Computers & Geosciences 214 (2026) 106197

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

GraphRAG and LLM-driven semantic exploration of critical mineral data

Armita Davarpanah [a,*], Hassan A. Babaie [b], W. Crawford Elliott [b], Yuanzhi Tang [c], Paul A. Schroeder [d]

[a] Environmental and Health Sciences Department, Spelman College, Atlanta, GA, 30314, USA
[b] Department of Geosciences, Georgia State University, Atlanta, GA, 30302, USA
[c] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA
[d] Department of Geology, Franklin College of Arts and Sciences, University of Georgia, Athens, GA, 30602, USA

ARTICLE INFO

Keywords:
Critical minerals
Knowledge graph
GraphRAG
Vector embeddings
Semantic retrieval

ABSTRACT

The identification of patterns in the distribution and relationships of critical minerals is fundamentally challenging due to the heterogeneity, scale, and semantic complexity of geoscience data. This study introduces a deterministic-first hybrid AI architecture that extends beyond standard retrieval-augmented generation (RAG) by explicitly separating symbolic reasoning, neural semantic retrieval, and generative synthesis into complementary and interpretable reasoning pathways. Unlike conventional RAG systems that rely primarily on text-based retrieval and opaque large language model (LLM) inference, our framework integrates ontology-grounded knowledge graph querying for authoritative reasoning, vector-based semantic retrieval for contextual generalization, and LLM synthesis for natural-language interaction. The architecture is instantiated through the construction of a large-scale Critical Minerals Knowledge Graph (CMKG) derived from the Critical Minerals in Ores (CMiO) dataset, comprising more than 29,000 sample records and implemented in Neo4j using a strict ontology and modular batching pipeline. Deterministic Cypher queries provide precise answers when explicit semantic relationships are available, while dense vector embeddings indexed with FAISS supply relevant context when exact matches are unavailable. An interactive Jupyter interface exposes these reasoning modes side-by-side to support transparency and interpretability. Evaluation using structured benchmark queries demonstrates that deterministic graph querying provides reliable and explainable results for structured scientific questions, while semantic and generative components contribute primarily to contextual support and interpretation. By unifying symbolic and neural reasoning within a controlled GraphRAG framework, this work contributes a generalizable AI architecture for transparent question answering over complex scientific knowledge graphs, with direct implications for critical mineral analysis and beyond.

1. Introduction

Critical minerals are essential to modern technologies, renewable energy systems, and national security, underpinning batteries, electronics, and clean energy infrastructure across global supply chains (Hofstra and Kreiner, 2020; Jowitt et al., 2020; Fortier et al., 2021; Schroeder et al., 2024; USGS, 2022, 2025; Zhu et al., 2025). Their roles have made them central to advanced manufacturing, decarbonization strategies, strategic resource planning, and sustainable development (USGS, 2022; UN, 2025). Among these, rare earth elements (REE), comprising the lanthanide series together with yttrium and scandium, represent a strategically important subset of critical minerals due to their indispensable technological applications and heightened supply-chain vulnerability. REE are commonly subdivided into light rare earth elements (LREE; La-Sm) and heavy rare earth elements (HREE; Gd-Lu, typically including yttrium owing to its similar geochemical behavior); a distinction that reflects differences in ionic radius, geochemical behavior, and economic significance across mineral systems (Haxel et al., 2002; Fortier et al., 2018). However, the discovery and characterization of the occurrences and relationships of critical mineral systems from large, heterogeneous databases remains a significant challenge due to the inherent complexity and variability of geoscience data (Hofstra and Kreiner, 2020; Hofstra et al., 2021). Geological datasets often span multiple spatial scales from atom-scale mineralogical observations to deposit-level attributes. These properties are stored in disparate formats, making integration and interpretation difficult

* Corresponding author.
E-mail address: adavarpa@spelman.edu (A. Davarpanah).

https://doi.org/10.1016/j.cageo.2026.106197
Received 24 January 2026; Received in revised form 16 May 2026; Accepted 18 May 2026
Available online 22 May 2026

(Zhou et al., 2021; Chen et al., 2024). Traditional approaches rely on manual data analysis or isolated relational databases, which limit the ability to uncover hidden relationships and generate actionable insights (Jatana et al., 2012).

Recent advances in artificial intelligence (AI) and knowledge representation offer new opportunities to address these challenges. Knowledge Graphs (KGs) provide a structured way to represent entities and their relationships, enabling semantic integration of diverse datasets (Ma, 2022; Qiu et al., 2023; Chen et al., 2024; Feng et al., 2024; Zhu et al., 2025). Large Language Models (LLMs) excel at natural language understanding and reasoning but lack direct access to domain-specific structured data and may produce ungrounded or non-reproducible outputs when used in isolation (Lin et al., 2024; Zhou and Li, 2025). Retrieval-Augmented Generation (RAG) bridges this gap by combining retrieval mechanisms with generative probabilistic models, allowing context-aware responses grounded in deterministic external knowledge sources (Peng et al., 2024; Han et al., 2025). However, traditional RAG typically relies on unstructured text retrieval, which limits its ability to exploit rich, structured relationships in multi-relational datasets. Graph-based Retrieval-Augmented Generation (GraphRAG) surpasses traditional RAG by leveraging graph-based reasoning for richer context, multi-hop reasoning, and improved interpretability. Despite these advances, most existing applications in geoscience focus on either graph-based reasoning or text-based AI, without fully leveraging their combined potential or clearly separating deterministic retrieval from generative synthesis (e.g., Ma et al., 2025).

Prior efforts in geoscience data integration have explored ontology-driven systems, relational databases, and standalone knowledge graphs to improve data interoperability (Brantley et al., 2021). Ontology-based approaches provide a strong semantic structure with clearly defined classes, relationships, and rules (Babaie et al., 2022; Davarpanah et al., 2024). However, they depend on rigid, predefined schemas and require structured queries, which can make them less flexible when dealing with evolving or heterogeneous data. Relational database systems remain widely used in geoscience (e.g., EarthChem Portal, 2025; Macrostrat Database, 2025), yet they handle heterogeneity poorly and limit flexible, multi-relational exploration (Jatana et al., 2012. Large-scale geoscience knowledge graphs (Chen et al., 2024; Zhu et al., 2025) enhance interoperability but lack natural-language interfaces and explanatory reasoning capabilities. Similarly, recent applications of LLMs in scientific domains demonstrate strong language capabilities but struggle with factual grounding and domain-specific accuracy when operating without structured data (Mousavi et al., 2025; Lin et al., 2024). Standard RAG implementations predominantly target unstructured text corpora and do not exploit graph-based structures for enhanced retrieval and reasoning (Lewis et al., 2020). These limitations highlight the need for a hybrid approach that combines the semantic richness of knowledge graphs with the generative power of LLMs, while ensuring that generative components remain grounded in deterministic, verifiable data sources. Emerging GraphRAG methodologies demonstrate the effectiveness of integrating graph-structured knowledge into retrieval pipelines, improving complex question answering and factual grounding (Peng et al., 2024). Yet, their application in scientific knowledge systems, particularly in geoscience, remains in early stages (Zhang et al., 2025).

This work introduces an AI-driven framework that integrates Knowledge Graphs with GraphRAG to enable intelligent exploration of critical mineral datasets. A Critical Minerals Knowledge Graph (CMKG) transforms tabular critical minerals data into an interconnected network, modeling relationships among samples, minerals, rocks, alterations, textures, deposits, environments, and countries. Built on CMKG, the proposed framework follows a deterministic-first hybrid architecture in which knowledge graph queries provide authoritative results, semantic retrieval supplies contextual support, and LLMs are used strictly for synthesis of retrieved evidence. This design ensures both precision and interoperability while enabling flexible natural language interaction. The system supports both structured queries and free-form input, delivering deterministic lookups, semantic retrieval, and synthesized responses for comprehensive insights.

The significance of this approach lies in bridging heterogeneous geoscience datasets with actionable, semantically grounded insights, addressing global challenges in sustainable resource discovery and supply chain security (Hofstra and Kreiner, 2020). Applications span mineral, academic research, and policy development, enabling stakeholders to extract insights from complex datasets efficiently. In addition, the framework supports analytical tasks such as identifying rare earth element-enriched deposits and multi-element associations, demonstrating its value for data-driven exploration. The innovation stems from merging graph-based reasoning with retrieval-augmented generation under a controlled and interpretable architecture, creating a dynamic knowledge ecosystem that retrieves relevant information and synthesizes it into coherent, context-rich explanations. By transforming static datasets into interactive, analytical environments, this platform accelerates discovery and supports informed decision-making in critical mineral systems (Zhang et al., 2025).

1.1. Related work

Recent research in artificial intelligence has explored integrating structured knowledge with neural and generative models to improve question answering and reasoning. Neuro-symbolic approaches aim to combine symbolic representations with neural models. Bosselut et al. (2019) propose COMET, which leverages pre-trained language models to generate new commonsense knowledge for loosely structured graphs such as ATOMIC and ConceptNet, demonstrating that generative models can support scalable knowledge graph completion. However, COMET focuses on text-based commonsense knowledge generation rather than deterministic reasoning over structured scientific data. Other works emphasize joint representation learning from text and knowledge graphs. For instance, Yasunaga et al. (2022) introduce DRAGON, a deep bidirectional pretraining framework that fuses language modeling with knowledge graph link prediction, achieving strong performance on question answering and complex reasoning tasks across multiple domains. While effective for representation learning, DRAGON does not address interactive querying, explicit graph execution, or provenance-aware reasoning over structured scientific knowledge graphs.

Retrieval-augmented generation (RAG) has become a dominant paradigm for grounding large language models in external knowledge, typically relying on vector-based retrieval over unstructured text (Lewis et al., 2020). Graph-enhanced variants extend this paradigm by incorporating structural information. LightRAG (Guo et al., 2024) integrates lightweight graph structures into text indexing and retrieval to improve contextual awareness and efficiency in LLM-based question answering but remains largely LLM-centric and focuses on retrieval quality rather than explicit symbolic reasoning. More broadly, GraphRAG frameworks formalize the integration of graph-structured data into retrieval-augmented generation pipelines. Han et al. (2025) provide a systematic overview of GraphRAG architecture, outlining key components for combining graph-based retrieval with LLM generation to support multi-hop reasoning and relational context. While GraphRAG establishes an important conceptual foundation, it primarily serves as a general framework and does not enforce deterministic query execution, numeric reasoning, or ontology-driven guarantees required for scientific domains.

In the geoscience domain, ontology-driven and large-scale knowledge graphs have been developed to improve semantic standardization and data interoperability. Zhu et al. (2025) present a global, ontology-driven geoscience knowledge graph for deep-time Earth research, emphasizing large-scale integration rather than task-specific AI reasoning or natural-language interaction. Similarly, semantic retrieval techniques based on dense vector embedding have been

applied to geoscientific text to enable similarity-based search; however, while effective for contextual retrieval, these approaches often lack precision for entity-specific queries. Large language models have also been explored for natural language interaction with scientific data, but LLM-only approaches remain prone to hallucination and lack transparency in their reasoning processes when not grounded in structured data.

The present work differs from these approaches in three key ways: (i) Ontology-grounded scientific KG with numeric attributes: Unlike prior approaches that rely primarily on text-derived or dynamically generated graphs, our framework operates over a strict, ontology-driven Critical Minerals Knowledge Graph (CMKG) built from structured geoscience datasets, including numeric and categorical attributes (e.g., ppm concentrations, deposit types). (ii) Deterministic-first hybrid reasoning: We explicitly separate deterministic Cypher-based KG querying (authoritative and exact) from semantic vector retrieval (coverage when exact matches fail), followed by LLM synthesis used strictly for integrating retrieved evidence. This contrasts with KG-RAG systems that emphasize LLM-centric evidence selection without a clear deterministic-first contract. (iii) Interpretability by design: Our architecture exposes side-by-side outputs from symbolic, neural, and generative components, ensuring provenance and reasoning transparency; features rarely prioritized in prior KG-RAG implementations.

2. Methodology

The project develops the Critical Minerals Knowledge Graph (CMKG), an integrated framework that combines sample-level geochemical, mineralogical, and lithological data with various deposit-level data from the CMiO (Critical Minerals in Ores) database. CMKG enables intelligent, semantic question answering by unifying vector embeddings, graph databases, and LLMs. Through this integration of structured and unstructured data, the system supports complex, domain-specific reasoning for critical mineral exploration. The framework follows a deterministic-first hybrid architecture in which structured graph queries provide authoritative results, semantic retrieval supplies contextual support, and LLMs are used strictly for synthesis of retrieved evidence. The methodology is implemented through a Jupyter-Python pipeline comprising the following major stages: (1) Data acquisition and preparation; (2) Vector embedding and semantic indexing; (3) Knowledge graph construction; (4) Development of an interactive user interface for Question Answering (QA); (5) Hybrid Graph Retrieval-Augmented Generation (GraphRAG) QA.

2.1. Data acquisition and preparation

The CMiO dataset was obtained as a CSV file from the Australian Government's Geoscience Data Portal (https://ecat.ga.gov.au/geonetwork/srv/eng/catalog.search#/metadata/145496). As of October 2025, it contains 29,033 records across 341 fields, representing geological, mineralogical, geochemical, and spatial data from various deposit types and environments (Hofstra et al., 2022; Case et al., 2025). The dataset captures multiple dimensions of geological information. These include: (i) Deposit level data: deposit names, environments, groups, types, province, and primary and secondary commodities such as lithium, rare earth elements, niobium, and tantalum. (ii) Sample level data: sample type, feature type, sampling method, material class, names and groups of stratigraphic units, earth material description including group and qualifier, mode of occurrence, alteration, texture, minerals, and sample description and preparation, and sampling method. (iii) Geochemical analyses: measurements of major oxides and trace elements, including the analytical methods used, detection limits applied, and timestamp of analysis. (iv) Location and submitter metadata: Spatial coordinates for samples and deposits, UTM zones, submitter name and descriptive notes.

This structured dataset serves as the foundation for semantic integration, intelligent retrieval, and reasoning using graph structures. Development and analysis were conducted in a Jupyter Python environment using packages such as pandas, numpy, langchain, and openai, with Neo4j desktop (1.6.1) (Neo4j, 2025) serving as the graph database backend. The schema data preparation process involved several key steps: cleaning null and inconsistent values, normalizing field names and measurement units, and standardizing categorical attributes along with geological taxonomies. Data types were cast according to an ontology-driven schema, while geochemical and text-based fields were parsed for embedding and graph ingestion. These measures ensured the data set was thoroughly cleaned, internally consistent, and semantically aligned for downstream processing.

2.2. Vector embedding and semantic indexing

To enable semantic understanding, textual and semi-structured fields were converted into vector embeddings using OpenAI's text-embedding-3-large model (OpenAI, 2024). Each record (row) was summarized into a concise text capturing its mineralogy, deposit type, geochemistry, and geological context. Embeddings were indexed using FAISS (Facebook AI Similarity Search) (Johnson et al., 2019; Meta AI, 2025) to support high-speed semantic retrieval based on cosine similarity. Processing steps included chunking records into coherent text segments, vectorizing contextual fields, and indexing with hybrid metadata filters (e.g., deposit environment, province). This embedding layer enables natural-language queries and context-aware searches, moving beyond rigid keyword matching to semantically informed retrieval. This semantic layer complements, but does not replace, deterministic graph querying, providing contextual coverage when exact structured matches are unavailable.

The embedding workflow begins by normalizing and tokenizing textual fields, ensuring consistent representation of mineral names, lithologies, and geochemical attributes. For each record, a composite text string is generated by concatenating key descriptors (e.g., sample UID, mineral assemblage, deposit type, geochemical signature, and spatial context). This text is then passed to the text-embedding-3-large model, which produces a dense vector representation capturing semantic relationships between terms. These embeddings encode contextual similarity rather than exact keyword matches, allowing queries like “REE-rich deposits in alkaline complexes” to retrieve relevant samples even if terminology differs.

When a user submits a question, the query text is also embedded using the same embedding model, producing a vector in the same semantic space as the stored record embeddings. FAISS uses an index optimized for cosine similarity to compare the query vector against all stored embeddings. Cosine similarity measures the angle between vectors rather than their magnitude; vectors that are nearly parallel (high cosine score) indicate strong semantic alignment. The top-k most similar embeddings are selected as candidates for retrieval, ensuring that results reflect conceptual closeness rather than exact wording. To improve precision, metadata filters (e.g., deposit environment, province, country) are applied during search, enabling hybrid queries that combine semantic similarity with structured constraints. This approach ensures scalable, semantically aware access to heterogeneous geoscience data while maintaining interpretability and relevance.

2.3. Ontology and knowledge graph construction

The CMiO ontology provides a structured semantic framework for representing critical mineral systems by formally defining key domain entities, attributes, and relationships. It models core concepts such as samples, deposits, minerals, commodities, analytical methods, and geological characteristics, along with their interconnections (e.g., sample-deposit associations, elemental compositions, and analytical results). The ontology incorporates controlled vocabularies and standardized property definitions to ensure consistency in representing

geochemical and geological data. By organizing heterogeneous dataset fields into a coherent schema, the CMiO ontology enables integration of multi-source information and supports structured querying and interpretation within the knowledge graph. This ontology-driven design ensures that domain knowledge is explicitly encoded, facilitating reliable data retrieval and semantic interoperability.

Knowledge graph (KG) is a structured, machine-interpretable representation of information in which entities (e.g., deposit type, minerals, deposit environment) are modeled as nodes and their relationships are explicitly encoded as typed edges, enabling semantic integration, reasoning, and flexible querying across heterogeneous data sources (Hogan et al., 2021; Noy et al., 2019; Ontotext, 2025; Zhu et al., 2025). By emphasizing meaning and relationships rather than tabular schemas, KGs support multi-scale data integration and interpretable analysis of complex systems. In the context of critical minerals, KG provides a unifying framework for linking mineralogical, geochemical, geological, and geospatial attributes into coherent representations of mineral systems.

Building on this framework, each relevant attribute (field) in the CMiO dataset was modeled as a node (e.g., SAMPLE_UID, DEPOSIT_TYPE, TEXTURE, MINERALS, Sc_PPM). Categorical properties (e.g., DEPOSIT_GROUP, PRIMARY_COMMODITIES) and numerical properties (e.g., TIO_2_WT_PERCENT, GA_PPM, OS_PPB) were modeled as first-class entities. Relationships were defined using dictionaries from the CSV file, linking elements such as Sample-Mineral, Sample-Deposit, and Sample-Texture to indicate associations. These dictionaries were curated to resolve naming inconsistencies and capture hierarchical distinctions (such as deposit type versus deposit name), ensuring a controlled vocabulary for graph constructions.

The relationship schema (Fig. 1) was serialized for reuse and efficiency, defining explicit edge types such as HAS_MINERAL, HAS_TEXTURE, and BELONGS_TO_DEPOSIT, along with their inverses for bidirectional queries. The CMKG converts CMiO's tabular data into a semantic network in Neo4j, where nodes and edges follow the ontology. Nodes include SampleRecord, Minerals, Texture, Alteration, Deposit Type, Deposit Name, Province, major oxides (e.g., TIO2_WT_PCT), trace elements (e.g., REE_Y_PPM, RH_PPB), and Stratigraphic Unit. Relationships link samples to minerals, textures, alterations, lithologies, analytical methods, and locations, forming a multi-relational structure for complex queries.

To ensure consistency, GPT-4o was used to normalize chemical symbols and mineral names, resolving issues like variant spellings and case sensitivity. These normalization steps were applied during preprocessing and do not affect deterministic query execution, ensuring that graph-based retrieval remains reproducible and independent of generative variability. Data ingestion was performed in batches using Neo4j's MERGE command to maintain schema integrity and prevent duplication of nodes and relationships. Each batch was validated against the ontology to ensure compliance with the predefined schema. The resulting graph captures multi-scale geological context, from sample-

```
sample_relationships = {
  "HAS_MINERAL": ("Minerals", "MINERALS"),
  "HAS_ALTERATION": ("Alteration", "ALTERATION"),
  "HAS_TEXTURE": ("Texture", "TEXTURE"),
  "FROM_COUNTRY": ("Country", "COUNTRY"),
  "HAS_LITHOLOGY": ("MaterialClass", "MATERIAL_CLASS"),
  "HAS_STRAT_NAME": ("StratUnitName", "STRAT_UNIT_NAME"),
  "HAS_STRAT_GROUPING": ("StratGrouping", "STRAT_GROUPING"),
  "HAS_MATERIAL_GROUP": ("EarthMaterialGroup", "EARTH_MATERIAL_GROUP"),
  "HAS_MATERIAL_QUALIFIER": ("EarthMaterialQualifier", "EARTH_MATERIAL_QUALIFIER"),
  "HAS_EARTH_MATERIAL": ("EarthMaterial", "EARTH_MATERIAL"),
  "HAS_MODE_OF_OCCURRENCE": ("ModeOccurrence", "MODE_OCCURRENCE"),
  "HAS_DESCRIPTION": ("SampleDescription", "SAMPLE_DESCRIPTION"),
  "PREPARED_BY": ("SamplePreparation", "SAMPLE_PREPARATION"),
  "HAS_SAMPLE_TYPE": ("SampleType", "SAMPLE_TYPE"),
  "HAS_MAJOR_OXIDE": ("MajorOxides", "MAJOR_OXIDES"),
  "HAS_TRACE_ELEMENTS": ("TraceElements", "TRACE_ELEMENTS"),
  "HAS_SUBMITTER": ("Submitter", "SUBMITTER"),
  "ANALYZED_ON": ("analysisDate", "ANALYSIS_DATETIME"),
  "HAS_LONGITUDE": ("SampleLongitude", "SAMPLE_LONGITUDE_WGS84"),
  "HAS_LATITUDE": ("SampleLatitude", "SAMPLE_LATITUDE_WGS84"),
  "TAKEN_FROM": ("Deposit", "DEPOSIT_UID"),
}
deposit_relationships = {
  "HAS_NAME": ("DepositName", "DEPOSIT_NAME"),
  "HAS_ENV": ("DepositEnvironment", "DEPOSIT_ENVIRONMENT"),
  "HAS_GROUP": ("DepositGroup", "DEPOSIT_GROUP"),
  "HAS_TYPE": ("DepositType", "DEPOSIT_TYPE"),
  "HAS_PRIMARY_COMMODITY": ("PrimaryCommodities", "PRIMARY_COMMODITIES"),
  "HAS_SECONDARY_COMMODITY": ("SecondaryCommodities", "SECONDARY_COMMODITIES"),
  "HAS_ALL_COMMODITIES": ("AllCommodities", "ALL_COMMODITIES"),
  "HAS_PROVINCE": ("Province", "PROVINCE"),
  "HAS_SAMPLE": ("SampleRecord", "SAMPLE_UID"),
}
```

Fig. 1. Relationship dictionaries for samples and deposits. Two dictionaries define semantic relationships for graph-based geological modeling. Each key specifies a relationship type (e.g., HAS_MINERAL, HAS_ALTERATION), and each value pairs a human-readable label with the standardized dataset field. The sample dictionary links samples to attributes such as minerals, lithology, geochemistry, spatial coordinates, and provenance (TAKEN_FROM). The deposit dictionary connects deposits to properties like name, environment, commodities, and associated samples (HAS_SAMPLE). These mappings ensure semantic consistency and enable complex queries across the graph.

level geochemistry to deposit-level classification, supporting both deterministic reasoning through Cypher queries and semantic linkage for advanced retrieval. This design enables flexible exploration of mineral systems, facilitating pattern discovery across compositional, and spatial dimensions.

2.4. Interactive user interface for QA

An interactive question builder, developed with ipywidgets, allows users to query the knowledge graph using dropdowns, filters, or free-text input. Key features include natural language query construction, terminology normalization (for example, mapping "rare earth elements" to "REE" and "platinum group elements" to "PGE"), and dynamic routing of queries through the GraphRAG pipeline to generate contextual responses. The interface exposes deterministic, semantic, and synthesized outputs separately, enabling users to assess provenance and interpretability of results. This interface democratizes access to complex geological information, empowering geoscientists and policymakers to explore intricate relationships without requiring specialized database expertise. The QA framework enables users to ask complex geological questions in natural language, explore multi-scale relationships among samples, deposits, and environments, visualize contextual geological patterns, and generate new insights into critical mineral processes. Ultimately, CMKG transforms the static geological dataset into a dynamic, knowledge-driven platform for exploration, analysis, and decision-making.

2.5. Hybrid GraphRAG based answering (QA)

As the volume and complexity of geologic data grow, traditional keyword searches are insufficient for deep reasoning and discovery because they cannot capture multi-relational context or semantic meaning. GraphRAG addresses this challenge by combining semantic search with structured graph reasoning, grounding LLM responses in verifiable sources such as CMiO. Unlike simple text retrieval, this approach integrates knowledge graphs with large language models to retrieve structured relationships and domain-specific context, then augments LLM outputs with this information.

The CMKG system implements GraphRAG by fusing the deterministic reasoning capabilities of graph databases with the generative strengths of LLMs. In this hybrid design, deterministic Cypher queries are executed first to obtain authoritative results, followed by semantic retrieval when exact matches are unavailable, and finally LLM-based synthesis that integrates retrieved evidence without introducing new unverified information. Semantic retrieval (via FAISS) and graph querying (via Neo4j) operate in tandem to deliver evidence-based answers. For example, queries such as "List deposits with lithium (Li) and gallium (Ga) as commodities" or "What are the LREE, HREE, REE, and REE_Y contents of sample AU.1,459,665?" are parsed, translated into Cypher queries, and executed on the graph to extract precise relationships. Retrieved subgraphs are then summarized by the LLM into contextual, human-readable answers, ensuring that responses are both accurate and interpretable.

The workflow includes several stages: Query interpretation: User input is embedded (vectorized) and matched to relevant nodes and relationships using semantic similarity. Hybrid retrieval: FAISS returns semantically similar matches, while Neo4j executes Cypher queries for deterministic graph-based results. Answer synthesis: The LLM fuses structured graph data with semantic text to produce coherent, context-rich explanations. Fallback mechanism: When graph data is incomplete or unavailable, the system defaults to semantic retrieval to maintain continuity. This hybrid approach effectively bridges symbolic reasoning (knowledge graph queries) and neural reasoning (LLM generation), combining precision with flexibility. By grounding generative outputs in structured, verifiable data, GraphRAG delivers responses that are accurate, explainable, and reproducible; capabilities essential for scientific domains where transparency and trust are paramount.

3. Results

3.1. Knowledge graph ingestion pipeline

To build the knowledge graph, the system processes all 29,033 rows from the CMiO.csv dataset. This is done in manageable chunks, applying a strict, ontology-derived schema (e.g., float, integer, string) to ensure each cell is correctly typed and validated. Data is ingested in batches for scalability, using environment variables like NEO4J_PASSWORD to securely connect to the Neo4j database via the Bolt protocol. The graph uses Cypher's MERGE operation on SAMPLE_UID and DEPOSIT_UID to avoid duplicate entries. Nodes and relationships are created based on a set of predefined relationship dictionaries (see Fig. 1), which guide the structure of the graph.

To improve computational efficiency, the ingestion process is divided into ten batches of 1000 rows each. Intermediate outputs are stored as Pickle files (Van Rossum, 2020), enabling modular and scalable processing. This design supports incremental updates to the CMiO dataset: new rows can be added without reprocessing the entire graph by detecting entries not yet ingested. After batching, the system reloads the pickled relationship schema and reconstructs the graph using either the Neo4j driver, a NetworkX representation, or an embedding model. Newly ingested data are seamlessly integrated into the existing graph, preserving consistency and structural continuity as the CMiO dataset evolves. This deterministic ingestion pipeline ensures reproducibility and consistency of the knowledge graph, which forms the authoritative foundation for downstream query evaluation.

Fig. 2 shows an example subgraph illustrating how nodes in the knowledge graph are related. This visualization represents only a small portion of the overall knowledge graph, depicting the connection between a mineral deposit and a sample taken from it, along with their geological and geochemical attributes. At the center-left, the red node labeled Deposit with DEPOSIT_UID "AUS.NSW.335,841" represents the Woodlawn deposit located in New South Wales, Australia. This deposit is linked to several descriptive attributes: DEPOSIT_GROUP: Volcanogenic massive sulfide (VMS); DEPOSIT_TYPE: Bimodal felsic VMS; DEPOSIT_ENVIRONMENT: Volcanic basin hydrothermal; ALL_COMMODITIES: Zn, Cu, Pb, (Ag, Au). These attributes list the primary (Zn, Cu, Pb) and secondary (Ag, Au) economic metals associated with the deposit. The blue node labeled Sample with SAMPLE_UID "AU.1,459,117" represents a sample collected from this deposit (through the TAKEN_FROM relationship). This sample is characterized by trace element data: REE: 136.0 ppm, HREE: 0.0 ppm, LREE: 136.0 ppm, and REE_Y: 136.0 ppm. These values indicate concentrations of rare earth elements (REE), including light REE (LREE), heavy REE (HREE), and yttrium (REE_Y). The edges in the graph define semantic relationships such as HAS_GROUP, HAS_TYPE, HAS_ENV, HAS_ALL_COMMODITIES, and HAS_TRACE_ELEMENTS, showing how these attributes connect to the deposit and sample. Overall, Fig. 2 serves as an illustrative example of how geological and geochemical data are integrated into a structured graph format, enabling complex queries and analysis of mineral systems.

3.2. Quantitative evaluation of QA performance

To evaluate the effectiveness of the proposed hybrid GraphRAG framework, a benchmark dataset of 20 structured and semi-structured queries was constructed from the CMiO dataset. This benchmark was selected as a subset of the structured queries shown in Fig. 4 and was designed to represent deterministic and semi-structured question types with verifiable ground truth answers, which were manually validated from source data.

The system was evaluated under four configurations: (1) Deterministic-only (Cypher queries), (2) Hybrid (graph + semantic

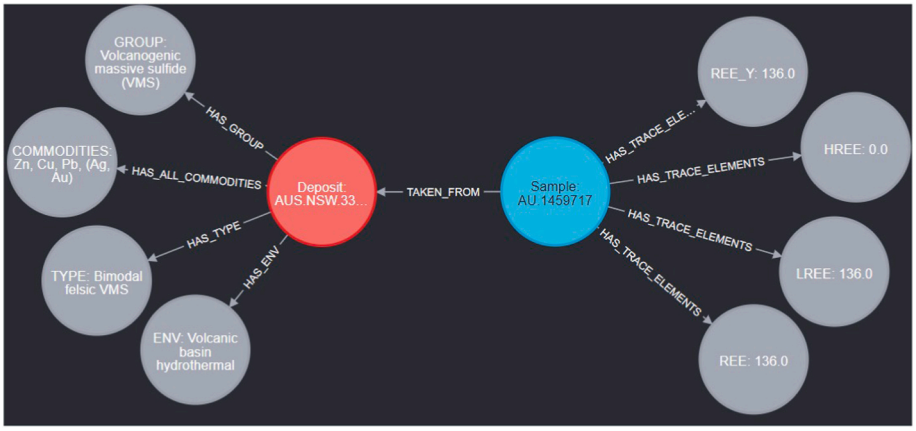

Fig. 2. An Example subgraph showing the relationships between a specific sample and its associated deposit attributes and trace element concentrations. The central blue node represents the Woodlawn SampleRecord (SAMPLE_UID: AU.1,459,717), which is connected to four trace element properties via HAS_TRACE_ELEMENTS relationships: REE (136.0 ppm), HREE (0.0 ppm), LREE (136.0 ppm), and REE_Y (136.0 ppm). The large red node represents the Woodlawn Deposit (DEPOSIT_UID: AUS. NSW.335,841), linked to the sample through the TAKEN_FROM relationship. The deposit node is further connected to four key attributes: Deposit Environment (Volcanic basin hydrothermal), Deposit Group (Volcanogenic massive sulfide (VMS)), Deposit Type (Bimodal felsic VMS), and All Commodities (Zn, Cu, Pb, (Ag, Au)) via relationships HAS_ENV, HAS_GROUP, HAS_TYPE, and HAS_ALL_COMMODITIES. This visualization highlights both geochemical data and geological context, providing an integrated view of sample-deposit associations. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

retrieval + LLM synthesis), (3) Semantic-only (FAISS retrieval), and (4) LLM-only (direct generation). Performance was measured using Exact Match (EM), Contains Match (CM), and latency. Results show that deterministic and hybrid approaches achieve comparable performance (EM = 88.9%), while semantic-only and LLM-only approaches fail to produce correct answers for structured queries (EM = 0.00). These findings demonstrate that (i) deterministic graph querying is essential for accurate structured scientific QA, (ii) the hybrid architecture preserves deterministic accuracy while improving interpretability, and (iii) semantic retrieval and LLM-only approaches are insufficient for precise attribute-level queries.

3.3. Comparative analysis of retrieval modes

A comparative analysis of deterministic graph querying, semantic retrieval, and LLM-based synthesis highlights the distinct strengths and limitations of each component. Deterministic graph querying consistently produced the most accurate and reproducible responses for structured factual questions involving sample identifiers, geochemical attributes, deposit metadata, and elemental concentrations. The strong entity-level evaluation results (EM = 88.9%, CM = 94.4%) demonstrate that deterministic Cypher execution is the primary driver of factual correctness within the framework. Semantic retrieval contributes contextual flexibility by retrieving geologically related documents and semantically similar records; however, semantic retrieval alone does not reliably produce exact attribute-level answers for structured scientific queries. LLM-only responses often generated fluent geological explanations but lacked sufficient grounding for precise factual retrieval.

The hybrid architecture combines these complementary capabilities by prioritizing deterministic graph retrieval while incorporating semantic context and optional natural-language synthesis. This design preserves the reproducibility and precision of deterministic querying while improving interpretability and user accessibility for interactive exploration.

3.4. Error analysis

A detailed error analysis was conducted across all benchmark categories. The majority of successful predictions occurred in deterministic entity-centric retrieval tasks, where Cypher-based graph querying consistently returned accurate geological attributes and sample metadata. Remaining failures were primarily associated with aggregation, comparative, and multi-hop reasoning tasks. These errors largely resulted from incomplete deterministic rule coverage, schema-linking limitations, and unresolved reasoning patterns involving multiple graph relationships. Additional failures were associated with heterogeneous data quality issues, including missing or incomplete attribute values within the source datasets. Importantly, the evaluation framework demonstrated that deterministic graph retrieval substantially reduces hallucination risk compared to unconstrained LLM generation. Because all entity-level answers are grounded directly in structured graph data, the framework provides transparent and reproducible retrieval behavior for factual geological queries. These findings indicate that deterministic graph querying is highly effective for precise geological information retrieval, while advanced reasoning over complex geological relationships remains an important area for future enhancement.

3.5. Domain relevance and exploration utility

Beyond quantitative evaluation, the proposed framework demonstrates practical relevance for mineral exploration workflows. The system enables structured access to geochemical and geological data, supporting tasks such as identification of rare earth elements (REE)-enriched samples and deposits, analysis of multi-element associations (e.g., Li-Ga or REE groupings), and integration of geochemical, geological, and spatial attributes. These capabilities extend traditional database querying by enabling multi-relational exploration within a unified knowledge graph. Rather than relying on manual data inspection, users can perform structured queries that combine multiple attributes across samples and deposits. While the system does not replace domain expertise, it provides a scalable and interpretable tool for data-driven analysis, supporting exploration workflows and facilitating hypothesis generation in critical mineral research.

Fig. 3A shows an example Cypher query used to retrieve information for a sample (AU.7371885) linked to a deposit (Hemlo Williams) through the TAKEN_FROM relationship. Fig. 3C and D displays the property panels returned by the query for the deposit and sample nodes. The deposit node includes attributes such as deposit name (Hemlo Williams), deposit type (Mesozonal orogenic gold), deposit group (Orogenic), and deposit environment (Metamorphic hydrothermal). The sample node shows properties including sample UID (AU.7371885),

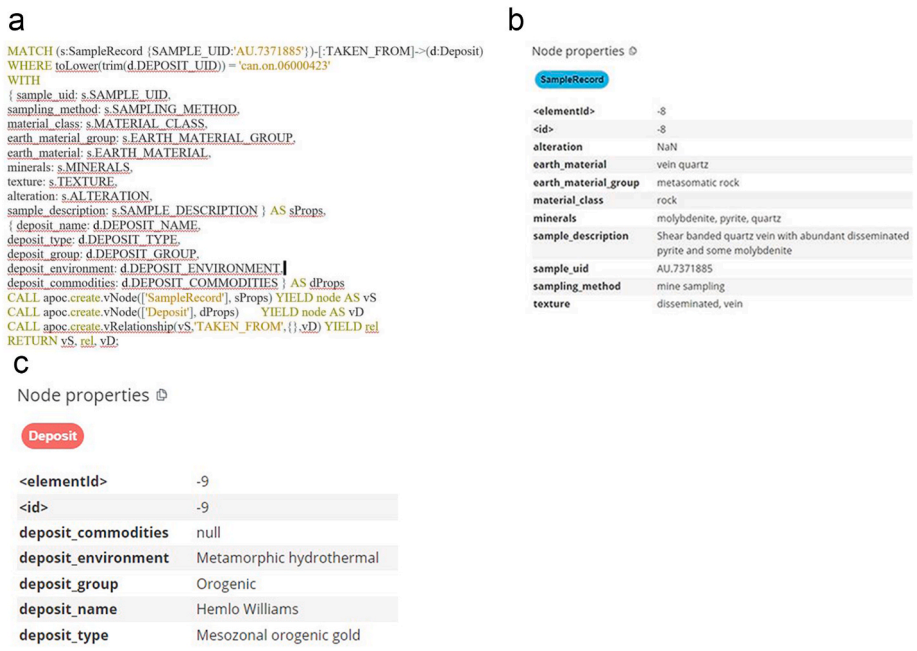

Fig. 3. Example Neo4j Cypher query and returned results. Fig. 3A (top) shows the Cypher query used to retrieve information for the sample node (UID: AU.7371885) linked to the deposit node (UID: CAN. ON.0600,042) through the TAKEN_FROM relationship. The query extracts selected attributes for both nodes. Fig. 3B (lower left) displays the returned property panel for the sample node, listing attributes such as sample UID (AU.7371885), sampling method (mine sampling), minerals (molybdenite, pyrite, quartz), texture (disseminated, vein), and a descriptive note about the shear banded quartz vein with disseminated pyrite and molybdenite. Fig. 3C (lower right) shows the returned property panel for the deposit node, including deposit name (Hemlo Williams), deposit type (Mesozonal orogenic gold), deposit group (Orogenic), and deposit environment (Metamorphic hydrothermal). Together, these panels demonstrate how Cypher queries retrieve and present both relationships and key geological properties within the knowledge graph. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

sampling method (mine sampling), minerals (molybdenite, pyrite, quartz), texture (disseminated, vein), and a descriptive note about the shear banded quartz vein with disseminated pyrite and molybdenite. Together, these panels demonstrate how Cypher queries retrieve and visualize relationships and key geological properties within the knowledge graph.

3.6. Example queries for QA

A subset of 20 structured queries from the full set shown in Fig. 4 was selected as a benchmark for quantitative evaluation (Section 3.2), while the remaining queries serve as demonstration examples of system capability. These queries were designed to span multiple dimensions of the dataset, including geological, geochemical, sample-level, deposit-level, and spatial aspects, ensuring comprehensive coverage of the knowledge graph's capabilities. The goal was to test both accurate data retrieval and reasoning across all components of the hybrid GraphRAG pipeline. The queries include sample-specific questions, such as retrieving elemental concentrations, analytical methods, and mineralogical composition for individual samples. Deposit-level queries explore relationships between deposits and their attributes, including commodities, deposit types, and associated provinces. In addition, advanced analytical queries were formulated to identify patterns, such as locating samples or deposits enriched in rare earth elements (REE) or associated with critical minerals, which are essential for resource assessment and exploration. These queries also form the basis of the quantitative benchmark described in Section 3.2, ensuring consistency between qualitative demonstrations and quantitative evaluation.

Each query is executed through the integrated pipeline, which combines three key components: Neo4j for structured graph queries, FAISS for semantic similarity search across unstructured text, and GPT-4o for contextual synthesis of retrieved information. This architecture enables the system not only to retrieve precise data from the graph but also to interpret and summarize complex relationships, providing users with both structured outputs and natural language explanations. This design allows direct comparison between deterministic, semantic, and hybrid responses, supporting systematic evaluation of each component's contribution.

Fig. 5 presents questions selected from the list in Fig. 4, along with their corresponding answers retrieved from the knowledge graph, illustrating how the QA system processes queries and returns structured information. These examples were selected because they have concise answers, making them suitable for display within limited space. While this figure shows only a small subset of possible queries, it demonstrates the system's ability to handle diverse question types related to geological and geochemical attributes.

3.7. Unified question answering (QA) architecture

The QA module integrates deterministic parsing, LLM synthesis, and semantic retrieval to enable robust, context-aware question answering. At its core, the system uses rule-based logic to extract user intent and generate structured responses. A suite of strongly typed builder functions handles tasks such as: Field and commodity extraction (e.g., resolving “Zn” to “zinc”); Numeric field detection (e.g., mapping “Fe2O3 total” to FE2O3TOT_WT_PERCENT); Method classification (distinguishing sampling vs. analytical methods); Compound queries (e.g., “samples with Zn and Cd”); Metadata and ranking queries (e.g., “top 5 deposits for Mn”); Geographic resolution (e.g., interpreting “NSW” as New South Wales and “Victoria” as separate states in Australia, ensuring they are treated as distinct regions). To ensure reproducibility, all LLM interactions were executed with fixed parameters (temperature = 0), and prompts were logged during evaluation, enabling consistent and repeatable outputs across runs.

The system maintains comprehensive dictionaries that map human-readable terms to standardized identifiers, ensuring consistent

```
example_questions = [
  "1. What is the deposit type of sample AU.1459712?",
  "2. Which commodities are present in sample AU.1459712?",
  "3. What is the SIO2 content of sample AU.1459712?",
  "4. Which sample has the highest SIO2 content?",
  "5. Provide all samples collected from New South Wales (NSW).",
  "6. What is the Fe2O3 content of sample AU.1459634?",
  "7. What analytical method was used for Fe2O3 from sample AU.1459634?",
  "8. What analytical method was used for Au related to sample AU.1467714?",
  "9. What sampling method was used for sample AU.1458591?",
  "10. What is the sample type of sample AU.1459712?",
  "11. What is the material class of sample AU.1459712?",
  "12. What is the stratigraphic unit name for sample AU.1459712?",
  "13. What are the earth material qualifier, earth material, and sample description of sample AU.1459717?",
  "14. What is the mode of occurrence of sample AU.1463247?",
  "15. What is the alteration of sample ca.824266?",
  "16. What is the texture of sample AU.7372076?",
  "17. What are the minerals for sample AU.7371776?",
  "18. List all samples associated with the Woodlawn deposit.",
  "19. Which samples have the highest Ag (silver) content?",
  "20. What are the commodities of the Panton deposit?",
  "21. What is the province for Woodlawn deposit?",
  "22. Which samples have the highest REE, LREE, HREE, or REE-Y content?",
  "23. Which deposits have PGE commodity?",
  "24. Which samples have PGE?",
  "25. List deposits with zinc (Zn) and cadmium (Cd) commodities",
  "26. List deposits with both tin (Sn) and tungsten (W) commodities",
  "27. List deposits with lithium (Li) and or gallium (Ga) commodities; also show sample values.",
  "28. Which samples contain lithium (Li) and gallium (Ga)?",
  "29. What is the sample description and deposit type of deposits with REE commodity",
  "30. What is the sample description and deposit type of deposits with Y commodity",
  "31. Which Deposits have the REE, LREE, HREE, or REE-Y commodities",
  "32. Which deposits have Dmd, Dmdi, Dmdg commodities?",
  "33. What are the LREE, HREE, REE, and REE_Y content of sample AU.1459665?",
  "34. What are the minerals of samples with Os, Rh, and Ru commodities?",
  "35. Which deposits contain highest content of Mn, Dy, Nd, Pr, and Ge?",
]
```

Fig. 4. Example set of structured queries curated to evaluate the knowledge graph system and guide user interaction. These queries cover multiple dimensions of the dataset, including geological attributes (e.g., deposit type, stratigraphic unit, material class), geochemical properties (e.g., elemental concentrations such as SiO_2, Fe_2O_3, REE, PGE), sample-level details (e.g., minerals, texture, sampling method), deposit-level relationships (e.g., commodities, deposit group, province), and advanced analytical queries (e.g., identifying deposits with critical minerals or rare earth element associations). By spanning these diverse aspects, the examples serve both as benchmarks for testing retrieval and reasoning capabilities and as templates to help users formulate their own questions.

interpretation of queries (for example, mapping "sample id" to SAMPLE_UID and "Cd" to Cadmium) and alignment with DataFrame column names. All keys are normalized by converting to lowercase, removing punctuation, and deduplicating entries to guarantee robust matching across variations in phrasing. A dispatcher function composes these builders by prioritizing specificity and relevance while providing fallback behavior when queries fall outside the supported scope. It supports multi-commodity, multi-attribute, and fallback scenarios to ensure reliable responses. The system also supports multi-column commodity groups (e.g., PGE, representing platinum group elements) and alias resolution, which allows queries to be interpreted accurately regardless of terminology.

An OpenAI LLM (GPT-4o) contextualizes deterministic outputs in a conversational format, instructed to rely primarily on structured logic and use semantic matches only for supplementary context. If requested data is unavailable, the LLM transparently communicates this, ensuring clarity and trust. The semantic layer retrieves contextually relevant passages using FAISS (dense vector retrieval), optionally backed by TF-IDF with cosine similarity. Integrated via LangChain and OpenAI embeddings, FAISS indexes textual fields from the dataset. It ranks top-k snippets by similarity, falling back to disk-based FAISS with embedding caching if needed. Semantic retrieval is complemented by graph-based reasoning via Neo4j. When Cypher queries return no results, fallback logic ensures continuity. LangChain's RetrievalQA and GraphCypherQAChain enable hybrid reasoning across unstructured and structured data. Final responses combine snippets from FAISS retrieval, results from Cypher graph queries, and a unified, markdown-formatted LLM-generated answer. Importantly, deterministic graph queries are executed prior to any LLM involvement, ensuring that authoritative results are derived directly from structured data before any generative synthesis is applied.

3.8. Interactive user interface (UI)

To support intuitive exploration and troubleshooting, the system provides a Jupyter-based interface built with ipywidgets. Designed for both technical and non-technical users, the UI abstracts underlying complexity and enables natural language interaction with the CMiO dataset. Key components include Dropdown menus for guided and example-based queries and free-form input for custom questions (Fig. 6).

Upon submission, each query, whether structured or free-text, passes through a three-stage pipeline (Fig. 7): Deterministic Response (Blue): Executes exact lookups using Neo4j and pandas to return authoritative data (e.g., deposit type, mineral composition, analytical method). LLM Response (Orange): Synthesizes deterministic and semantic outputs using GPT-4o, generating concise, contextual summaries grounded in geological relevance. Semantic Search Response (Green): Uses FAISS-based vector retrieval to identify contextually relevant passages when exact matches are unavailable. Results are displayed in side-by-side, color-coded answer boxes (Fig. 7), allowing users to compare deterministic accuracy, semantic relevance, and LLM interpretation. This layered presentation enhances transparency and interpretability. The

Answer to selected samples (only 1-3 short answers provided for brevity):

1. What is deposit type of sample AU.1459712?: Bimodal felsic VMS
3. What is the SIO2 content of sample AU.1459712? SIO2_WT_PERCENT: 69.93
4. Which sample has the highest SIO2 content? Value: 98.0 SAMPLE_UID(s): AU.3310602
8. What analytical method was used for Au related to sample AU.1467714?
Numeric Result: AU_PPB: 14.0 (method: ICPMS_FAU)
9. What sampling method was used for sample AU.1458591? SAMPLING_METHOD: drilling
10. What is the sample type of sample AU.1459712? SAMPLE_TYPE: borehole specimen
11. What is the material class of sample AU.1459712? MATERIAL_CLASS: rock
12. What is the stratigraphic unit name for sample AU.1459712?
STRAT_UNIT_NAME: ore horizon STRAT_GROUPING: ore horizon
13. What are the earth material qualifier, earth material, and sample description of sample AU.14597?
EARTH_MATERIAL_QUALIFIER: basic fine grained EARTH_MATERIAL: basic volcanic rock
SAMPLE_DESCRIPTION: hghly altrd fne-graind basic rck,anlyses poss affctd by weathrng
14. What is the mode of occurrence of sample AU.1463247? MODE_OCCURRENCE: dyke
15. What is the alteration of sample ca.824266? ALTERATION: quartz-sericite alteration
16. What is the texture of sample AU.7372076? TEXTURE: disseminated
17. What are the minerals for sample AU.7371776?
MINERALS: chalcopyrite, galena, pyrite, sphalerite
18. List all samples associated with the Woodlawn deposit. AU.1457754; AU.1457762; AU.1457778
19. Which samples have the highest Ag (silver) content? Value: 223000.0 SAMPLE_UID(s): ca.829638
20. What are the commodities of the Panton deposit? PRIMARY_COMMODITIES: Pt, Ni, Pd;
SECONDARY_COMMODITIES: Cu, Au, Cr, PGE
21. What is the province for Woodlawn deposit? PROVINCE: Lachlan Orogen
22. Which samples have the highest REE, LREE, HREE, or REE-Y content?
Sample(s) with highest REE_PPM: Value: 113645.22 SAMPLE_UID(s): AU.7372485
Sample(s) with highest LREE_PPM: Value: 112673.0 SAMPLE_UID(s): AU.7372485
23. Which deposits have PGE commodity? Deposits with PGE as commodities
(highest values per analyte):Bingham Canyon: PT_PPB: 3.0 | PD_PPB: 3.0
24. Which samples have PGE? PGE (members: IR, OS, PD, PT, RH, RU) (showing 1 of 7835):
AU.1467714 – PD_PPB: 170.9 | PT_PPB: 181.2
25. List deposits with zinc (Zn) and cadmium (Cd) commodities. Deposits (through samples)
with Highest CD and ZN as commodities: Badger: CD_PPM: 620.0 | ZN_PPM: 268000.0
Balmat-Edwards District: CD_PPM: 1410.0 | ZN_PPM: 447000.0
28. Which samples contain lithium (Li) and gallium (Ga)? showing 1 of 23770:
AU.1459712 – GA_PPM: 10.0 | LI_PPM: 5.0
29. What is the sample description and deposit type of deposits with REE commodity?
SAMPLE_UID:AU.7372484|DEPOSIT_TYPE: Carbonatite laterite REE|SAMPLE_DESCRIPTION:
OSNACA Low grade or waste sample.Pale brown colloform sphalerite, abundant coarse galena and minor pyrite
30. What is the sample description and deposit type of deposits with Y commodity?
SAMPLE_UID: AU.7371917 | DEPOSIT_TYPE: NYF pegmatite | SAMPLE_DESCRIPTION:
OSNACA Pegmatite or iron ore sample. Banded felsic rock with fluorite
33. What are the LREE, HREE, REE, and REE_Y content of sample AU.1459665?
Sample AU.1459665 contents – LREE_PPM: 159.0; HREE_PPM: 0.0; REE_PPM: 159.0;
Y_PPM: 53.0; REE_Y_PPM: 212.0
35. Which deposits contain highest content of Mn, Dy, Nd, Pr, and Ge?
DY_PPM (highest): Bear Lodge: 4360.0
GE_PPM (highest): Central Tennessee District: 506.0
MN_PPM (highest): Greens Creek: 326000.0
ND_PPM (highest): Pea Ridge: 42400.0
PR_PPM (highest): Pea Ridge: 14068.86

Fig. 5. Selected example questions from Fig. 4 and their corresponding answers, illustrating the system's ability to retrieve precise values such as elemental concentrations, sample descriptions, and relationships between samples and deposits. The questions span multiple dimensions, including sample-level details (e.g., SiO_2 content, sampling method, minerals, texture), deposit-level attributes (e.g., commodities, deposit type, province), and advanced analytical queries (e.g., identifying samples with highest REE or PGE content). For brevity, only one to three representative answers are shown for each question.

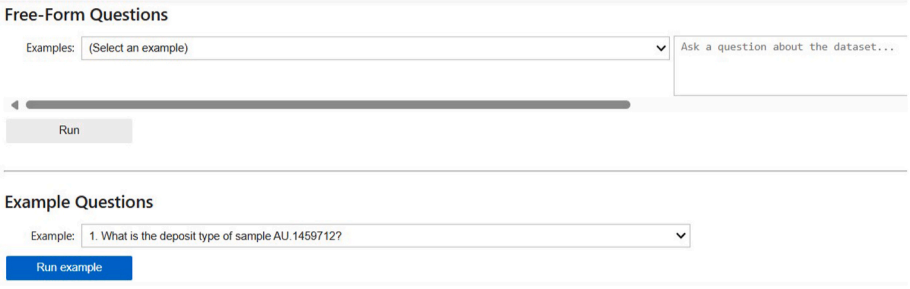

Fig. 6. Interface for querying the dataset: Users can either type custom questions in the Free-Form section or select from predefined Example Questions (Fig. 4) to quickly run queries.

side-by-side presentation provides an explicit mechanism for users to evaluate consistency across retrieval modes, supporting transparency and facilitating qualitative validation of system outputs. By integrating structured logic, semantic retrieval, and generative synthesis, the interface enables users to pose complex queries such as "Which sample has the highest REE?" and receive responses that are both accurate and geologically meaningful. For example, it can return the sample's unique identifier (SAMPLE_UID), the value of its REE_PPM property, and details about its associated deposit, including name and other attributes.

This hybrid approach delivers scalable, user-friendly access to

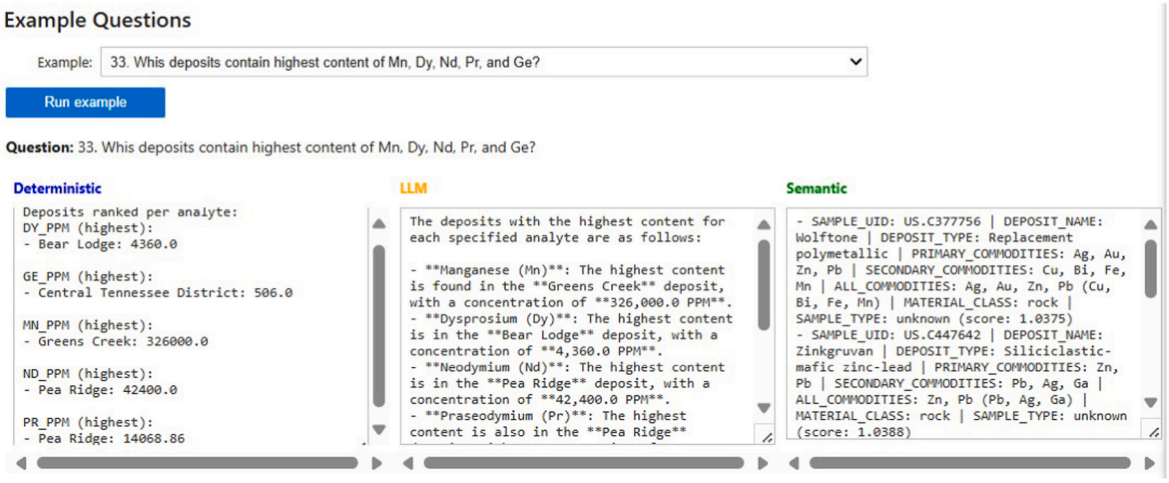

Fig. 7. Results of a single query answered using three complementary methods. The deterministic approach (left) returns exact, ranked results based on explicit analyte values. The LLM-based approach (center) synthesizes these results into a natural-language explanation highlighting the top deposits for each element (Mn, Dy, Nd, Pr, and Ge). The semantic similarity approach (right) retrieves contextually relevant deposits based on embedding similarity, providing additional descriptive metadata and relevance scores. Together, the figure illustrates how deterministic accuracy, interpretive reasoning, and semantic context contribute distinct yet complementary insights for query.

intricate mineral datasets, empowering data-driven research and informed decision-making in the management of critical minerals. The interface also enables manual inspection of system behavior, allowing users to identify discrepancies between deterministic, semantic, and LLM-generated outputs, which complements the quantitative evaluation presented in Section 3.2.

3.9. Free-form natural language queries

The system also supports open-ended, user-driven questions, enabling flexible interaction with the Critical Minerals Knowledge Graph (CMKG) beyond structured example queries. A set of free-form natural language queries was developed to mimic real-world geological exploration scenarios (Fig. 8). These queries allow users to pose

Free-form Examples = {
"Sample metadata": [
"1. US.C389729: What is the sample description and deposit type?",
"2. What is the sampling method and sample type of AU.1459702?",
"3. Provide the stratigraphic unit (STRAT_UNIT_NAME) and grouping (STRAT_GROUPING) for sample US.C447050.",
"4. For sample AU.1636046, list the minerals and mode of occurrence.",
"5. What are the earth material and earth material qualifier for AU.1459660?",],
"Analytical & numeric": [
"6. What analytical method was used for US.C389729 to determine Be?",
"7. Show Fe2O3 total for AU.1636091 (value, detection limit, and method).",
"8. For ca.824143, list SIO2 wt%, detection limit, and method.",
"9. What are Cd (ppm), detection limit, and method for US.SACM25720?",
"10. Provide LREE_PPM, HREE_PPM, REE_PPM, and REE_Y_PPM for sample AU.1457811.",],
"Commodities": [
"11. List deposits with both zinc (Zn) and cadmium (Cd) commodities.",
"12. Which deposits contain Au and Ag commodities?",
"13. What is the sample description and deposit type of samples with REE commodity?",
"14. Show samples with yttrium (Y) commodity along with sample descriptions and deposit types.",],
"Deposit-level": [
"15. List samples associated with the Whitetail Ridge deposit.",
"16. What is the deposit type and environment of the Nashville deposit?",
"17. Provide the province and deposit group for Bull Hill deposit.",
"18. Which deposit has the highest average Cd (ppm)?",
"19. What is the deposit source for Barr deposit?",],
"Location & coordinates": [
"20. List samples collected from New South Wales (NSW).",
"21. Show samples collected from Victoria (VIC).",
"22. What are the sample latitude and longitude for US.C389729?",
"23. Provide the deposit latitude and longitude for the Yeager deposit.",
"24. What is the sample UTM (easting, northing, zone) for US.C447050?",],
"Drill, dates, and provenance": [
"25. For AU.7368463, what are the top and base depths (m)?",
"26. What is the analysis date (ANALYSIS_DATETIME) for US.C389729?",
"27. Who is the submitter and what is the sample source for AU.1458571?",
"28. What is the last update date for US.C389729?",
"29. Show any comments for sample AU.7372565.",],
"IDs & names": [
"30. What is the DEPOSIT_UID and DEPOSIT_LOCAL_ID of the Buick deposit?",
"31. For sample US.C380137, provide the SAMPLE_NAME and SAMPLE_LOCAL_ID.",
"32. What is the FEATURE_UID of sample US.C447050?",]}

Fig. 8. Thematic grouping of free-form example queries. This figure shows representative free-form questions organized into six categories: Sample Metadata, Analytical and Numeric, Commodities, Deposit-Level, Location and Coordinates, and Drill/Dates/Provenance. Examples include queries such *as* "What is the sampling method and sample type of AU.1,459,702?", "Provide LREE_PPM, HREE_PPM, REE_PPM, and REE_Y_PPM for sample AU.1457811", "List deposits with both Zn and Cd as commodities", and "Show samples collected in Victoria (VIC)". These examples illustrate the system's ability to handle diverse geological questions ranging from precise analytical values to broader contextual insights.

questions without predefined templates, testing the system's adaptability across structured and unstructured data. The free-form queries complement the structured benchmark by evaluating system robustness under realistic, unconstrained user inputs.

Free-form questions were organized into six functional categories that reflect the diverse information needs of geologists and mineral exploration professionals: Sample Metadata: Sample identifiers, collection methods, and contextual details; Analytical and Numeric: Geochemical measurements, elemental concentrations, and statistical summaries; Commodities: Specific critical minerals or element groups; Deposit-Level: Deposit types, classifications, and related geological features; Location and Coordinates: Spatial queries involving geographic context; Drill, Dates, and Provenance: Temporal and source-related information, including drilling history and sample chronology. This categorization ensures that the evaluation reflects real-world exploration workflows, including both precise analytical queries and broader contextual categorization.

Although the system supports open-ended natural language interaction, the strongest quantitative performance was observed for deterministic entity-centric retrieval tasks. Consequently, the framework is most reliable when natural language questions can be mapped to explicit graph entities and structured geological attributes. More complex comparative and multi-hop reasoning scenarios remain areas for ongoing development.

To further assess performance, we compared three methods for answering a free-form query about rare earth element concentrations (LREE_PPM, HREE_PPM, REE_PPM, and REE_Y_PPM) for sample AU.1457811 (Fig. 9): Deterministic: Returned exact numerical values from the dataset; LLM: Interpreted the question using natural language to extract relevant values; Semantic: Provided contextual deposit information with metadata-based scoring and interpretive insights. This comparison highlights the strengths and trade-offs of each approach, showing how CMKG can deliver both precise data retrieval and broader semantic exploration, depending on user needs. These results illustrate that deterministic retrieval ensures numerical accuracy, while semantic retrieval and LLM synthesis provide complementary contextual support without contributing to factual correctness. Together, these experiments demonstrate that the CMKG framework supports flexible interaction beyond predefined query templates while maintaining reliable, structured data retrieval.

4. Discussion

The results demonstrate that the Critical Minerals Knowledge Graph (CMKG) successfully transforms the CMiO dataset into a scalable, queryable, and user-friendly analytical environment that supports complex geological and geochemical questions. The ingestion pipeline effectively handled more than 29,000 rows of heterogeneous data by applying a strict, ontology-derived schema and a modular batching strategy. This approach ensured type consistency, prevented duplication through MERGE operations, and enabled incremental updates without reprocessing the entire dataset. By serializing intermediate relationship dictionaries and supporting reconstruction through Neo4j, NetworkX, and embedding models, the pipeline provides the flexibility required for long-term, iterative graph growth. This robust foundation is essential for ensuring that downstream question-answering remains stable as the CMiO dataset evolves. Importantly, this deterministic ingestion process ensures that all downstream results are reproducible and grounded in a consistent, validated knowledge graph structure.

The curated set of 35 example queries highlights the system's ability to retrieve, relate, and contextualize geological and geochemical information across sample-level, deposit-level, and spatial dimensions. These structured tests show that the CMKG can resolve typical geoscience questions, such as elemental assays, analytical methods, mineral associations, and commodity groupings, while also handling more complex reasoning tasks, including identifying deposits with rare earth element signatures or tracing sample–deposit connections. Running these queries through the hybrid GraphRAG pipeline underscores the complementary strengths of deterministic Cypher retrieval, semantic similarity search, and LLM-based synthesis. Deterministic querying consistently yields authoritative, field-level answers, while FAISS provides contextual descriptions when relevant textual information exists, and LLM synthesis integrates both into a coherent, geologically meaningful narrative. These qualitative observations are consistent with the quantitative evaluation (Section 3.2), which shows that accuracy is primarily driven by deterministic graph querying, while the hybrid architecture preserves this accuracy and enhances interpretability.

The unified QA architecture is central to this performance. By combining deterministic parsing, typed builder functions, alias and synonym dictionaries, geographic disambiguation, and fallback semantic retrieval, the system can reliably interpret queries phrased in

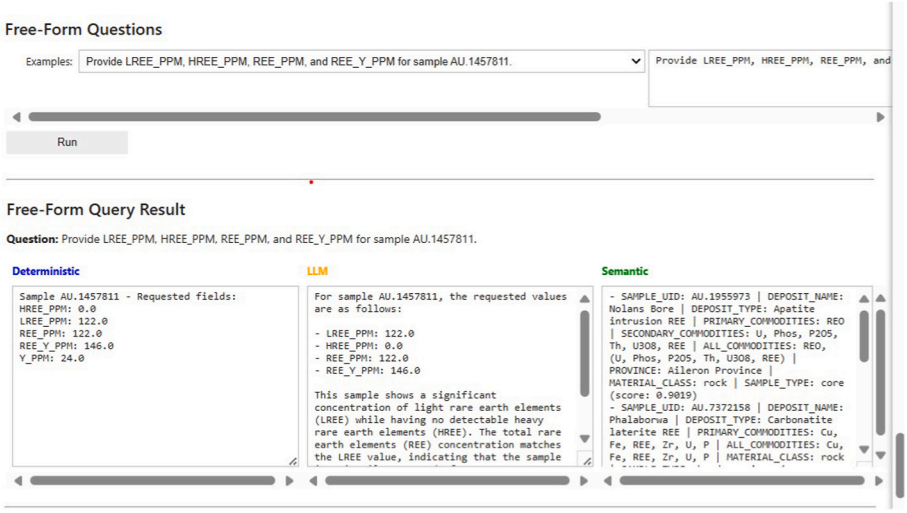

Fig. 9. Comparison of three query-processing approaches for a free-form analytical request. The example query, "Provide LREE_PPM, HREE_PPM, REE_PPM, and REE_Y_PPM for sample AU.1457811" is resolved using deterministic retrieval, LLM-based interpretation, and semantic similarity search. The deterministic panel (left) returns exact numerical values directly from the dataset. The LLM panel (center) reformulates these values into a concise natural-language summary, adding interpretive context regarding rare earth element distributions. The semantic panel (right) retrieves contextually related samples and deposit-level metadata, such as deposit name, type, province, associated commodities, and relevance scores, providing broader geological context beyond the requested numeric fields. Together, the figure illustrates how the CMKG system integrates precise data access with interpretive reasoning and contextual discovery.

diverse, domain-specific language. The ability to resolve shorthand chemical symbols, multi-element commodity groups, analytical terminology, and state abbreviations indicates a high degree of linguistic robustness. The dispatcher function ensures that deterministic methods take precedence when structured data are available, while semantic retrieval and LLM interpretation provide resilience when queries fall outside predefined patterns. This layered logic results in consistently stable behavior, even when user phrasing varies widely or queries span multiple conceptual domains. This explicit prioritization of deterministic retrieval addresses concerns regarding the reliability of LLM-based systems by ensuring that factual outputs are derived from structured data rather than generated heuristically.

The interactive Jupyter-based interface further demonstrates how these capabilities can be made accessible to a broad user community. Its design, featuring guided queries, free-form input, live examples, and instructional prompts, lowers the barrier to engaging with graph-based data structures. The three-panel output display is particularly important: by visually separating deterministic, semantic, and LLM responses, the interface increases transparency and allows users to understand how each method contributes to the final answer. This explicit differentiation supports scientific accountability, enabling researchers to distinguish empirically retrieved values from context-based semantic reasoning or LLM-generated narrative. The system thus provides both accessibility for non-experts and interpretive depth for advanced users. This transparency mechanism also supports qualitative validation of system outputs, complementing the quantitative benchmark evaluation.

The evaluation of free-form queries highlights the system's flexibility in addressing real-world geological questions beyond predefined templates. Queries span seven key categories, sample metadata, analytical values, commodities, deposit-level details, locations, drill/provenance data, and identity-based lookups, covering tasks such as retrieving sampling methods, analytical results, commodity associations, deposit sources, coordinates, analysis dates, and unique identifiers. Together, these categories reflect common workflows in mineral exploration and geoscience research. The system's ability to return precise numeric values, identify commodity associations, resolve geographic regions, and provide deposit-level context demonstrates that the hybrid architecture generalizes beyond controlled examples. The three-way comparison of deterministic, semantic, and LLM responses for rare earth element concentrations is particularly illustrative: deterministic logic returns exact values; the LLM reframes these values in a geological context; and semantic retrieval supplies deposit-level insight and relevance scoring. Together, these responses show how the CMKG can serve both as a factual data retrieval engine and as a tool for interpretive geological reasoning. This ability to handle both structured and open-ended queries addresses the challenge of out-of-scope or "out-of-bag" questions raised by the reviewers.

Overall, the results indicate that the CMKG's integrated architecture, spanning ingestion, structured querying, semantic search, LLM reasoning, and interactive visualization, successfully enables multimodal exploration of a large and complex mineral dataset. The system's performance across structured examples and open-ended queries confirms that knowledge graph–based representations, when combined with hybrid retrieval and LLM synthesis, provide a powerful foundation for answering diverse geoscience questions. This approach improves data accessibility, enhances interpretive capabilities, and supports both expert and non-expert users in navigating critical mineral information. The demonstrated robustness, scalability, and transparency suggest strong potential for future extensions, including more advanced graph analytics, uncertainty quantification, and integration with additional geoscience datasets. Nevertheless, current limitations include dependence on the completeness of the underlying schema and the availability of structured attributes, which may affect performance for queries involving missing or sparsely represented data.

This work builds on prior efforts by overcoming the limitations of manual analysis, isolated relational databases, and standalone knowledge graphs. Traditional systems, while useful for structured storage, lack the flexibility to integrate heterogeneous attributes and support multi-relational reasoning. Our approach advances beyond these constraints by embedding the CMiO dataset into a semantically rich knowledge graph and coupling it with GraphRAG for hybrid retrieval and LLM-driven synthesis. This combination enables natural language interaction, multi-hop reasoning, and contextual interpretation; capabilities absent in conventional database-centric workflows. By transforming static tables into an interactive, explainable knowledge ecosystem, the framework not only improves data accessibility but also accelerates pattern discovery and decision-making in critical mineral research. These contributions represent a significant step toward scalable, transparent, and semantically grounded exploration of complex geoscience datasets. In contrast to LLM-only systems, the proposed framework ensures that all generated responses remain anchored to verifiable data sources, thereby reducing the risk of hallucination.

The Critical Minerals Knowledge Graph (CMKG) constructed in this work serves as a foundational layer for advanced graph analytics that can uncover deeper insights into mineral systems. Future work will leverage this structure for techniques such as community detection to identify clusters of related deposits, link prediction to infer missing relationships, centrality analysis to rank influential minerals or deposits, and graph embeddings for machine learning tasks like anomaly detection and resource classification. These capabilities will enable predictive modeling and hypothesis generation, extending the system beyond its current question-answering framework. Future work will also explore integration with additional datasets and incorporation of temporal and multi-source data to address more complex geological scenarios.

5. Conclusions

This study demonstrates that a knowledge graph–driven architecture, combined with hybrid retrieval and LLM-based reasoning, can substantially enhance access to and interpretation of complex critical mineral datasets. CMKG knowledge graph establishes a robust foundation for representing geological, geochemical, spatial, and contextual relationships in an integrated framework by converting the CMiO dataset into a semantically structured graph of more than 29,000 records. The ingestion pipeline, featuring strict schema enforcement, modular batching, and incremental update capabilities, ensures scalability, reliability, and long-term maintainability as new data are incorporated. The deterministic-first design ensures that all core results are reproducible and grounded in structured data.

The system consistently provides accurate, context-aware responses across 35 structured queries and a broad set of free-form natural language questions. Deterministic graph querying delivers authoritative factual retrieval, semantic search contributes relevant contextual information when direct matches are unavailable, and an LLM-based synthesis layer unifies these outputs into coherent geological explanations. This hybrid approach enables users to navigate mineralogical, geochemical, and spatial relationships in ways that traditional relational databases or standalone LLMs cannot achieve. The side-by-side comparison of deterministic, semantic, and generative outputs further enhances transparency and fosters trust in the system's reasoning.

Quantitative evaluation demonstrates that deterministic graph querying is the primary driver of reliable factual retrieval within the CMKG framework. The benchmark evaluation achieved strong performance for entity-centric geological questions, with Exact Match and Contains Match scores of 88.9% and 94.4%, respectively, for entity lookup tasks. These results confirm that structured Cypher-based retrieval provides accurate and reproducible access to geological attributes stored within the knowledge graph. At the same time, the evaluation highlights the continued difficulty of aggregation, comparative, and multi-hop reasoning over heterogeneous geological datasets. These findings emphasize both the strengths and current limitations of hybrid geological question-answering systems and provide a transparent

foundation for future work involving advanced reasoning, schema expansion, and automated graph inference.

The interactive interface makes these capabilities accessible to both technical and non-technical users, allowing exploration of CMKG without requiring knowledge of Cypher, SQL, or the underlying schema. By accommodating diverse question types, ranging from elemental assays and deposit attributes to provenance, sampling history, and geographic queries, the system supports real-world workflows in mineral exploration, research, and resource assessment.

Overall, the CMKG framework illustrates the value of combining structured graph representations with modern retrieval and language technologies to unlock rich scientific insights from large geoscience datasets. The integration of strict data governance, hybrid question answering, and interpretable user interfaces positions the CMKG as a practical tool for critical mineral research and provides a scalable foundation for future extensions, including automated analytics, cross-dataset integration, and advanced mineral systems modeling. By explicitly separating deterministic retrieval from generative synthesis, the framework provides a transparent and reproducible alternative to purely LLM-based approaches.

6. Code availability statement

The Jupyter notebook used to construct the Critical Minerals Knowledge Graph (CMKG), including the full data ingestion pipeline, example queries, hybrid QA architecture, and the interactive Jupyter-based interface, is openly available in a public GitHub repository at: https://github.com/adavarpa/Critical-Minerals-Knowledge-Graph-CMKG- The repository includes the complete Jupyter notebook used for data processing, graph construction, and system evaluation, ensuring that all analyses presented in this work can be reproduced and extended by researchers.

CRediT authorship contribution statement

Armita Davarpanah: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – original draft. **Hassan A. Babaie:** Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – review & editing. **W. Crawford Elliott:** Data curation, Formal analysis, Resources, Validation, Writing – review & editing. **Yuanzhi Tang:** Resources, Validation, Writing – review & editing. **Paul A. Schroeder:** Resources, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

https://github.com/adavarpa/Critical-Minerals-Knowledge-Graph-CMKG-

References

Babaie, H.A., Davarpanah, A., Elliott, W.C., 2022. Ontology of the complex rare-earth elements mineral system. In: Ma, X., Mookerjee, M., Hsu, L., Hills, D. (Eds.), Recent Advancement in Geoinformatics and Data Science, vol. 558. Geological Society of America Special Paper, pp. 29–44. https://doi.org/10.1130/2022.2558(03.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., Choi, Y., 2019. COMET: commonsense transformers for automatic knowledge graph construction. AAAI Conference on Artificial Intelligence v2. https://ar5iv.labs.arxiv.org/html/1906.05317.

Brantley, Susan L., Wen, Tao, Agarwal, Deborah, Catalano, Jeffrey G., Schroeder, Paul A., Lehnert, Kerstin, Varadharajan, Charuleka, Pett-Ridge, Julie, Engle, Mark, Castronova, Anthony M., Hooper, Richard P., Ma, Xiaogang, Jin, Lixin, McHenry, Kenton, Aronson, Emma, Shaughnessy, Andrew R., Derry, Louis A., Richardson, Justin, Bales, Jerad, Pierce, Eric M., 2021. The future low-temperature geochemical datascape envisioned by the U.S. geochemical community. Comput. Geosci. https://doi.org/10.1016/j.cageo.2021.104933.

Case, G.N.D., Graham, G.E., Lawley, C.J.M., Bastrakov, E., Huston, D.L., Hofstra, A.H., Lisitsin, V., Hawkins, S.G., Wang, B., 2025. Critical minerals in ores (CMiO) database. U.S. Geological Survey Fact Sheet 2025–3002. https://doi.org/10.3133/fs20253002.

Chen, Y., Tian, M., Wu, Q., Tao, L., Jiang, T., Qiu, Q., Huang, H., 2024. A deep learning-based method for deep information extraction from multimodal data for geological reports to support geological knowledge graph construction. Earth Sci Inform 17, 1867–1887. https://doi.org/10.1007/s12145-023-01207-0.

Davarpanah, A., Babaie, H.A., Elliott, W.C., 2024. Knowledge-based query system for the critical minerals. Applied computing and Geosciences 22 (2024), 100167. https://doi.org/10.1016/j.acags.2024.100167.

EarthChem Portal, 2025. EarthChem: a hub for geochemical, geochronological, and petrological data. Interdisciplinary Earth Data Alliance (IEDA). Retrieved from. https://earthchem.org/.

Feng, Q., Zhao, T., Liu, C., 2024. A "pipeline"-based approach for automated construction of geoscience knowledge graphs. Minerals 14 (12). https://doi.org/10.3390/min14121296. Article 1296.

Fortier, S.M., Nassar, N.T., Graham, G.E., Hammarstrom, J.M., Day, W.C., Mauk, J.L., Seal, I.I.R.R., 2021. USGS critical minerals review 2021. Min. Eng. 74 (5), 34–48. https://pubs.usgs.gov/publication/70233196.

Fortier, S.M., Nassar, N.T., Lederer, G.W., Brainard, Jamie, Gambogi, Joseph, McCullough, E.A., 2018. Draft critical mineral list—Summary of methodology and background information—U.s. geological survey technical input document in response to secretarial order No. 3359. U.S. Geological Survey Open-File Report 2018–1021 15. https://doi.org/10.3133/ofr20181021.

Guo, Z., Xia, L., Yu, Y., Ao, T., Huang, C., 2024. LightRAG: simple and fast retrieval-augmented generation. arXiv:2410.05779. https://arxiv.org/html/2410.05779v1#bib.

Han, H., Wang, Y., Shomer, H., Guo, K., Ding, J., Lei, Y., Halappanavar, M., Rossi, R.A., Mukherjee, S., Tang, X., He, Q., Hua, Z., Long, B., Zhao, T., Shah, N., Javari, A., Xia, Y., Tang, J., 2025. Retrieval-augmented generation with graphs (GraphRAG). arXiv preprint arXiv:2501.00309. https://arxiv.org/abs/2501.00309.

Haxel, G.B., Hedrick, J.B., Orris, G.J., Stauffer, P.H., Hendley, J.W.I.I., 2002. Rare Earth elements: critical resources for high technology (USGS fact sheet 087-02). U.S. Department of the Interior, U.S. Geological Survey. https://doi.org/10.3133/fs08702.

Hofstra, A., Lisitsin, V., Corriveau, L., Paradis, S., Peter, J., Lauzière, K., Lawley, C., Gadd, M., Pilote, J., Honsberger, I., Bastrakov, E., Champion, D., Czarnota, K., Doublier, M., Huston, D., Raymond, O., VanDerWielen, S., Emsbo, P., Granitto, M., Kreiner, D., 2021. Deposit classification scheme for the critical minerals mapping initiative global geochemical database. U.S. Geological Survey Open-File Report 2021–1049 60. https://doi.org/10.3133/ofr20211049.

Hofstra, A.H., Kreiner, D., Granitto, M., Emsbo, P., Lisitsin, V., Corriveau, L., et al., 2022. Update on the deposit classification scheme for the critical minerals in ores database. (U.S. Geological Survey, preliminary report.).

Hofstra, A.H., Kreiner, D.C., 2020. Systems-deposits-commodities-critical minerals table for the Earth mapping resources initiative. USGS Open-File Report. https://doi.org/10.3133/ofr20201042, 2020–1042.

Hogan, A., et al., 2021. Knowledge graphs. ACM Comput. Surv. 54 (4). https://doi.org/10.1145/3447772. Article 71.

Jatana, N., Puri, S., Ahuja, M., Kathuria, I., Gosain, D., 2012. A survey and comparison of relational and non-relational database. Int. J. Eng. Res. Technol. 1 (6), 1–5.

Johnson, J., Douze, M., Jégou, H., 2019. Billion-scale similarity search with GPUs. IEEE Trans. Big Data 7 (3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572.

Jowitt, S., Mudd, G.M., Thompson, J.F., 2020. Future availability of non-renewable metal resources and the influence of environmental, social, and governance conflicts on metal production. Commun. Earth Environ. 1–8. https://doi.org/10.1038/s43247-020-0011-0. Nature.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., Kiela, D., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv preprint arXiv:2005.11401. https://arxiv.org/abs/2005.11401.

Lin, Z., Deng, C., Zhou, L., Zhang, T., Xu, Y., Xu, Y., He, Z., Shi, Y., Dai, B., Song, Y., Zeng, B., Chen, Q., Miao, Y., Xue, B., Wang, S., Fu, L., Zhang, W., He, J., Zhu, Y., Wang, X., Zhou, C., 2024. GeoGalactica: a Scientific Large Language Model in Geoscience arXiv preprint arXiv:2401.00434.

Ma, X., 2022. Knowledge graph construction and application in geosciences: a review. Comput. Geosci. 161 (2022), 105082. https://doi.org/10.1016/j.cageo.2022.105082, 1-15. ISSN 0098-3004.

Ma, J., Zhou, Y., he, L., Zhang, Q., Atif Bilal, M., Zhang, Y., 2025. Enhancing geological knowledge engineering with retrieval-augmented generation: a case study of the Qin–Hang metallogenic belt. Minerals 2025 15 (10), 1023. https://doi.org/10.3390/min15101023.

Macrostrat Database, 2025. Macrostrat: Geological Map and Stratigraphic Data Platform. University of Wisconsin, Madison. Available at: https://macrostrat.org.

Meta, A.I., 2025. FAISS: a library for efficient similarity search and clustering of dense vectors. https://faiss.ai/.

Mousavi, W.M., Alghisi, S., Riccardi, G., 2025. LLMs as repositories of factual knowledge: limitations and solutions. IEEE/ACM Trans. Audio Speech Lang. Process. XX (X), 1–13. https://doi.org/10.48550/arXiv.2501.12774.

Neo4j, 2025. https://neo4j.com/download/.

Noy, N.F., Gao, Y., Jain, A., Narayanan, A., Patterson, A., Taylor, J., 2019. Industry-scale knowledge graphs: lessons and challenges. Commun. ACM 62 (8), 36–43. https://doi.org/10.1145/333116.

Ontotext, 2025. What is knowledge graph? https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/.

OpenAI, 2024. Text-Embedding-3-Large. OpenAI platform documentation. https://platform.openai.com/docs/models/text-embedding-3-large.

Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Tang, S., 2024. Graph retrieval-augmented generation: a survey. arXiv preprint arXiv:2408.08921. https://arxiv.org/abs/2408.08921.

Qiu, Q., Wang, B., Ma, K., Lü, H., Tao, L., Xie, Z., 2023. A practical approach to constructing a geological knowledge graph: a case study of mineral exploration data. J. Earth Sci. 34 (5), 1374–1389. https://doi.org/10.1007/s12583-023-1809-3.

Schroeder, P.A., Elliott, W.C., Tang, Y., Lemke, L., 2024. Facilitating the critical mineral future: valorization of kaolin mining waste through partnerships. GSA Today 34, 60–61. https://doi.org/10.1130/GSATG599GW.1.

UN, 2025. Harnessing the potential of critical minerals for sustainable development. World Economic Situation and Prospects 2025.

USGS, 2022. 2022 final list of critical minerals. Fed. Regist. 87 (37), 10381–11038. https://www.usgs.gov/news/national-news-release/us-geological-survey-releases-2022-list-critical-minerals.

USGS, 2025. 2025 final list of critical minerals. Federal register. Fed. Regist.: Final 2025 List of Critical Minerals.

Van Rossum, G., 2020. The Python library reference (v3.8.2). Python Software Foundation. https://docs.python.org/3/library/pickle.html.

Yasunaga, M., Ren, H., Bosselut, A., Liang, P., Leskovec, J., 2022. Deep bidirectional language–knowledge pretraining. NeurIPS. https://github.com/michiyasunaga/dragon.

Zhang, L., Cheng, Z., Hao, Z., Zuo, K., Lui, K., 2025. Integrating inflow control valve control with LSTM networks for oil production forecasting in horizontal intelligent well application. J. Pet. Explor. Prod. Technol. 15, 5. https://doi.org/10.1007/s13202-025-01964-2.

Zhou, C., Wang, H., Wang, C., Hou, Z., Zheng, Z., Shen, S., Cheng, Q., Feng, Z., Wang, X., Lv, H., Fan, J., Hu, X., Hou, M., Zhu, Y., 2021. Geoscience knowledge graph in the big data era. Sci. China Earth Sci. 64, 1105–1114. https://doi.org/10.1007/s11430-020-9750-4.

Zhou, B., Li, K., 2025. Fusing geoscience large language models and lightweight RAG for enhanced geological question answering. Geosciences 2025 15 (10), 382. https://doi.org/10.3390/geosciences15100382.

Zhu, Y., Wang, Q., Wang, S., Sun, K., Wang, X., Lv, H., Hu, X., Zhang, J., Wang, B., Qiu, Q., Yang, J., Zhou, C., 2025. Methodology, progress and challenges of geoscience knowledge graph in international big science program of deep-time digital Earth. J. Geogr. Sci. 35, 1132–1156. https://doi.org/10.1007/s11442-025-2361-0.