/answer pairs – four correct, high-quality responses spanning extremal retrieval, aftershock association, narrative-grounded mechanism reasoning, and a correct refusal, plus the two residual errors detailed in Section 6 – to illustrate concretely what the model’s output looks like; extended examples are given in the SM.

Results: Narrative-Text Enrichment

To assess the effect of narrative enrichment, each catalog was re-indexed with contextual literature—Baihetan Reservoir background for Qiaojia, Shelly (2020) for Ridgecrest, and Guan et al. (2024) for Maduo—and re-evaluated. The effect is catalog-dependent and non-monotonic (Category-D scores in SM Table S6). Qiaojia’s Category-D score rises (0.55→1.35) but mainly reflects a more cautious tone matching the text’s “contested” framing rather than genuine use of its specific named entities. For Ridgecrest, the score remained essentially unchanged (1.45 to 1.40). When asked about industrial activity, the model simply refused to answer instead of using the newly provided text by Shelly (2020) to confidently deny the claim. While a blank refusal is the correct response when using only the raw catalog—since the bare catalog lacks mechanism data—it represents a failure to utilize the enriched text, which contained the evidence needed to explicitly rule out the mechanism. While Maduo’s score increased (0.75 to 1.65), the enriched text introduced a major new mistake: the model confused the natural deep-fluid processes described by Guan et al. with human-induced hydraulic fracturing. This error did not occur when the model relied on the catalog alone. Additionally, under default settings, the model hallucinated the answer for the “largest-magnitude event”. It correctly named the real-world May 2021 mainshock, but because the provided catalog data did not start until June 1, the model pulled this fact from its outside memory rather than the data—a “coincidentally correct” error that persisted even after prompt engineering.

Table 1.

Representative example queries and GraphRAG answers, spanning correct, high-quality responses and the two residual error classes of Section 6 (translated from the original Chinese-language queries; retrieval mode and condition in parentheses; all answers from the fully re-verified post-fix benchmark).

Question	Answer (abridged)	Outcome
When, where, and at what magnitude did the largest event in the Qiaojia-Dongchuan catalog occur (Global Search, catalog-only)?	"The largest-magnitude event in the catalog occurred on 18 October 2022 at 22:12:35.080..., with magnitude 4.2."	Correct (true: $M_{4.18}$, 18 Oct 2022 22:12)
Which events are directly related to the largest-magnitude event in the 2019 Ridgecrest sequence (Local Search, catalog-only)?	"The largest-magnitude event ... is the $M_{7.10}$ earthquake on 2019-07-06 at 03:19:52.860. This event is a major mainshock, followed by a series of aftershocks directly related to it..."	Correct
What is the approximate rupture/triggering mechanism of the 2021 Maduo $M_{7.4}$ sequence (Global Search, narrative-enriched)?	"...generated by the KMPJ fault, clearly identified as the main seismogenic fault, ... interacting with secondary faults (the Maduo-Gande and Jiuzhi faults)... the aftershock sequence is complex, with a bifurcating geometry..."	Correct; grounded in the enrichment text rather than pretrained priors
Is the 2019 Ridgecrest seismicity associated with industrial activity such as hydraulic fracturing, wastewater injection, or reservoir impoundment (Global Search, catalog-only)?	"I am sorry but I am unable to answer this question given the provided data."	Correct refusal: the purely tectonic catalog contains no such evidence; default prompts had answered from background knowledge
What is the trend in seismicity rate over time in the 2019 Ridgecrest sequence (Global Search, catalog-only)?	"...on July 4, 36 events were recorded, with a maximum magnitude of 4.6. Activity then peaked on July 6, with multiple nearby epicenters recorded, maximum magnitude 4.97..."	Wrong: omits the true $M_{7.1}$ mainshock on its own day (Section 6)
What is the largest-magnitude event in the Maduo sequence (Global Search, narrative-enriched)?	"...the sequence's main event is the $M_{7.4}$ earthquake of 22 May 2021... recorded as EVENT 2021-05-22T00:00:00..."	Wrong: catalog contains no May-2021 data (Section 6)

Results: Iterative Prompt Engineering

Following the identification of mechanistically distinct failure modes in Sections 3 and 4, we deployed iteratively verified prompt and configuration adjustments to address four specific issues identically across all conditions. These targeted errors consisted of magnitude fabrication via field confusion, historical-earthquake conflation, cross-catalog place-name contamination, and a scope-collapse bug in community detection. Each intervention was tailored to a diagnosed root cause; for instance, the scope-collapse issue was traced to a default GraphRAG clustering parameter that inadvertently discarded the majority of catalog entities. Comprehensive diagnostic evidence, exact prompt modifi-

cations, and mechanistic details are provided in SM Text S5.

We re-indexed every affected condition and re-ran the full benchmark, individually re-verifying all answers (Figure 3; Table 2). All three targeted fabrication modes were effectively eliminated: no impossible magnitude value appeared in any of the 600 re-verified post-fix answers (versus repeated " $M_{10.4}$ "/" $M_{12.8}$ " before), no cross-catalog place-name contamination was found, and historical-earthquake conflation (previously ~ 1 in 5 relevant answers) disappeared. Despite the prompt fixes, two distinct types of hallucinations remained: presenting out-of-catalog events as real data, and inventing aggregate counts (detailed in Section 6). While scores for physical mechanism questions (Category

D) improved dramatically—jumping from 0.55 to 2.90 for Qiaojia and 1.45 to 2.85 for Ridgecrest—precise statistics (Category B) remained persistently weak. Category B scored between 0.50 and 1.04 out of 3, making it the lowest-performing category in five out of six test conditions. The main issue is that the model severely undercounts totals and rates. This happens because a retrieval-then-synthesize system tries to answer math questions using a few text snippets rather than scanning the entire database—a fundamental architectural flaw rather than a prompt issue (detailed further in SM Text S5.4).

The vector-RAG baseline (Table 2) shows that GraphRAG’s advantage concentrates precisely where its community reports are designed to help: holistic summarization (Category A: 1.85 vs. 1.30 on Qiaojia) and temporal-stage comparison (Category E: 1.70 vs. 1.10 on Ridgecrest), where top-*k* chunk retrieval cannot synthesize catalog-wide structure and mostly returns hedged partial descriptions or refusals. Category B is comparably weak for both architectures (0.85–1.05 across the Qiaojia and Ridgecrest catalog-only conditions; lower still for GraphRAG on Maduo), confirming this limitation is common to retrieval-then-synthesize systems rather than specific to GraphRAG. The baseline’s perfect Category-D scores (3.00) reflect uniform honest refusals scored as correct because the catalogs contain no mechanism evidence; GraphRAG’s 2.85–2.90 comes from substantive grounded answers, so Category D is uninformative here – the architectural comparison rests on Categories A and E. On Maduo, however, the baseline’s uniformly honest refusals outscore GraphRAG’s degraded index (average 1.40 vs. 0.80): the graph layer’s value is conditional on a healthy index. The ranking is sensitive to the refusal-scoring convention (Table 2, note), so we report both. Notably, the baseline reproduces the famous-mainshock intrusion of Section 6 – misdating Ridgecrest’s *M*7.1 from pretrained knowledge while explicitly admitting the retrieved chunks lack it – evidence that this hazard is intrinsic to LLM synthesis, not to GraphRAG.

Residual Catalog-Specific Errors

Our per-answer verification of the fully fixed pipeline (Section 5) also surfaced two narrower, catalog-specific errors, concentrated on the two catalogs tied to a globally famous mainshock – though not exclusively: a mechani-

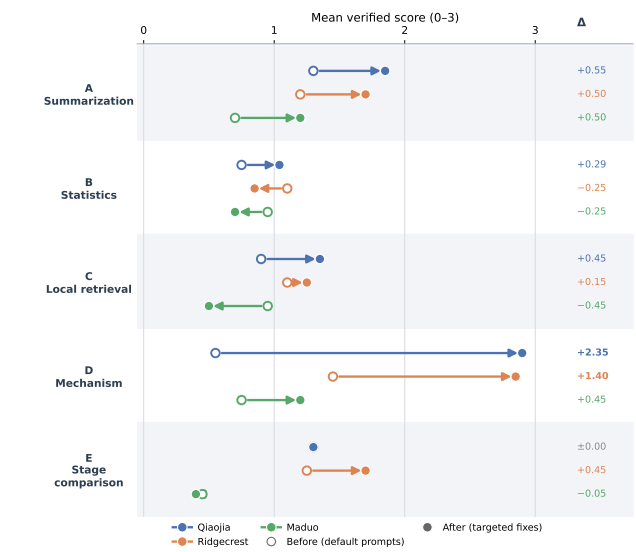


Figure 3. Per-category verified benchmark scores before (open circles; default prompts) and after (filled circles; targeted fixes) the Section 5 prompt-engineering scheme, catalog-only condition. Arrows show the direction of change for each catalog; the right column (Δ) gives the per-catalog score change. Physical-mechanism questions (Category D) improve most; precise-statistics questions (Category B) remain architecture-limited.

Alt text: Dumbbell (dot-and-arrow) plot. X-axis: mean verified score, 0 to 3. Y-axis: five question categories (A–E), each with three colored rows for Qiaojia (blue), Ridgecrest (orange), and Maduo (green). Each row shows an open circle (default prompts) joined by an arrow to a filled circle (after fixes), pointing right for gains and left for losses; a right-hand Δ column lists each numeric change. Category D shows the longest rightward arrows, up to +2.35; Category B shows short arrows, including two negative changes.

cal scan of all 200 post-fix Qiaojia answers for dates or coordinates outside the catalog’s range found six (3%, all Category-C local-retrieval questions, in both conditions) listing wholly invented event records; the same scan flags zero Ridgecrest answers. We report these to inform responsible deployment, not as a limitation of the pipeline’s core query capability demonstrated in Sections 3–5.

On **Ridgecrest**, the true *M*7.1 mainshock (6 Jul 2019, the busiest day) is placed on 4 or 5 July instead in 50–70% of relevant re-verified answers, often merged with the *M*6.4 foreshock date; narrative enrichment did not reduce this. The correct answers in Section 3 and Table 1 are drawn from the complementary 30–50%: the same question type yields the correct date in a substantial minority of answers, which is precisely what makes this error hazardous – spot-checking a few correct answers cannot exclude it. On **Maduo**, whose catalog begins 1 June 2021 with no May data, the real out-of-catalog 22-May-2021 mainshock is nonetheless presented as the catalog’s own largest event in up to 60%

Table 2.

Verified category-average scores after the targeted fixes (Section 5), all catalogs/conditions, and the vanilla vector-RAG baseline (catalog-only).

Condition	A	B	C	D	E	Avg
Qiaojia catalog-only	1.85	1.04	1.35	2.90	1.30	1.69
Qiaojia +narrative	1.85	0.90	1.30	1.60	1.50	1.43
Ridgecrest catalog-only	1.70	0.85	1.25	2.85	1.70	1.67
Ridgecrest +narrative	1.23	0.90	1.05	2.60	1.10	1.38
Maduo catalog-only	1.20	0.70	0.50	1.20	0.40	0.80
Maduo +narrative	0.65	0.50	0.90	2.15	1.15	1.07
Qiaojia vector-RAG baseline	1.30	1.00	1.15	3.00	1.00	1.49
Ridgecrest vector-RAG baseline	1.45	1.05	1.20	3.00	1.10	1.56
Maduo vector-RAG baseline	1.00	1.00	1.00	3.00	1.00	1.40

A: summarization; B: statistics; C: local retrieval; D: mechanism; E: stage comparison. Baseline rows share chunks, embeddings, LLM, and questions with the corresponding catalog-only condition (Section 2); their Category-D scores reflect uniform honest refusals scored as correct. Under the stricter convention scoring refusals on answerable questions 0 rather than 1, the baseline averages fall to 0.98, 1.20, and 0.69.

of narrative-enriched answers (versus $\leq 15\%$ catalog-only), alongside an independently fabricated “78,832 aftershocks” figure recurring in $\sim 30\%$ of enriched mechanism/stage-comparison answers. We interpret the catalog-dependent asymmetry as evidence that an LLM’s pretrained knowledge of a globally famous, heavily reported event can, in a minority of answers, override retrieved catalog-grounded evidence during synthesis – a narrower and more specific hazard than the general “ungrounded generation” hallucination described in the broader literature (Ji et al., 2023), and one that our Section 5 HISTORICAL_EVENT fix, designed for a structurally similar problem (historical-earthquake conflation), did not fully generalize to suppress. These hazards resist casual review: in our scoring audit (Section 2), two fabricated-record answers were initially rated fully correct by an unaided human pass and were caught only when their dates and coordinates were checked against the catalog’s actual range. Extended examples are given in the SM.

Discussion and Conclusions

This study demonstrates that GraphRAG can be applied directly, end-to-end, to earthquake catalogs of markedly different scale, region, and origin – with no hand-designed schema, no custom extraction rules, and no per-catalog

reconfiguration – to produce structurally complete knowledge graphs supporting flexible natural-language query, with verified instances of precise record retrieval and grounded mechanism reasoning (Table 1), though answer reliability varies by catalog and question category (Table 2). This is, to our knowledge, the first systematic demonstration of GraphRAG’s schema-free construction capability transferring across independent earthquake catalogs at essentially zero ontology-engineering cost, although downstream answer quality does not transfer uniformly (post-fix catalog-only averages of 1.69, 1.67, and 0.80/3).

Beyond capability, we show that reliability is not fixed once GraphRAG is applied out of the box, but can be measurably improved with modest, seismology-oriented prompt engineering: our four targeted fixes (Section 5) eliminated three concrete error types from the re-verified answer set entirely and raised mean scores from 0.76–1.22/3 (SM Table S5) to 0.80–1.69/3 (Table 2) – catalog-only gains of +0.73 for Qiaojia and +0.45 for Ridgecrest, but an essentially flat +0.04 for Maduo, where gains in summarization and mechanism reasoning were offset by declines in statistics, retrieval, and stage comparison. These findings point to a practical and affordable strategy for deploying LLM query tools on earthquake catalogs: start with the default

GraphRAG pipeline, test it against a small ground-truth benchmark, and refine the prompts instead of building a new system from scratch. However, there is a catch: while this approach successfully eliminated the targeted hallucinations across all test cases, it only produced significant overall score improvements for two of the three catalogs. Narrative-text enrichment, by contrast, reduced post-fix overall scores for Qiaojia (1.69→1.43/3) and Ridgecrest (1.67→1.38/3); it raised Maduo's (0.80→1.07/3), but at the cost of more frequent out-of-catalog mainshock intrusions there (up to 60% of enriched answers versus ≤15% catalog-only), so it should be adopted only with per-catalog verification (Section 4).

Deployment is limited by two main caveats, the first being that precise-statistics queries (Category B) performed poorly across all conditions even after applying fixes, scoring just 0.50–1.04 out of 3. This category scored the lowest in five out of six evaluation conditions, which points to a fundamental structural flaw in retrieval-then-synthesize architectures when handling aggregate or ranking tasks. Therefore, rather than relying on prompt engineering to fix this issue, we recommend routing these specific mathematical questions directly to deterministic database or graph queries. The second caveat involves the residual errors detailed in Section 6, which primarily affected catalogs associated with well-known mainshocks. These errors show that verifying the output for each specific catalog remains a necessary step before deployment, especially when analyzing well-known events. A promising future fix would be explicitly instructing the model to never use its pre-trained background knowledge to "correct" the retrieved data, though we did not have the chance to design and test that approach in this study. Because all our results rely on a single model (gpt-4o-mini), the rate of hallucinations and the effectiveness of our fixes might vary if a different model is used. However, since the vector-RAG baseline reproduced the exact same hallucination (Section 6), we know this specific flaw stems from the LLM itself, not GraphRAG's architecture. Ultimately, neither limitation invalidates our main conclusion: GraphRAG is a highly useful tool for automatically building earthquake-catalog knowledge graphs, performing strongly for Qiaojia and Ridgecrest, and more modestly for Maduo. When the index is healthy (as with Qiaojia and Ridgecrest), GraphRAG's community-report layer noticeably outperforms a standard vector-RAG baseline on complex, holistic questions. Conversely, when the

index degrades (as with Maduo), the baseline matches or even beats GraphRAG. Overall, combining this architecture with targeted prompt engineering yields genuine, verifiable improvements in reliability.

Data and Resources

All data and codes are available at <https://doi.org/10.5281/zenodo.21459373>. The Supplementary Material accompanying this article provides full pipeline and benchmark parameters, per-category and per-catalog score tables, entity-type and community-size distributions, exact prompt-engineering modifications, and extended question/answer examples supporting Sections 3–6.

Declaration of Competing Interests

The authors acknowledge that there are no conflicts of interest recorded.

Acknowledgments

Y.Z. and S.M.M were supported by Harvard Milton Fund. The authors thank the developers of the GraphRAG framework and the seismological data providers whose catalogs and publications made this study possible. We also thank the members of Plantcore.AI, whose insights during our discussions helped inspire this work.

References

- Davis, W. and C. R. Hunt (2024). Knowledge graphs for seismic data and metadata. *Appl. Comput. Geosci.* **21**, 100151.
- Edge, D., H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson (2024). From local to global: A graph RAG approach to query-focused summarization.
- Guan, P. H., J. S. Lei, and D. P. Zhao (2024). Machine-learning based location of the 2021 Mw 7.4 Maduo earthquake sequence: Insight into intraplate seismogenesis. *Tectonophysics* **888**, 230458.
- Gupta, H. K. (2002). A review of recent studies of triggered earthquakes by artificial water reservoirs with special emphasis on earthquakes in Koyna, India. *Earth-Sci. Rev.* **58**(3–4), 279–310.
- Gutenberg, B. and C. F. Richter (1944). Frequency of earthquakes in California. *Bull. Seismol. Soc. Am.* **34**(4), 185–188.
- Hogan, A., E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A. C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann (2021). Knowledge graphs. *ACM Comput. Surv.* **54**(4), 1–71.

- Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**(12), 248.
- Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, Volume 33, pp. 9459–9474.
- Li, W. and Y. Zhou (2026). Toward knowledge-enhanced geohazard intelligence: A review of knowledge graphs and large language models. *GeoHazards* **7**(2), 40.
- Mousavi, S. M. and G. C. Beroza (2022). Deep-learning seismology. *Science* **377**(6607), eabm4470.
- Mousavi, S. M., W. L. Ellsworth, W. Zhu, L. Y. Chuang, and G. C. Beroza (2020). Earthquake transformer – an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nat. Commun.* **11**, 3952.
- Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Am. Stat. Assoc.* **83**(401), 9–27.
- Orantes-Jiménez, D. (2025). Harnessing large language models to build knowledge graphs from earthquake news. *Int. J. Digit. Earth* **18**(2), 2594950.
- Pan, S., L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans. Knowl. Data Eng.* **36**(7), 3580–3599.
- Qiu, P. Y., L. K. Pang, Y. Luo, Y. H. Liu, H. Q. Xing, K. Liu, and G. L. Zhuang (2024). Earthquake event knowledge graph construction and reasoning. *Geomat. Nat. Hazards Risk* **15**(1), 2383768.
- Ross, Z. E., M. A. Meier, E. Hauksson, and T. H. Heaton (2018). Generalized seismic phase detection with deep learning. *Bull. Seismol. Soc. Am.* **108**(5A), 2894–2901.
- Ross, Z. E., D. T. Trugman, E. Hauksson, and P. M. Shearer (2019). Searching for hidden earthquakes in Southern California. *Science* **364**(6442), 767–771.
- Shelly, D. R. (2020). A high-resolution seismic catalog for the initial 2019 Ridgecrest earthquake sequence: Foreshocks, aftershocks, and faulting complexity. *Seismol. Res. Lett.* **91**(4), 1971–1978.
- Traag, V. A., L. Waltman, and N. J. van Eck (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233.
- Utsu, T., Y. Ogata, and R. S. Matsu'ura (1995). The centenary of the Omori formula for a decay law of aftershock activity. *J. Phys. Earth* **43**(1), 1–33.
- Yao, L., F. Ren, and K. Du (2025). From knowledge graph construction to retrieval-augmented generation: A framework for comprehensive earthquake emergency support. *Geo-Spat. Inf. Sci.* **29**(1).
- Yu, R., S. Luo, R. Ghosh, L. Li, Y. Xie, and X. Jia (2025). RAG for geoscience: What we expect, gaps and opportunities.
- Zhu, W. Q. and G. C. Beroza (2019). PhaseNet: A deep-neural-network-based seismic arrival-time picking method. *Geophys. J. Int.* **216**(1), 261–273.