Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GeoExtractor: A Framework for Structured Information Extraction from Geoscientific Literature

Zhong Peng
Zhejiang Lab
Ziqi Song
Zhejiang Lab
Yufei Ye
Zhejiang Lab
Shuang Li
Zhejiang Lab
Zongyuan Xiang
Zhejiang Lab
Jiang Yang
cscyangj@zhejianglab.org
Zhejiang Lab

Article

Keywords:

Posted Date: July 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10136696/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

GeoExtractor: A Framework for Structured Information Extraction from Geoscientific Literature

Zhong Peng[1+], Ziqi Song[1+], Yufei Ye[1], Shuang Li[1], Zongyuan Xiang[1], and Jiang Yang[1*]

[1]Earth and Space Science Computing Research Center, Zhejiang Lab, Hangzhou, 311100, China
[+]These authors contributed equally to this work.
[*]cscyangj@zhejianglab.org

ABSTRACT

Earth science research increasingly depends on the integration of large volumes of heterogeneous data distributed across publications. And a substantial amount of quantitative information remains embedded in unstructured literature, making large-scale data compilation labor-intensive and difficult to maintain. Here, we introduce GeoExtractor, a large language model (LLM)-powered multi-agent framework designed to automatically extract structured information from geoscience literature. GeoExtractor employs a hierarchical extraction strategy that combines an iterative decision–retrieval–generation workflow with a two-stage retrieval mechanism, enabling the recovery of complex multi-instance scientific records from heterogeneous document sources. We evaluate the framework on four benchmark datasets spanning diverse geoscience sub-disciplines. The framework consistently outperforms existing extraction methods, particularly in tasks involving multi-field extraction and complex scientific reasoning. To demonstrate its practical utility, we compiled a literature-derived database containing 2259 zircon analyses from 179 publications of the Central Asian Orogenic Belt (CAOB) and the resulting dataset reproduces previously recognized tectonic patterns previously recognized for the region. These results demonstrate that LLM-based multi-agent systems can facilitate the transformation of unstructured geoscientific literature into structured knowledge resources, substantially reducing the effort required for database construction and accelerating data-driven research across Earth science disciplines.

Introduction

Earth science research generates large volumes of quantitative information describing the composition, age, properties, and evolution of the Earth system. Although much of this information is reported in published articles, a substantial proportion remains embedded within unstructured text, tables, figures, and supplementary materials. Consequently, the reuse of published data often depends on labor-intensive manual compilation, limiting the efficiency of large-scale synthesis and comparative analyses across studies[1]. The rapid growth of geoscientific literature has further widened the gap between data production and data integration, creating a major challenge for emerging data-intensive approaches in Earth system science[2].

To improve data accessibility, several community database initiatives have been developed for specific geoscience domains, including PetDB[3], GEOROC[4,5], EarthChem[6]. However, they rely heavily on manual curation, cover only a fraction of the published literature, and often operate within disciplinary silos with limited interoperability[7,8]. Moreover, most existing databases are restricted to relatively homogeneous data types within individual sub-disciplines[9]. By contrast, cross-disciplinary quantitative knowledge remains overwhelmingly dispersed across unstructured literature, lacking unified data schemas and semantic standards, and is consequently difficult to systematically integrate and reuse[10]. As a result, the incorporation of published data continues to lag behind the pace of scientific publication, leaving large volumes of potentially valuable information inaccessible for computational analysis[11].

Compared with many scientific information extraction tasks, geoscientific knowledge is inherently organized around entities such as rock samples, stratigraphic units, fossil occurrences, drill cores, or geographic locations. These entities serve as the primary carriers of scientific observations, measurements, and interpretations, while the associated information is often distributed across multiple components of a publication[12,13]. Consequently, reconstructing a complete scientific record frequently requires integrating evidence from heterogeneous sources within a document rather than extracting information from isolated text passages[14,15]. Moreover, geoscientific records commonly exhibit hierarchical structures in which a single entity is associated with multiple attributes, measurements, analytical methods, uncertainties, and contextual metadata (e.g., GEOROC). Such information is rarely reported in a uniform format and may span different levels of granularity. These characteristics make direct document-level or sentence-level extraction insufficient for accurately recovering structured scientific knowledge. Instead, effective extraction requires an entity-centered process that can first identify target entities, then iteratively retrieve and assemble their associated attributes from distributed evidence sources. This challenge highlights the need for extraction

frameworks capable of reconstructing hierarchical scientific records rather than merely recognizing isolated entities or relations.

Recent advances in large language models (LLMs) have expanded the capabilities of automated scientific information extraction. LLM-based systems have demonstrated strong capabilities in tasks such as entity recognition, relation extraction, information synthesis, and structured data generation[16–25]. In parallel, retrieval-augmented generation (RAG) and iterative retrieval methods have shown promise for accessing distributed evidence across long documents and supporting multi-step reasoning[26–35]. Despite these advances, most existing approaches have been developed for question answering, document understanding, or general-purpose information retrieval. Their ability to reconstruct complex, entity-centered scientific records from heterogeneous geoscientific literature remains largely unexplored. In particular, approaches capable of generalizing across multiple Earth science sub-disciplines while producing structured, traceable, and verifiable outputs remain limited[36].

Here we present GeoExtractor, a modular multi-agent framework for automated extraction of structured knowledge from geoscience literature. GeoExtractor formulates information extraction as a hierarchical, entity-centered reconstruction process in which entities are first identified and then linked to their associated attributes through structured extraction. The framework integrates an iterative decision–retrieval–generation workflow together with a two-stage retrieval mechanism to enable robust recovery of complex scientific records from heterogeneous document sources. We evaluate GeoExtractor using a benchmark spanning four geoscience sub-disciplines (paleomagnetism, petroleum geology, geochemistry, and fire history), and compare its performance with representative extraction baselines. As a large-scale application, we further use GeoExtractor to construct a CAOB database containing zircon U–Pb–Hf–O and whole-rock Sr–Nd isotope data, demonstrating its potential for accelerating geoscientific data synthesis and knowledge discovery.

Results

The GeoExtractor Framework

GeGeoExtractor adopts a modular multi-agent architecture composed of six functional modules: document processing, primary key extraction, decision, retrieval, generation, and verification. The framework is built upon three key design principles. First, extraction is organized hierarchically. Rather than directly extracting all schema fields simultaneously, GeoExtractor first identifies primary-key instances that define individual records and then extracts the associated secondary-key fields for each instance. This decomposition transforms a complex record-level extraction task into a sequence of smaller and more manageable operations, reducing interference among heterogeneous fields and improving extraction stability in long scientific documents. Second, secondary key extraction is performed through an iterative decision–retrieval–generation loop. For each record, the decision module determines the next field to be extracted, the retrieval module gathers the most relevant evidence from the document, and the generation module produces the field value. The process is repeated until all required secondary-key fields have been completed. This field-level iterative strategy enables the framework to dynamically adapt to varying document structures and information distributions while maintaining consistent extraction quality across large schemas. Third, the retrieval module employs a two-stage retrieval mechanism. The first stage is designed to handle the majority of routine extraction scenarios through efficient semantic retrieval of locally relevant evidence. When the retrieved context is insufficient or the target field requires information distributed across multiple document sections, a second-stage retrieval process is activated to perform multi-hop evidence gathering and reasoning. This hierarchical retrieval strategy balances efficiency and coverage, allowing GeoExtractor to address both straightforward extraction tasks and more challenging cases that require cross-paragraph information integration.

Task definition. Geoscientific data are typically organized around domain-specific entities, each associated with a set of constrained attributes. Although research objects and terminology vary substantially across sub-disciplines, geoscientific datasets generally share a hierarchical entity–attribute structure. Given a geoscientific article together with user-defined field definitions, field descriptions, and parent–child relationships between fields, GeoExtractor produces structured records in which attribute values are explicitly linked to their corresponding entities. The extraction process begins by identifying domain entities that serve as primary-key instances within the target database. These entities function as unique identifiers for individual records. The framework then extracts the associated secondary-key attributes and establishes explicit links between attribute values and their parent entities. This formulation aligns the extraction procedure with the underlying structure of geoscientific datasets.For example, in the GEOROC database, a rock sample represents the primary-key entity, whereas attributes such as geographic location, Lu concentration, and Hf concentration correspond to secondary-key fields. GeoExtractor must therefore identify individual sample records while simultaneously assigning the correct attribute values to each sample.

Benchmark Evaluation

Benchmark design and construction

We constructed a cross-disciplinary geoscience information extraction benchmark to systematically evaluate the generalization capability of automated extraction methods across diverse sub-disciplines and heterogeneous data conditions. The benchmark

comprises four datasets representing paleomagnetism, petroleum geology, geochemistry, and fire history, thereby covering a broad range of geoscientific data types and reporting styles.

Table 1 summarizes the benchmark statistics. The datasets vary substantially in scale and structure: PaleoMag contains 316 records across 11 fields from 13 articles, while GCD contains only 3 records across 8 fields from 3 articles. This variation reflects the diversity of real-world extraction scenarios, ranging from publications containing hundreds of records to studies reporting only a few observations.

Table 1. Benchmark dataset summary statistics.

Dataset	Publications	Entity Records	Attribute Types	Attribute Values
PaleoMag	13	316	11	3,476
SweetSpot	10	197	16	3,152
GEOROC	7	125	8	1,000
GCD	3	3	8	24

The PaleoMag and SweetSpot datasets are derived from expert-curated compilations in paleomagnetism and petroleum exploration, respectively, and include high-quality ground-truth annotations verified by domain specialists. The GEOROC dataset was constructed by extracting a subset of records from the publicly available GEOROC Expert Datasets[37] and covers igneous rock samples with major and trace element compositions. The GCD dataset was derived from the Global Charcoal Database[38] and represents a challenging test case where each article corresponds to exactly one record, testing the method's robustness under low sample density conditions.

The benchmark encompasses publications with varying levels of structural and semantic complexity. Document lengths range from several to dozens of pages, and the number of fields to be extracted ranges from 8 to 16. In terms of modality, the benchmark includes unstructured text, structured and semi-structured tables, and figure captions, requiring methods to handle both text-based contextual reasoning and complex table understanding tasks such as large multi-page tables, hierarchical and cross-column headers, and semantic alignment between table bodies and annotation text.

Extraction performance

We evaluated GeoExtractor against two baseline methods: (1) a full-text end-to-end generation approach that prompts an LLM to extract all fields in a single pass from the entire document, and (2) LangExtract[39], an open-source, example-driven information extraction framework that relies on user-provided examples for in-context learning. All methods used DeepSeek-v4-flash as the base LLM to ensure fair comparison, with identical field definitions and prompt specifications. Detailed model configurations and extraction procedures are provided in the Methods section.

Table 2. Extraction performance of GeoExtractor compared to baseline methods across four benchmark datasets. All methods were implemented using DeepSeek-v4-flash as the underlying LLM.

Method	PaleoMag			SweetSpot			GEOROC			GCD		
	Recall	Prec	F1	Recall	Prec	F1	Recall	Prec	F1	Recall	Prec	F1
GeoExtractor	0.9748	0.9583	**0.9665**	0.7488	0.7979	**0.7726**	0.5434	0.8773	**0.6711**	0.8261	0.3654	0.5067
Full-text	0.8203	0.9840	0.8947	0.5807	0.8012	0.6733	0.5400	0.7649	0.6331	0.8261	0.7600	**0.7917**
LangExtract	0.3418	0.4119	0.3736	0.3744	0.2927	0.3286	0.3578	0.2362	0.2362	0.0500	0.5000	0.0909

Table 2 shows the precision, recall and F1 score for all methods across the four benchmark datasets. GeoExtractor achieved the highest F1 score on three of four datasets. The performance differences can be attributed to two aspects of the framework design. First, the iterative decision–retrieval–generation loop improves information coverage in multi-field tasks: the full-text baseline shows declining recall on datasets with many fields per document (0.5807 on SweetSpot; 0.8203 on PaleoMag), whereas GeoExtractor maintained high recall (0.7488 on SweetSpot; 0.9748 on PaleoMag). Second, the two-stage progressive retrieval mechanism facilitates multi-hop reasoning that is difficult to achieve using a single-pass full-text generation approach. We illustrate this with GEOROC's `province` field, which requires inferring the geological province of a sample through the chain `sample` → `sample location` → `province`: a two-hop association that cannot be resolved from any single passage. GeoExtractor extracted 71 of the 126 expected province values, whereas the full-text baseline extracted only 4, demonstrating GeoExtractor's ability to assemble evidence distributed across different sections of an article. These results suggest that generic example-driven extraction frameworks may be less effective for hierarchical geoscientific extraction tasks.

Error Analysis

To understand the nature and sources of extraction failures, we conducted a detailed error analysis on the GeoExtractor outputs, decomposing errors into omissions (missed values present in the source text) and wrong extractions (incorrect or hallucinated values). The dominant error type differed markedly across datasets. In GEOROC and SweetSpot, omissions overwhelmingly dominated, accounting for 90.6% and 84.1% of all errors, respectively, indicating that GeoExtractor frequently fails to locate and extract values that are present in the source text—particularly for descriptive fields such as geological province names, stratigraphic formations, and mineral compositions that require comprehension of free-text geological descriptions rather than extraction from standardized tables.

Field-level analysis further revealed a clear performance gradient with respect to field type. Highly structured numeric fields (latitude, longitude, and elevation) were consistently extracted with near-perfect accuracy across all datasets, benefiting from their standardized formatting and unambiguous semantics. In contrast, descriptive text fields (geological province names, stratigraphic formations, mineral compositions, and vegetation descriptions) showed the highest error rates. Notable examples include a 100% error rate for pressure coefficient and wax content in SweetSpot (both entirely missed across all records), where these fields were inconsistently formatted across articles and often embedded in running text rather than structured tables.

Application: Data compilation for the Central Asian Orogenic Belt

Database construction and tectonic background

To evaluate the practical utility of GeoExtractor for large-scale database construction, we applied the framework to compile a comprehensive isotopic–geochronological dataset for CAOB, one of the world's largest accretionary orogenic systems. A total of 179 CAOB-related publications were retrieved from the Web of Science and converted from PDF to markdown format using Table2LaTeX-RL[40]. GeoExtractor was then configured to extract 2259 zircon analyses, including zircon U–Pb ages, zircon Hf–O isotopic compositions, whole-rock Sr–Nd isotopic data, and associated sample metadata such as sample identifiers, geographic coordinates, rock types, and tectonic affiliations. Traditional compilation of isotopic and geochronological datasets typically requires researchers to manually inspect text, tables, figures, and supplementary materials and subsequently harmonize extracted information into structured formats. In contrast, GeoExtractor automatically reconstructed entity-centered records from heterogeneous document components through its iterative extraction workflow. On average, the framework processed a publication in approximately 15 minutes, compared with an estimated 8 hours required for manual extraction. For the 179-publication corpus, this corresponds to a reduction from roughly 1,400 hours of manual effort to less than 45 hours of automated processing time, demonstrating the potential of GeoExtractor to accelerate large-scale geoscientific database construction by nearly two orders of magnitude.

The CAOB[41–46], also known as the Altaids[47–49], is one of the world's largest and longest-lived accretionary orogens, extending over 5,000 km from the Ural Mountains in the west to the Pacific margin in the east[45,47]. It is bounded by the East European, Siberian, Karakum, Tarim, and North China cratons (Fig. 1) and comprises a complex collage of continental fragments, magmatic arcs, and subduction-accretion complexes assembled through multiple episodes of subduction, accretion, and collision from the Neoproterozoic to the Mesozoic[44,50,51]. Deciphering the crustal evolution of the CAOB requires systematic integration of vast quantities of isotopic and geochronological data dispersed across hundreds of individual publications—a task that has traditionally required years of manual compilation and that exemplifies the data synthesis bottleneck in modern Earth system science[52–54].

Temporal distribution of magmatism and crustal age structure

The zircon U-Pb age distribution (Fig. 2) reveals that the most intense magmatic activity in the CAOB is concentrated at 500-300 Ma, corresponding to the subduction and closure of the Paleo-Asian Ocean and subsequent arc and post-collisional magmatism. A secondary peak at 1000-800 Ma reflects the early stages of Paleo-Asian Ocean opening and microcontinent rifting. Sparse age records at 2000-1500 Ma document the early formation of microcontinent basements. This U-Pb age distribution is comparable to previous findings[55,56].

In contrast, whole-rock Nd model ages are predominantly concentrated at 1.2-0.8 Ga, significantly predating the zircon U-Pb crystallization ages and displaying almost no Archean peaks. This systematic offset indicates that Paleozoic magmas were not directly sourced from juvenile mantle melts but were primarily inherited from Neoproterozoic-Early Paleozoic juvenile accreted crust or metasomatized mantle wedge materials that had been extracted from the mantle substantially earlier than the magmatic crystallization events.

Zircon Hf model ages exhibit more extended Precambrian tails, reaching 3-4 Ga (Fig. 2), indicating incorporation of ancient crustal components into the magmatic systems. These ancient contributions likely originate from subducted sediment melting, eroded microcontinent fragments, or deep crustal reworking. The consistency between whole-rock-scale (Nd) and mineral-scale (zircon Hf) records—both pointing to a dominant juvenile crust contribution with subordinate ancient material—constitutes a robust, multi-scale isotopic signature of the CAOB's crustal architecture.

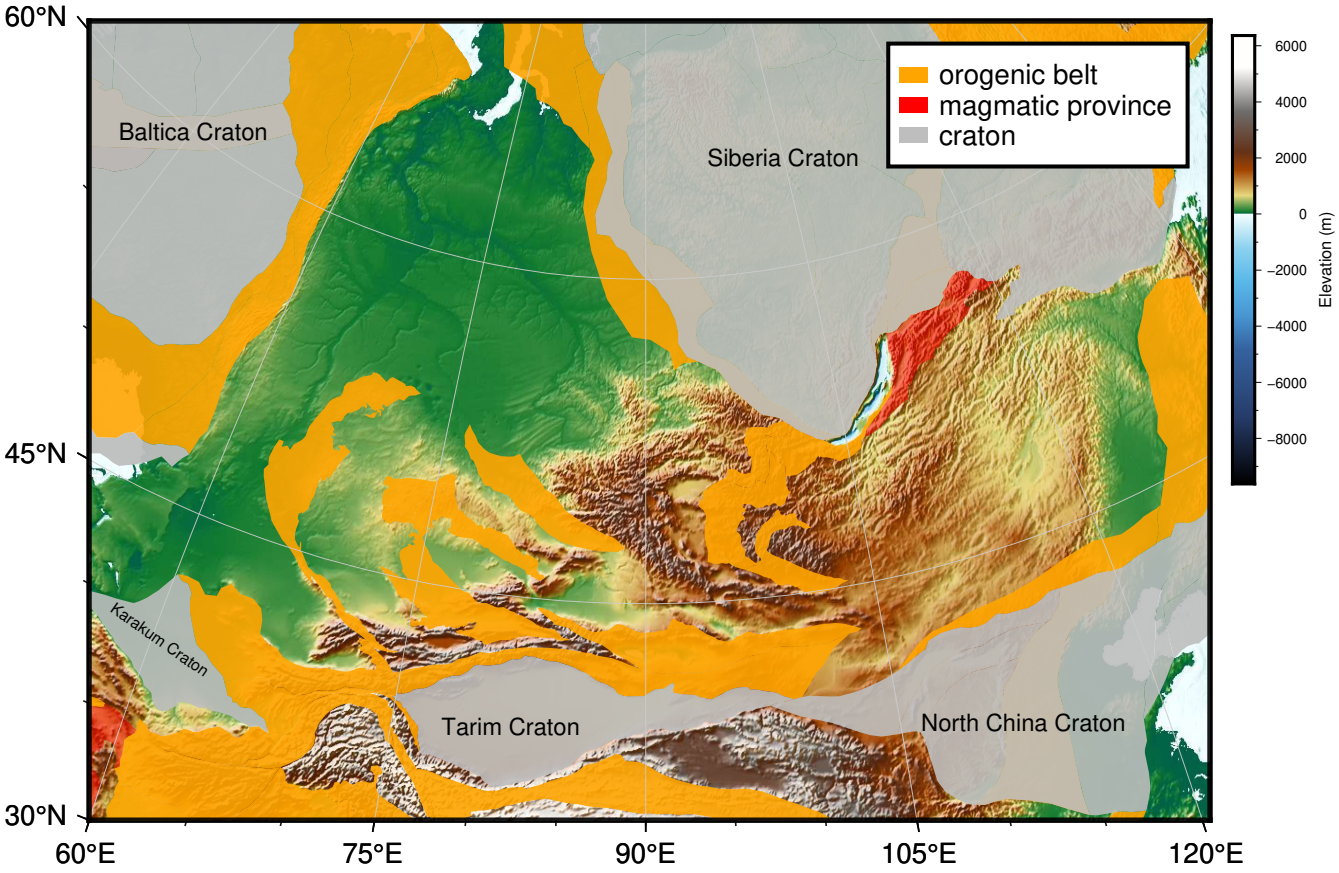


Figure 1. Tectonic map of CAOB, showing the distribution of major tectonic units, bounding cratons, and key suture zones. The CAOB extends from the Ural Mountains to the Pacific margin and is bounded by the East European, Siberian, Karakum, Tarim, and North China cratons. The orogen is composed of continental fragments, magmatic arcs, and subduction-accretion complexes. Topographic base map from NOAA National Geophysical Data Center.

Source mixing constrained by Sr-Nd isotopes

The $\varepsilon\mathrm{Nd}(t)$ values of CAOB igneous rocks span a wide range from $+8$ to -22 (Fig. 3), indicating a typical multi-endmember mixing system. The majority of samples cluster at $\varepsilon\mathrm{Nd}(t) = +3$ to $+9$, consistent with derivation from depleted mantle or juvenile accreted crust. Samples with $\varepsilon\mathrm{Nd}(t)$ near zero are temporally interleaved with positive-εNd samples, demonstrating that accreted crust underwent continuous reworking after its formation rather than experiencing a single rapid growth episode followed by quiescence.

Initial $^{87}\mathrm{Sr}/^{86}\mathrm{Sr}$ ratios are predominantly concentrated in the range of 0.703-0.707, with no systematic temporal evolution (Fig. 3). The absence of a secular trend toward more radiogenic values indicates that crustal growth and crustal recycling were synchronous processes throughout the Paleozoic, consistent with the sustained subduction-accretion regime that characterizes the CAOB. Sporadically elevated $^{87}\mathrm{Sr}/^{86}\mathrm{Sr}$ values record the localized involvement of more evolved continental components, reflecting the lateral heterogeneity inherent to accretionary orogens formed by the amalgamation of disparate terranes.

Crust-mantle interaction revealed by zircon Hf-O isotopes

Zircon $\varepsilon\mathrm{Hf}(t)$ values (Fig. 4) exhibit three distinct populations that jointly constrain the magma source characteristics. First, strongly positive values ($+10$ to $+20$) are concentrated at 200-500 Ma and indicate near-primitive mantle-derived melts or reworking of newly formed juvenile accreted crust. Second, intermediate values (0 to $+10$) reflect mixed juvenile-ancient crustal sources, recording the intense arc-continent interaction that typifies accretionary orogens. Third, markedly negative values (-10 to -50) are predominantly associated with ages of 1500-3500 Ma, suggesting contributions from Precambrian crustal components to the magmatic systems.

Zircon $\delta^{18}\mathrm{O}$ values are mainly concentrated at 5.5-8.5‰, slightly above the mantle zircon baseline of 5.3 ± 0.6‰, indicating that magmas were predominantly derived from juvenile crust or mantle systems with limited upper crustal contamination. A distinct $\delta^{18}\mathrm{O}$ plateau centered at 400-500 Ma marks the magmatic peak during the Devonian-Silurian, recording the climactic

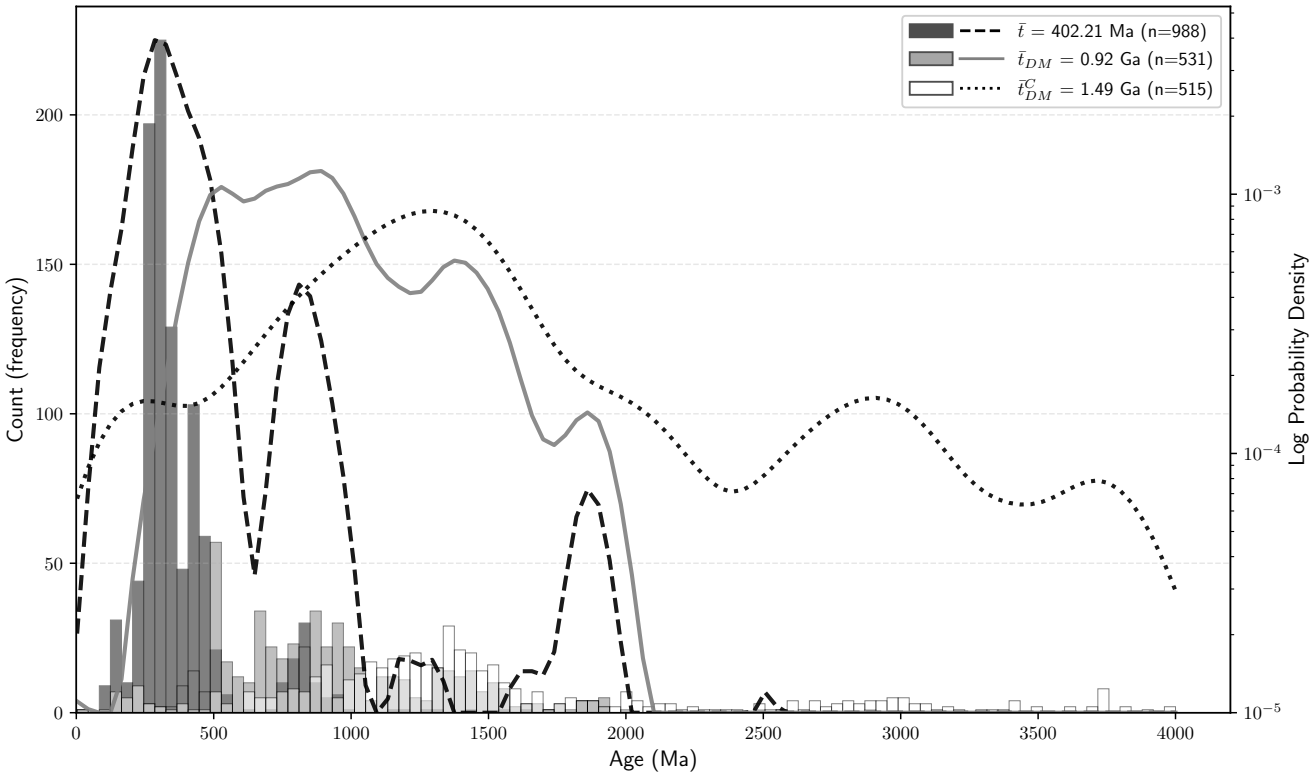


Figure 2. Histograms and probability density curves for the CAOB magmatic rocks. **a**, Zircon U-Pb ages, showing a dominant peak at 500-300 Ma, a secondary peak at 1000-800 Ma, and sparse ages at 2000-1500 Ma. **b**, Zircon Hf model ages, exhibiting Precambrian tails extending to 3-4 Ga. **c**, Whole-rock Nd model ages, concentrated at 1.2-0.8 Ga with minimal Archean contributions.

phase of Paleo-Asian Ocean subduction.

Integrated understanding of CAOB crustal evolution

Taken together, the large-scale isotopic database constructed via automated extraction supports the view that the CAOB is not a reactivated cratonic margin but a long-lived, multi-stage subduction-accretion system with continuous crustal reworking[44,47,50,52,53]. The integrated multi-isotope evidence reveals a coherent picture: Paleozoic magmatism, while concentrated at 300-500 Ma, was overwhelmingly sourced from Neoproterozoic juvenile accreted crust that formed during earlier stages of the orogenic cycle, with ongoing but subordinate incorporation of Precambrian ancient crustal materials through sediment subduction, microcontinent erosion, and deep crustal anatexis. This multi-stage crustal evolution—with juvenile crust production, rapid reworking, and ancient material recycling operating simultaneously—is fundamentally different from the episodic growth models proposed for cratonic settings.

This case study demonstrates the unique advantages of automated, cross-disciplinary literature knowledge extraction for reconstructing regional crustal evolution. Data that traditionally required years of manual compilation from hundreds of publications were assembled and analyzed in days, enabling systematic isotopic characterization at a scale previously impractical. The resulting database provides a unified foundation for testing quantitative tectonic models and for future data-driven exploration of CAOB metallogenic patterns.

Discussion

This study demonstrates that a modular LLM-based agent framework can effectively extract structured quantitative information from geoscientific literature without task-specific fine-tuning or extensive rule engineering. The effectiveness of GeoExtractor primarily arises from decomposing document-level extraction into hierarchical, field-centered sub-problems, which allows the model to localize reasoning at the level of individual attributes rather than processing entire documents in a single pass. This design, combined with an iterative decision–retrieval–generation process and a two-stage retrieval mechanism, enables adaptive

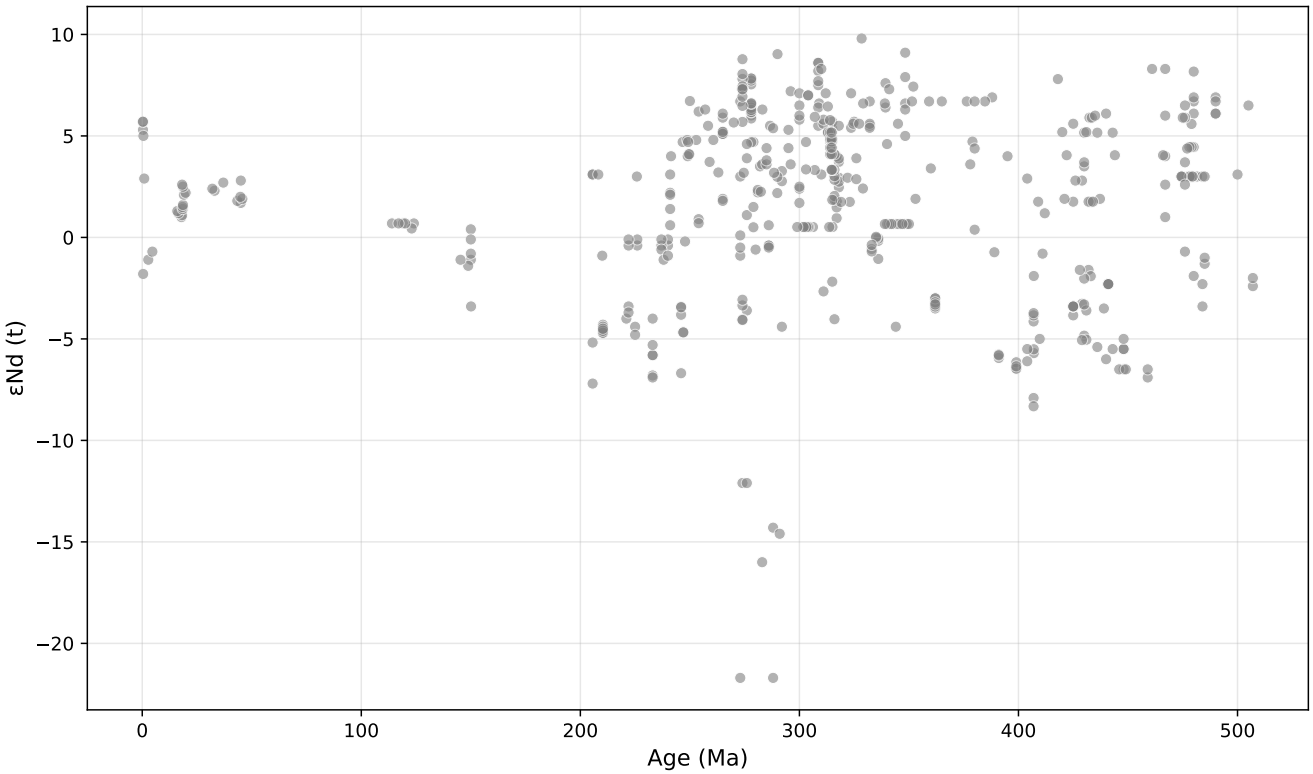


Figure 3. Sr-Nd isotopic data for igneous rocks from the Central Asian Orogenic Belt. **a**, $\varepsilon\mathrm{Nd}(t)$ versus age, showing values spanning $+8$ to -22 with the majority clustering at $+3$ to $+9$. **b**, Initial $^{87}\mathrm{Sr}/^{86}\mathrm{Sr}$ versus age, concentrated at 0.703-0.707 with no systematic temporal trend.

allocation of computational effort, with simple fields resolved through lightweight retrieval and complex fields triggering deeper multi-hop evidence aggregation.

Compared with end-to-end generation approaches and example-dependent systems such as LangExtract, GeoExtractor reduces reliance on global document understanding and curated demonstrations, instead leveraging field-driven control to generalize across geoscience sub-disciplines. The two-stage retrieval strategy further contributes to robustness by separating efficient coverage-oriented retrieval from selective reasoning-oriented retrieval, ensuring that computational resources are focused where they are most needed.

From a practical perspective, GeoExtractor substantially reduces the time required for literature-based data compilation, processing each article in approximately 15 minutes including retrieval and generation steps. While manual extraction remains more flexible in handling ambiguous cases, it typically requires several hours per article. These results suggest that automated extraction frameworks can significantly improve the scalability of geoscientific data synthesis, while still requiring human verification for quality control in downstream database construction.

Several limitations should be acknowledged. First, on datasets where each article contains only a single record (GCD), GeoExtractor sometimes over-generates candidate records, and the simple full-text baseline outperforms GeoExtractor by a substantial margin (F1 0.7917 vs. 0.5067). This limitation stems from the chunk-based processing paradigm: when the entire article describes a single entity, the iterative retrieval cycle may interpret distinct chunks as referring to distinct entities. A potential remedy is to incorporate a global entity-count estimation step before initiating the iterative extraction cycle. More broadly, this result demonstrates that GeoExtractor's advantages are most pronounced in multi-record scenarios—which constitute the majority of practical geoscience extraction tasks—and that alternative approaches may be preferable for single-record-per-article datasets. Second, the current benchmark covers four sub-disciplines of geoscience; validation on a broader range of domains, including sedimentology, hydrology, and atmospheric science, would further establish cross-disciplinary generalizability. Third, the ground-truth annotations for the benchmark were produced by a single annotator; inter-annotator agreement rates are therefore not available, and the annotation quality ceiling is unknown. Establishing multi-annotator ground truth with reported agreement metrics would strengthen the reliability of the benchmark. Fourth, the current framework processes text and tables but does not natively interpret figures, maps, or stratigraphic columns—modalities that contain

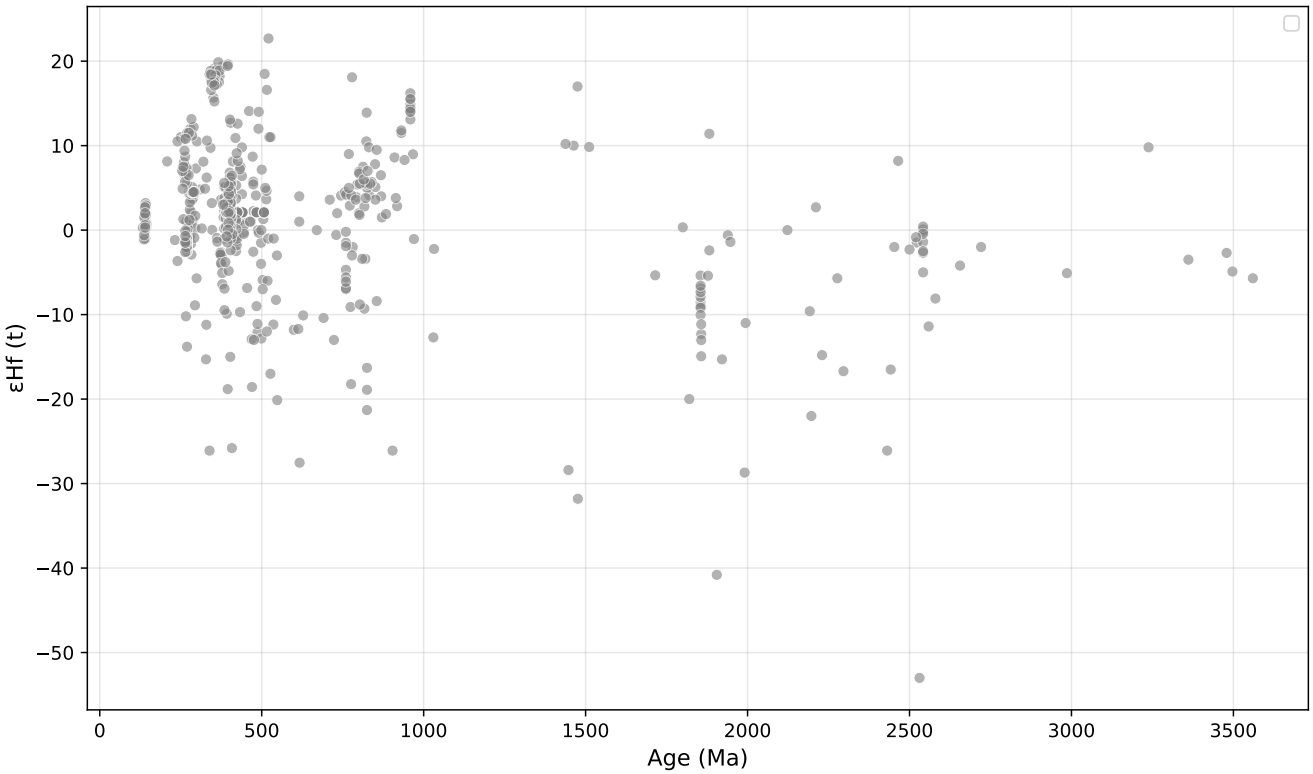


Figure 4. Zircon Hf-O isotopic data for CAOB magmatic rocks. **a**, $\varepsilon\mathrm{Hf}(t)$ versus U-Pb age, showing three populations: highly positive ($+10$ to $+20$, 200-500 Ma), intermediate (0 to $+10$), and strongly negative (-10 to -50, 1500-3500 Ma). **b**, $\delta^{18}\mathrm{O}$ versus age, with values concentrated at 5.5-8.5‰ and a plateau at 400-500 Ma.

substantial information in geoscience publications. Integrating vision-language models for multi-modal extraction represents an important direction for future work. Additionally, standard isotope geochemistry quality-control concerns—such as discordance filtering for zircon U-Pb ages, harmonization of Hf model age calculation parameters, and treatment of inter-laboratory calibration offsets—are not addressed in the current automated pipeline and represent important considerations for production use. Sixth, the LangExtract baseline comparison employs the framework in its simple-example configuration, results with domain-specific examples may differ. Finally, the extraction quality depends on the capabilities of the underlying LLM; improvements in base model reasoning, context length, and factual accuracy will directly benefit the framework.

Future work should pursue several directions. Integrating figure and table understanding capabilities, potentially through vision-language models, would enable true multi-modal extraction from geoscience publications. Incorporating domain-specific knowledge graphs could improve entity linking and cross-document data harmonization. Scaling the framework to operate on continuous publication streams through automatically extracting and updating databases as new papers appear, could enable living geoscience knowledge bases that grow with the literature. Extending the approach to related data-rich disciplines such as environmental science, soil science, and oceanography would further demonstrate the generality of the iterative agent extraction paradigm.

In summary, GeoExtractor establishes a practical methodology for converting the vast, unstructured knowledge embedded in geoscience literature into structured, queryable databases suitable for computational analysis. By enabling systematic, scalable literature data synthesis, this approach addresses a critical bottleneck in data-driven Earth system science and provides a foundation for next-generation geoscience knowledge infrastructure.

Methods

evaluation metrics

To quantitatively evaluate extraction performance, we employ precision, recall, and F1 score at the field level. The evaluation is conducted by comparing model outputs against human-annotated ground truth for each field value:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP (true positive) denotes correctly extracted field values, FP (false positive) denotes spuriously extracted values (hallucinated or incorrect), and FN (false negative) denotes missed field values that exist in the ground truth but were not extracted by the model. This field-level metric balances omission errors (recall) and hallucination errors (precision), providing a comprehensive measure of joint entity-attribute extraction capability. All experimental results in this study are reported using field-level F1 score.

GeoExtractor architecture

GeoExtractor is implemented as a modular multi-agent system. All LLM-driven modules use DeepSeek-v4-flash as the base model, accessed via API. The framework is model-agnostic and can incorporate alternative LLMs.

Document processing. Input articles are first converted to markdown format. For PDF sources, we use Table2LaTeX-RL[40] to preserve table structure during conversion. The markdown text is separated into text and table components. Text is segmented into chunks of 1,024 characters using a sliding window with no overlap. Tables are preserved as whole chunks regardless of size. All chunks are indexed and stored in a chunk library for retrieval.

Primary key extraction. The primary key extraction module processes the full article text in a single LLM call. The prompt instructs the LLM to identify all instances of the primary key, outputting each instance with a unique identifier and a brief natural-language description. For example, for a geochemistry dataset, the LLM identifies sample IDs and their contextual descriptions (rock type, location hints). The extracted primary keys serve as anchors for subsequent secondary key extraction.

Decision module. At each iteration, the decision module receives: (1) the current cumulative extraction results, (2) the list of target secondary key fields with their completion status, and (3) the set of remaining unmarked chunks. The module evaluates whether all target fields are sufficiently covered. If fields remain incomplete, it generates a natural-language query describing the information still needed. The query is passed to the retrieval module. Two termination conditions halt the iterative cycle: (1) n consecutive rounds ($n \leq 2$) produce no new incremental field values; (2) all chunks in the library have been marked as processed.

Retrieval module. The retrieval module combines sparse retrieval (BM25) and dense retrieval (text embeddings from a sentence transformer model). The query from the decision module is first rewritten by an LLM-based query rewriting model to optimize semantic coverage. The rewritten query is then issued to both retrievers in parallel. Retrieved results are merged by taking the union of the top-k results from each retriever ($k = 10$ for BM25, $k = 10$ for dense), with duplicates removed.

In Stage 1, all chunks retrieved and passed to the generation module are marked. Marked chunks are excluded from subsequent retrieval rounds (implemented via an ID-based filter on the retrieval results). This ensures that each retrieval round targets only previously unprocessed chunks.

In Stage 2, triggered when Stage 1 fails to achieve field completion, the retrieval module performs up to 3 additional rounds of progressive retrieval. In each round, the query and retrieved chunks are input to the LLM, which outputs: (a) important chunk indices within `<important_ids>` tags, identifying the most relevant chunks from the current retrieval set; (b) a refined query within `<query>` tags for the next round. The refined query drives the next round of hybrid retrieval. Important chunks from all Stage 2 rounds are accumulated and passed to the generation module.

Generation module. The generation module is a single LLM call that receives: (1) the query specifying which fields to extract, (2) the set of retrieved chunks, and (3) field definitions with descriptions. It outputs structured field values in JSON format. The generation is field-level: the LLM extracts only the specified target fields from the provided chunks, without access to historical extraction results or process state. After each generation round, results are merged with the cumulative extraction state by updating field values (new values added, conflicts resolved by keeping the most recently extracted value). The merged state is returned to the decision module.

Verification module. After the iterative cycle terminates, the final aggregated results and all marked text chunks are passed to the verification module. The LLM is prompted to check each extracted field value against the source chunks, outputting a binary verification result (grounded or ungrounded) for each field. Values flagged as ungrounded are removed from the final output. This post-hoc verification step reduces hallucination by filtering extractions that cannot be traced to source text.

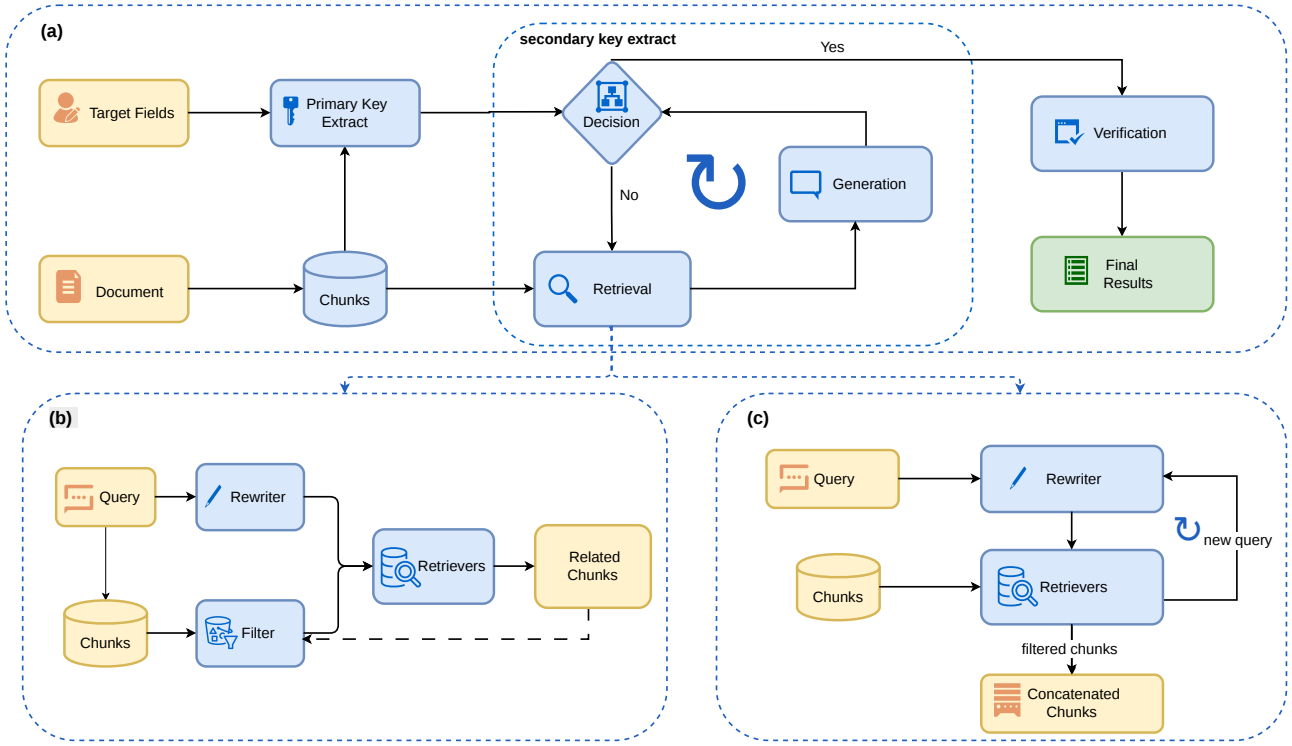


Figure 5. Architecture of the GeoExtractor framework. **a**, Overall modular agent architecture comprising six functional modules: document processing, primary key extraction, decision, retrieval, generation and validation. The decision module dynamically evaluates extraction completeness and controls iterative flow. **b**, Stage 1 retrieval: coverage-driven iterative retrieval with marked-chunk filtering to ensure non-redundant full-text coverage across iterations. **c**, Stage 2 retrieval: progressive multi-hop retrieval for complex fields requiring cross-paragraph reasoning, with LLM-driven query generation and important chunk extraction.

Ablation analysis

To quantify the contribution of each architectural component, we conducted an ablation study evaluating four configurations: Stage 1 only, Stage 1 + Stage 2, Stage 1 + Verification, and the final configuration (Stage 1 + Stage 2 + Verification). Results are reported in Table 3.

Stage 1 alone, providing coverage-driven iterative retrieval without progressive multi-hop retrieval or hallucination verification, achieved F1 score of 0.9325 (PaleoMag), 0.7623 (SweetSpot), 0.5305 (GEOROC), and 0.5067 (GCD). This demonstrates that the core iterative retrieval mechanism with chunk filtering already provides competitive performance on tasks dominated by simple field extraction.

Adding Stage 2 (progressive multi-hop retrieval[57]) improved F1 on GEOROC from 0.5305 to 0.6526, which is the largest relative improvement (+23.0%) among all datasets. On PaleoMag, adding Stage 2 also improved F1 from 0.9325 to 0.9680. These gains confirm that the progressive retrieval mechanism is particularly effective for fields requiring multi-hop reasoning, such as those in geochemical datasets where information is distributed across sections. The improvement was negligible on GCD (F1 unchanged at 0.5067), consistent with the single-record nature of that dataset.

Adding the verification module to Stage 1 improved F1 increase from 0.9325 to 0.9732. However, on SweetSpot, the combination of Stage 1 + Verification without Stage 2 resulted in lower F1 (0.6562 vs. 0.7623). This degradation occurs because the verification module may suppress correct but uncertain extractions when they are not corroborated by multiple chunks.

The full configuration (Stage 1 + Stage 2 + Verification) achieved the best overall balance, with particularly strong performance on PaleoMag (F1 = 0.9665) and GEOROC (F1 = 0.6711). The results demonstrate that each component contributes meaningfully to the overall system, with Stage 2 being essential for reasoning-intensive fields and the verification module providing additional performance for well-represented fields.

Table 3. Ablation study: contribution of each GeoExtractor component. F1 score reported for each configuration across benchmark datasets.

Configuration	PaleoMag	SweetSpot	GEOROC	GCD
Stage 1 only	0.9325	0.7623	0.5305	0.5067
Stage 1 + Stage 2	0.9680	0.7701	0.6526	0.5067
Stage 1 + Verification	0.9732	0.6562	0.5680	0.5067
Final (Stage 1 + Stage 2 + Verif.)	0.9665	0.7726	0.6711	0.5067

Baseline methods

Two baseline methods were implemented for comparison, both using DeepSeek-v4-flash as the base LLM to ensure fair comparison.

Full-text generation. The entire article text is provided to the LLM in a single prompt, together with field definitions identical to those used by GeoExtractor. The LLM is instructed to extract all field values in a single pass. For articles exceeding the model's context window, the text is truncated to the maximum allowable length.

LangExtract. We configured LangExtract[39], Google's open-source example-driven extraction framework, with the same field definitions. A simple geoscience-specific example (a table containing target fields) was provided; the framework was used with its default generic extraction configuration. This choice reflects the practical scenario in which a user applies a general-purpose extraction tool to a new domain without investing effort in domain-specific example curation—precisely the scenario that GeoExtractor is designed to address.

CAOB database construction

For the CAOB case study, we searched Web of Science for publications containing CAOB-related geochronology and isotope geochemistry data. Search terms included combinations of 'Central Asian Orogenic Belt', 'CAOB', 'Altaids', 'zircon U-Pb', 'Hf isotope', 'Nd isotope', and 'Sr isotope'. PDF files were obtained through institutional access and converted to markdown using Table2LaTeX-RL[40].

Data availability

The benchmark datasets and CAOB extracted database in this study have been deposited in the Zenodo database under accession code (https://doi.org/10.5281/zenodo.20822042).

Code availability

The source code for GeoExtractor is available from the corresponding author upon reasonable request.

References

1. Ma, X. Knowledge graph construction and application in geosciences: A review. *Comput. & Geosci.* **161**, 105082, DOI: https://doi.org/10.1016/j.cageo.2022.105082 (2022).
2. Reichstein, M. *et al.* Deep learning and process understanding for data-driven earth system science. *Nature* **566**, 195–204, DOI: https://doi.org/10.1038/s41586-019-0912-1 (2019).
3. Lehnert, K., Su, Y., Langmuir, C. H., Sarbas, B. & Nohl, U. A global geochemical database structure for rocks. *Geochem. Geophys. Geosystems* **1**, DOI: https://doi.org/10.1029/1999gc000026 (2000).
4. Kallas, L. *et al.* Advancing fair geochemical data: 25 years of georoc database service. *EGU Gen. Assembly 2025* DOI: https://doi.org/10.5194/egusphere-egu25-11467 (2025).
5. Wörner, G. *et al.* Georoc 2.0: A globally connected geochemical database to facilitate interdisciplinary, data-driven research. *EGU Gen. Assembly 2024* DOI: https://doi.org/10.5194/egusphere-egu24-15541 (2024).
6. Profeta, L., Lehnert, K., Song, L. & Figueroa, J. D. Earthchem communities: Building geochemical data best practices with researcher engagement. *EGU Gen. Assembly 2021* DOI: https://doi.org/10.5194/egusphere-egu21-16459 (2021).
7. Lehnert, K., Klöcking, M. *et al.* Overcoming fragmentation of geochemical data resources: Collaboration between earthchem, astromat, georoc, and metbase. *EGU Gen. Assembly 2023* DOI: https://doi.org/10.5194/egusphere-egu23-16683 (2023).

8. Klöcking, M. *et al.* Community recommendations for geochemical data, services and analytical capabilities in the 21st century. *Geochimica et Cosmochimica Acta* **351**, 66–77, DOI: https://doi.org/10.1016/j.gca.2023.04.019 (2023).

9. Li, Z.-X., Eglington, B. & Wang, T. Toward a big-data approach for reconstructing regional to global paleogeography and tectonic histories: Preface. *Earth-Science Rev.* **261**, 105030, DOI: https://doi.org/10.1016/j.earscirev.2024.105030 (2025).

10. Mahecha, M. D. *et al.* Earth system data cubes unravel global multivariate dynamics. *Earth Syst. Dyn.* **11**, 201–234, DOI: https://doi.org/10.5194/esd-11-201-2020 (2020).

11. Zhou, C. *et al.* Geoscience knowledge graph in the big data era. *Sci. China Earth Sci.* **64**, 1105–1114, DOI: https://doi.org/10.1007/s11430-020-9750-4 (2021).

12. Yu, P., Deng, C., Ji, H. & Wen, Y. Utilizing large language models for geoscience literature information extraction. *EGU Gen. Assembly 2024* DOI: https://doi.org/10.5194/egusphere-egu24-16101 (2024).

13. Guo, Z. *et al.* Sm-nd isotope data compilation from geoscientific literature using an automated tabular extraction method. *Sci. Data* **12**, 203, DOI: https://doi.org/10.1038/s41597-024-04229-5 (2025).

14. Qiu, Q., Duan, Y., Ma, K., Tao, L. & Xie, Z. Information extraction and knowledge linkage of geological profiles and related contextual texts from mineral exploration reports for geological knowledge graphs construction. *Ore Geol. Rev.* **163**, 105739, DOI: https://doi.org/10.1016/j.oregeorev.2023.105739 (2023).

15. Feng, Q., Zhao, T. & Liu, C. A "pipeline"-based approach for automated construction of geoscience knowledge graphs. *Minerals* **14**, DOI: https://doi.org/10.3390/min14121296 (2024).

16. Song, Z., Hwang, G.-Y., Zhang, X., Huang, S. & Park, B.-K. A scientific-article key-insight extraction system based on multi-actor of fine-tuned open-source large language models. *Sci. Reports* **15**, DOI: https://doi.org/10.1038/s41598-025-85715-7 (2025).

17. Jami, H. C. *et al.* Ccu-llama: A knowledge extraction llm for carbon capture and utilization by mining scientific literature data. *Ind. & Eng. Chem. Res.* DOI: https://doi.org/10.1021/acs.iecr.4c01656 (2024).

18. Duan, D., Peng, J., Zhang, Y. & Zhang, C. Scinlp: A domain-specific benchmark for full-text scientific entity and relation extraction in nlp. In *Proceedings of EMNLP 2025*, 14484–14497, DOI: https://doi.org/10.18653/v1/2025.emnlp-main.732 (Association for Computational Linguistics, 2025).

19. Wang, X., Huey, S. L., Sheng, R., Mehta, S. & Wang, F. Scidasynth: Interactive structured data extraction from scientific literature with large language model. *Campbell Syst. Rev.* **21**, DOI: https://doi.org/10.1002/cl2.70073 (2025).

20. Ateia, S., Kruschwitz, U., Scholz, M., Koschmider, A. & Almohaishi, M. Llm-based information extraction to support scientific literature research and publication workflows. In *International Conference on Theory and Practice of Digital Libraries (TPDL 2025)*, 90–99, DOI: https://doi.org/10.1007/978-3-032-06136-2_9 (Springer Nature Switzerland, 2025).

21. Li, S. *et al.* Exploring llms for scientific information extraction using the sciex framework. Preprint at https://arxiv.org/abs/2512.10004 (2025).

22. Swain, M. C. & Cole, J. M. Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904, DOI: https://doi.org/10.1021/acs.jcim.6b00207 (2016).

23. Mavrač. Chemdataextractor 2.0: Autopopulated ontologies for materials science. *J. Chem. Inf. Model.* **61**, 4280–4289, DOI: https://doi.org/10.1021/acs.jcim.1c00446 (2021).

24. Li, J., Li, M., Yang, Q. & Luo, S. Reactionseek: Llm-powered literature data mining and knowledge discovery in organic synthesis. *Nat. Commun.* **17**, DOI: https://doi.org/10.1038/s41467-026-70180-1 (2026).

25. Zhang, W. *et al.* Fine-tuning large language models for chemical text mining. *Chem. Sci.* **15**, 10600–10611, DOI: https://doi.org/10.1039/d4sc00924j (2024).

26. Asai, A., Wu, Z., Wang, Y., Sil, A. & Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. Preprint at https://arxiv.org/abs/2310.11511 (2023).

27. Zhuang, Z. *et al.* Efficientrag: Efficient retriever for multi-hop question answering. Preprint at https://arxiv.org/abs/2408.04259 (2024).

28. Li, R. *et al.* Knowtrace: Bootstrapping iterative retrieval-augmented generation with structured knowledge tracing. In *Proceedings of KDD 2025*, 1470–1480, DOI: https://doi.org/10.1145/3711896.3737015 (ACM, 2025).

29. Zhao, P. *et al.* Retrieval-augmented generation for ai-generated content: A survey. Preprint at https://arxiv.org/abs/2402.19473 (2024).

30. Lin, H. *et al.* Fishing for answers: Exploring one-shot vs. iterative retrieval strategies for retrieval augmented generation. Preprint at https://arxiv.org/abs/2509.04820 (2025).

31. Li, J. *et al.* Tree of reviews: A tree-based dynamic iterative retrieval framework for multi-hop question answering. Preprint at https://arxiv.org/abs/2404.14464 (2024).

32. Jang, Y., Lee, K.-i., Bae, H., Lee, H. & Jung, K. Itercqr: Iterative conversational query reformulation with retrieval guidance. Preprint at https://arxiv.org/abs/2311.09820 (2023).

33. Dhole, K. D., Chandradevan, R. & Agichtein, E. Generative query reformulation using ensemble prompting, document fusion, and relevance feedback. Preprint at https://arxiv.org/abs/2405.17658 (2024).

34. Fang, J., Meng, Z. & Macdonald, C. Kirag: Knowledge-driven iterative retriever for enhancing retrieval-augmented generation. Preprint at https://arxiv.org/abs/2502.18397 (2025).

35. Singh, A., Ehtesham, A., Kumar, S., Khoei, T. T. & Vasilakos, A. V. Agentic retrieval-augmented generation: A survey on agentic rag. Preprint at https://arxiv.org/abs/2501.09136 (2025).

36. Yu, R. *et al.* Rag for geoscience: What we expect, gaps and opportunities. Preprint at https://arxiv.org/abs/2508.11246 (2025).

37. Ball, P. W. Global distribution and composition of neogene-quaternary intraplate volcanic rocks, DOI: https://doi.org/10.5880/DIGIS.E.2024.006 (2022).

38. Power, M. J., Marlon, J. R., Bartlein, P. J. & Harrison, S. P. Fire history and the global charcoal database: A new tool for hypothesis testing and data exploration. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **291**, 52–59, DOI: https://doi.org/10.1016/j.palaeo.2009.09.014 (2010).

39. Goel, A. Langextract. *Zenodo* DOI: https://doi.org/10.5281/zenodo.17015090 (2025).

40. Ling, J. *et al.* Table2latex-rl: High-fidelity latex code generation from table images via reinforced multimodal language models. Preprint at https://arxiv.org/abs/2509.17589 (2025).

41. Jahn, B.-m., Wu, F. & Chen, B. Granitoids of the central asian orogenic belt and continental growth in the phanerozoic. *Earth Environ. Sci. Transactions The Royal Soc. Edinb.* **91**, 181–193, DOI: https://doi.org/10.1017/S0263593300007367 (2000).

42. Khain, E. V. *et al.* The palaeo-asian ocean in the neoproterozoic and early palaeozoic: new geochronologic data and palaeotectonic reconstructions. *Precambrian Res.* **122**, 329–358, DOI: https://doi.org/10.1016/S0301-9268(02)00218-8 (2003).

43. Xiao, W., Windley, B. F., Hao, J. & Zhai, M. Accretion leading to collision and the permian solonker suture, inner mongolia, china: Termination of the central asian orogenic belt. *Tectonics* **22**, 2002TC001484, DOI: https://doi.org/10.1029/2002TC001484 (2003).

44. Xiao, W. *et al.* A tale of amalgamation of three permo-triassic collage systems in central asia: Oroclines, sutures, and terminal accretion. *Annu. Rev. Earth Planet. Sci.* **43**, 477–507, DOI: https://doi.org/10.1146/annurev-earth-060614-105254 (2015).

45. Windley, B. F., Alexeiev, D., Xiao, W., Kröner, A. & Badarch, G. Tectonic models for accretion of the central asian orogenic belt. *J. Geol. Soc.* **164**, 31–47, DOI: https://doi.org/10.1144/0016-76492006-022 (2007).

46. Han, Y. & Zhao, G. Final amalgamation of the tianshan and junggar orogenic collage in the southwestern central asian orogenic belt: Constraints on the closure of the paleo-asian ocean. *Earth-Science Rev.* **186**, 129–152, DOI: https://doi.org/10.1016/j.earscirev.2017.09.012 (2018).

47. Ş. Evolution of the altaid tectonic collage and palaeozoic crustal growth in eurasia. *Nature* **364**, 299–307, DOI: https://doi.org/10.1038/364299a0 (1993).

48. Yakubchuk, A. Architecture and mineral deposit settings of the altaid orogenic collage: a revised model. *J. Asian Earth Sci.* **23**, 761–779, DOI: https://doi.org/10.1016/j.jseaes.2004.01.006 (2004).

49. Wilhem, C., Windley, B. F. & Stampfli, G. M. The altaids of central asia: A tectonic and evolutionary innovative review. *Earth-Science Rev.* **113**, 303–341, DOI: https://doi.org/10.1016/j.earscirev.2012.04.001 (2012).

50. Xiao, W. *et al.* Paleozoic multiple subduction-accretion processes of the southern altaids. *Am. J. Sci.* **309**, 221–270, DOI: https://doi.org/10.2475/03.2009.02 (2009).

51. Xiao, W. *et al.* Paleozoic multiple accretionary and collisional processes of the beishan orogenic collage. *Am. J. Sci.* **310**, 1553–1594, DOI: https://doi.org/10.2475/10.2010.12 (2010).

52. Jahn, B.-m., Wu, F.-Y. & Chen, B. Massive granitoid generation in central asia: Nd isotope evidence and implication for continental growth in the phanerozoic. *Episodes* **23**, 82–92, DOI: https://doi.org/10.18814/epiiugs/2000/v23i2/001 (2000).

53. Kröner, A. *et al.* Reassessment of continental growth during the accretionary history of the central asian orogenic belt. *Gondwana Res.* **25**, 103–125, DOI: https://doi.org/10.1016/j.gr.2012.12.023 (2013).

54. Safonova, I. *et al.* A new concept of continental construction in the central asian orogenic belt. *Episodes* **34**, 186–196, DOI: https://doi.org/10.18814/epiiugs/2011/v34i3/005 (2011).

55. Wang, L. *et al.* Detrital zircon u–pb and hf isotopic dataset for the central asian orogenic belt, northern china. *Geosci. Data J.* **11**, 426–432, DOI: https://doi.org/10.1002/gdj3.214 (2024). https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/gdj3.214.

56. Cao, G. & Chen, H. Monazite as an indicator for orogenic gold deposits: Constraints from detrital zircon th/u ratio and isotope data in the central asian orogenic belt. *Geosci. Front.* **16**, 102155, DOI: https://doi.org/10.1016/j.gsf.2025.102155 (2025).

57. Jiang, P. *et al.* s3: You don't need that much data to train a search agent via rl. Preprint at https://arxiv.org/abs/2505.14146 (2025).

Acknowledgements

We thank the maintainers of GEOROC, and GCD for making their datasets available.

Funding

This work is supported by the Deep Earth Probe and Mineral Resources Exploration - National Science and Technology Major Project (Project No.2024ZD1001203).

Author contributions statement

J. Y. and Z. S. conceived and designed the study. Z. P. performed all experimental tasks, analyzed the data, and drafted the manuscript. Z. S. and J. Y. contributed to the study design, manuscript writing, and critical revision. Y. Y., S. L., and Z. X. developed the core codes and performed the software implementation. All authors have read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional Information

The prompt used for the full-text baseline extraction method is as follows:

> You are an assistant for information-extracting tasks. Use the following retrieved context to extract the value of fields.
> Context:
> {context}
> Question: Extract {primary_name} (primary key) and corresponding {num} attributions: {attributions}.
> The meaning of each fields are as follows:
> {descriptions}
> for each primary key, extract the attribute values (keep the raw form in the context) one by one. processing method:
> - please check the data one by one to ensure there is no misalignment!
> - for sample attributions that don't exist in the context, return "don't know",
> - The final output format is as follows:
> {output_format}