

GraphRAG on Technical Documents – Impact of Knowledge Graph Schema

Henri Scaffidi 

School of Computer Science and Software Engineering, The University of Western Australia, Australia

Melinda Hodkiewicz 

School of Mechanical Engineering, The University of Western Australia, Australia

Caitlin Woods 

School of Computer Science and Software Engineering, The University of Western Australia, Australia

Nicole Roocke 

Minerals Research Institute of Western Australia, Australia

Abstract

Retrieval Augmented Generation (RAG) is seeing rapid adoption in industry to enable employees to query information captured in proprietary data for their organisation. In this work, we test the impact of domain-relevant knowledge graph schemas on the results of Microsoft's GraphRAG pipeline. Our approach aims to address the poor quality of GraphRAG responses on technical reports rich in domain-specific terms. The use case involves technical reports about geology, chemistry and mineral processing published by the Minerals Research Institute of Western Australia (MRIWA). Four schemas are considered: a simple five-class minerals domain expert-developed schema, an expanded minerals domain schema, the Microsoft GraphRAG auto-generated schema, and a schema-less GraphRAG. These are compared to a conventional baseline RAG. Performance is evaluated using a scoring approach that accounts for the mix of correct, incorrect, additional, and missing content in RAG responses. The results show that the simple five-class minerals domain schema extracts approximately 10% more entities from the

MRIWA reports than the other schema options. Additionally, both the five-class and the expanded eight-class minerals domain schemas produce the most factually correct answers and the fewest hallucinations. We attribute this to the minerals-specific schemas extracting more relevant, domain-specific information during the Indexing stage. As a result, the Query stage's context window includes more high-value content. This contributes to the observed improvement in answer quality compared to the other pipelines. In contrast, pipelines with fewer domain-related entities in the KG retrieve less valuable information, leaving more room for irrelevant content in the context window. Baseline RAG responses were typically shorter, less complete, and contained more hallucinations compared to our GraphRAG pipelines. We provide a complete set of resources at <https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main>. These resources include links to the MRIWA reports, a set of questions (from simple to challenging) along with domain-expert curated answers, schemas, and evaluations of the pipelines.

2012 ACM Subject Classification Information systems → Specialized information retrieval

Keywords and phrases RAG, minerals, local search, global search, entity extraction, competency questions

Digital Object Identifier 10.4230/TGDK.3.2.3

Category Research

Supplementary Material *InteractiveResource*: <https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain> [30]

Funding *Nicole Roocke*: MRIWA Grant No. M10530. DOI: <https://doi.org/10.71342/574784944778>

Received 2025-03-14 **Accepted** 2025-08-13 **Published** 2025-10-15



© Henri Scaffidi, Melinda Hodkiewicz, Caitlin Woods, and Nicole Roocke;
licensed under Creative Commons License CC-BY 4.0

Transactions on Graph Data and Knowledge, Vol. 3, Issue 2, Article No. 3, pp. 3:1–3:24



Transactions on Graph Data and Knowledge

TGDK Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Recent advancements in Large Language Models (LLMs) have revolutionised AI’s capability to process and understand natural language, demonstrating strong performance in complex practical tasks. However, LLMs are often limited to their pre-trained knowledge and therefore falter when additional technical domain-specific information is required. Retrieval Augmented Generation (RAG) systems address this issue by enabling the LLM to access and incorporate information from a retrieval source – structured or unstructured knowledge such as documents, tables, or graphs. This process grounds the LLM’s responses in additional, domain-specific information that complements its pre-trained dataset. RAG operates through two steps: retrieval and generation. Retrieval involves searching through the retrieval source to rank the information based on its relevance to the query. The most relevant portions are then passed to the LLM for the generation step, where they are combined with the original query to produce more targeted responses. Knowledge graphs (KGs) store explicit knowledge as structured data and are used to enhance the retrieval step [11, 25].

Industry’s challenge is that unstructured technical reports contain unstandardised formatting, a mix of text, graphics, and tables, and language requiring domain expertise for interpretation. A specific example of this is MRIWA (Minerals Research Institute of Western Australia). This organisation has a 40-year archive of such reports holding valuable data and insights for the state’s minerals industry and research community. Work to understand the impact of different KG schemas on the performance of KG-RAG on complex technical texts is in its infancy. Complex (non-academic) technical texts written by and for industry users are challenging to source as organisations are often reluctant to release them. To use these reports for reproducible KG-RAG research, the documents must be accompanied by KG schemas, question and answer sets, and pipeline answer evaluations by domain experts. This project seeks to address these challenges by evaluating the impact of different KG schemas on the quality of generated responses to a set of curated queries for MRIWA’s minerals research technical reports. We release all the artifacts at <https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main> to support FAIR¹ research.

The paper is organised as follows. Section 2 reviews the literature on RAG, identifies gaps, and defines the research question. Section 3 describes the structure and content of MRIWA’s reports. Section 4 outlines the process of data selection, pipeline functionality, and the experimental setup. Section 5 analyses the results of the experiment.

2 Literature Review

2.1 Overview of Retrieval Augmented Generation

The RAG paradigm is still evolving and can be broadly classified into three processes: baseline RAG, advanced RAG and modular RAG [25]. Baseline RAG has two steps: retrieval and generation. In the retrieval step, text from the document is segmented into chunks, then encoded using an embedding model and stored in a vector database. The RAG system uses the same encoding model for the user query and computes a similarity score between the vectors for the query and the indexed chunk. The top-k chunks with the highest similarity score are used to generate a response to the prompt. Baseline RAG, or “Naïve” RAG, often struggles to identify appropriate chunks, especially when the query requires large amounts of information (exceeding the top-k restriction) to be retrieved and then aggregated across documents, and when the query requires the system to synthesise insights not explicitly stated in reports [11].

¹ FAIR stands for Findable, Accessible, Interoperable, and Reusable.

Advanced RAG uses pre-retrieval and post-retrieval strategies to improve indexing and streamline the retrieval process. For example, there is growing interest in utilising additional content, specifically structured data such as knowledge graphs, to enhance the retrieval stage in RAG. This is called GraphRAG [26, 27]. Specifically, altering the retrieval source structure, from unstructured text to graph-based knowledge, for example, can improve the quality of retrieved content, which helps to reduce LLM hallucinations [16] and improve summarisation performance [7]. A modular RAG framework introduces new modules into the pipeline, allowing for substitution or reconfiguration of specific steps, making the process more adaptable to specific use-cases beyond traditional sequential retrieval and generation [11].

2.2 Domain-Specific Knowledge Graph Schemas and Construction

A KG is a collection of triples, where each triple represents a relation (edge) between two entities (nodes): $relation(ENTITY1, ENTITY2)$. The KG schema defines the types of entities and relations permissible in the KG. GraphRAG relies on a KG to retrieve relevant information. However, manually curating a KG is labor-intensive and time-consuming, often relying on crowd-sourced efforts like Freebase [3] and Wikidata [35]. As a result, automated KG construction methods are crucial for efficiently extracting structured knowledge from unstructured or semi-structured data, significantly reducing the manual burden [42]. The challenge lies in designing a KG schema that balances domain specificity with adaptability, ensuring the KG can incorporate new concepts, evolving terminology, and varied technical document structures.

Broadly, there are two approaches to KG schema development: top-down and bottom-up. The top-down approach develops a schema based on domain knowledge and the use case, sometimes informed by existing ontologies if available [24]. Bottom-up approaches identify concepts of interest using entity discovery processes to discover strings referring to semantic entities and then classify them into types, which collectively define the schema [36, 41, 42]. In practice, a combination of both approaches is usually deployed.

Within the minerals domain are several technical disciplines, each with its own processes, norms, and data models. These disciplines contribute at different stages along the mineral processing value chain – find and define the deposit (geology, geochemistry, geostatistics), consider options for processing (geometallurgy), identify risks and quantify the value (mining finance), mine planning and execution (mining engineering), mineral processing (metallurgists) to produce the saleable product (marketing), environmental management and closure (environmental engineers). These disciplines draw on the sciences – including geology, chemistry, physics, and mathematics – as well as all branches of engineering and business. While there has been work on developing schemas and ontologies in subdomains such as geology [6, 12], there is no publicly available KG schema or ontology that spans the entire minerals domain value chain. An ontology is a formal and semantically rich specification used when reasoning and interoperability are required.

2.3 KG-RAG Retrieval Processes

As the literature on RAG has grown, more studies have sought to integrate KGs into the pipeline, particularly as the field acknowledges the limitations of vector search over unstructured text [11]. Broadly, these efforts can be categorised into two approaches: using LLM-generated KG queries [10, 17, 18, 37, 39], and machine learning-based KG querying [15, 16, 23].

2.4 Microsoft’s GraphRAG

Microsoft’s GraphRAG [7], introduced in 2024, uses LLMs to generate the KG from text. Graph machine learning is utilised to build a bottom-up hierarchical cluster of the KG using an LLM to pre-summarise each cluster and produce a set of community reports. The KG and community reports are retrieved via two search options: **local search** and **global search**. Local search identifies semantically related entities within the KG, based on the prompt, and retrieves closely connected information. This search method is suited to answer targeted queries about specific entities [19]. Global search retrieves all community reports at a specific level of the hierarchical cluster and conducts Map-Reduce to synthesise a response. This search method is designed to answer high-level aggregation queries about the dataset.

Given the novelty of GraphRAG (released July 2, 2024), there is no research testing it in the minerals domain. As shown in Table 1, the published literature on Microsoft GraphRAG is limited. In our analysis, surveys were excluded, and only studies that evaluated Microsoft GraphRAG’s question–answering performance were included. Across the 13 identified studies, most used broad, general-domain public datasets such as UltraDomain or HotpotQA, while only a small proportion applied GraphRAG to narrow industrial and technical domain corpora. LLM-based judging was the predominant evaluation approach, with few examples of human assessment or automated retrieval metrics. Citation counts are generally low, and impact factors, where applicable, are modest. These trends highlight a gap in the literature, which our work addresses by evaluating GraphRAG’s performance in a real-world context within the minerals domain using human subject-matter expert evaluation.

■ **Table 1** Summary statistics from published studies evaluating Microsoft GraphRAG ($n = 13$).

Metric	Value
Average citations (Google Scholar, Aug 2025)	1.15
Average impact factor (where reported, $n = 2$)	6.15
Papers using general-domain public datasets rather than narrow industrial and technical domain corpora	9 out of 13
Papers using LLMs as evaluators of pipeline performance	9 out of 13

Many organisations are seeking to use RAG to query their internal documents, and given their existing commitment to the Microsoft ecosystem, Microsoft’s GraphRAG has become a relevant product worth investigating.

2.5 Research Question

Based on the gaps in the literature discussed in this section we propose the following research question: Given a set of minerals research reports that are accessed in a GraphRAG pipeline, how do different KG schemas of the report’s subject matter impact the quality of responses to local and global search queries?

3 Data

The Minerals Research Institute of Western Australia (MRIWA) is a statutory government body whose purpose is to foster and promote minerals research for the benefit of the State of Western Australia. The organisation provides financial assistance to industry and research entities, with project outcomes typically presented in technical reports. MRIWA has accumulated a portfolio of over 300 such reports, each detailing project undertakings and findings [22].

Currently, querying the reports is limited to keyword search on project abstracts on the MRIWA website, enabling users to identify relevant reports which are then available for download. This approach is often ineffective, as information in the full reports may not be adequately captured in their abstracts. Once reports are downloaded, the process of locating specific information in the full reports can be challenging due to their technical content, length (the mean number of pages is 163, with approximately 7% of reports more than 400 pages long) and inconsistent formats. These issues create a challenge for MRIWA, their stakeholders, and research community members to extract data and insights that are relevant to the sector.

To illustrate the complexity of the report content, 50 text samples are analysed. Each sample is randomly selected from a page in the MRIWA report repository. The samples are a minimum of one sentence in length and consist solely of body text. Based on the samples, the academic content in the reports has an average Flesch Reading Ease of 45 out of 100, where plain English typically scores between 60 and 70 [32]. Specialised domain expertise is often required to interpret this technical content. Below are examples of typical text found in MRIWA’s reports:

“In the footwall Li zone, the absence of petalite and fluid inclusion work indicate formation of (feldspar-free) quartz-spodumene assemblage...” [38].

“The interaction with gangue may be advantageous (for example, pyrite may enhance copper leaching by increasing the fluid progress into the core, as was presented by Dutrizac et al. (1971)), or be detrimental (such as reagent consumption in carbonate dissolution (Sinclair and Thompson, 2015)), depending on the gangue type.” [13]

“The low shaft power needed for this type of vertical flow liquid-liquid contacting ensures a uniform drop size in the dispersion, resulting in very low entrainment (Ju et al. 1991).” [28]

MRIWA’s reports exist as either scanned PDFs or converted Word documents and their content includes tables, figures, and text. With no standardised format or template, significant formatting inconsistencies exist across all reports. Figure 1 highlights these inconsistencies, including variations in figure captions, column layouts, headers and footers, and section headings.

MRIWA seeks to support their stakeholders to query these reports using natural language, unlocking information in these historical documents to foster new insights for the WA minerals industry.

In addition to the MRIWA reports, the data set used for this project includes a set of competency questions and subject matter expert ideal answers.

3.1 Competency Questions and Ideal Answers

The validation input for our analysis consists of a set of competency questions to assess the effectiveness of a minerals domain RAG pipeline. The questions were developed by MRIWA’s CEO and several members of the MRIWA staff. As representatives of the typical users of the RAG pipeline, they provided questions that reflect real-world information needs in the minerals research sector. The quantity of information required to answer each question and the complexity of each question varies. We have categorised these questions into five groups based on the CRAG [40] classification scheme: simple, simple with condition, set, aggregation, and post-processing-heavy (see Table 2 and Table 3). The CRAG paper defines eight question types, five of which apply to our question set. The three CRAG question types not covered by the CEO’s questions are comparison, multi-hop, and false-premise.

For each competency question the MRIWA team developed an ideal answer². These ideal answers are designed specifically for use with our evaluation approach outlined in Section 4.5.

² https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/data/mriwa_cqa

■ **Table 2** Number of MRIWA competency questions by CRAG [40] classification.

Question Type	Description	Count
Simple	Questions asking for simple facts that are unlikely to change overtime, such as the birth date of a person and the authors of a book.	4
Simple with Condition	Questions asking for simple facts with some given conditions, such as stock prices on a certain date and a director's recent movies in a certain genre.	2
Set	Questions that expect a set of entities or objects as the answer (e.g., "what are the continents in the southern hemisphere?").	4
Aggregation	Questions that require aggregation of retrieval results to answer (e.g., "how many Oscar awards did Meryl Streep win?").	2
Post-Processing-Heavy	Questions that need reasoning or processing of the retrieved information to obtain the answer (e.g., "how many days did Thurgood Marshall serve as a Supreme Court justice?").	3

■ **Table 3** MRIWA competency questions grouped by query type.

Simple
Identify which MRIWA reports reference MERIWA or MRIWA.
Which MRIWA reports has Commonwealth Scientific Industrial Research Organisation been involved with in any capacity (including being listed in references)?
Which MRIWA report is related to the East Kimberley region?
Which MRIWA reports relate to leaching?
Simple with Condition
Which MRIWA reports has Commonwealth Scientific Industrial Research Organisation been involved with as a researcher?
Which MRIWA reports has Commonwealth Scientific Industrial Research Organisation been involved with as a sponsor?
Set
Extract all references to MERIWA and MRIWA from the MRIWA reports.
Identify any references to nickel or Ni in the MRIWA reports.
Which elements are considered in the MRIWA reports?
Which regions of Western Australia are referenced in the MRIWA reports?
Aggregation
Which MRIWA report author has been involved in more than one report/project?
What is the average number of references in each MRIWA report?
Post-Processing-Heavy
Which MRIWA reports relate to exploration?
Which MRIWA reports relate to mining extraction?
Which MRIWA reports relate to mineral processing?

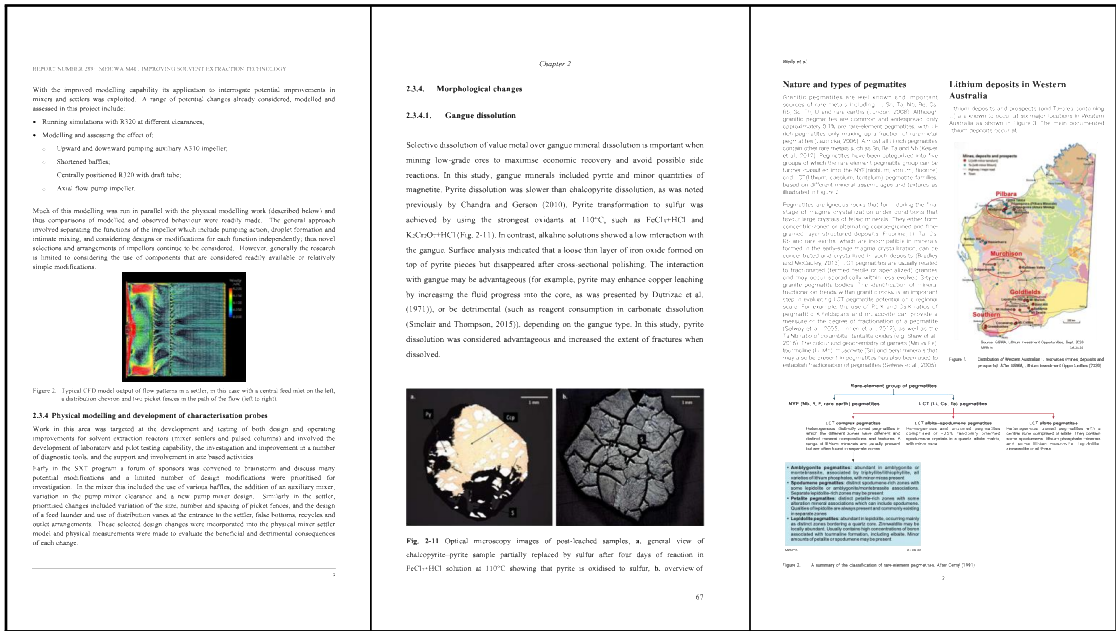


Figure 1 Examples of formatting inconsistencies across typical pages in MRIWA reports M0401 [28] (left), M0488 [13] (middle), and M0532 [38] (right). There is inconsistent formatting of figure captions, column layouts, headers and footers, and section headings, for example.

Some of the queries test GraphRAG’s ability to find and return concepts not explicitly referenced in the query. For example, the question “Which MRIWA report is related to the East Kimberley region” requires the model to identify towns, mine sites, and geological locations in the East Kimberley region, not rely on the regular expression *East Kimberley*. Similarly, for “reports relating to leaching” we hope to find reports relating to leaching from a minerals processing perspective (the context of the question) and which mention any chemical process that extracts valuable minerals from ore by dissolving with a solvent.

4 Process

This section describes sample selection, KG schema development, and the experimental setup, which is based on four KG schema options, each examined using the local and global search strategies in Microsoft GraphRAG.

4.1 MRIWA Report Sample Selection

A subset of MRIWA’s reports is sampled based on domain-expert opinion informed by natural language processing to count unique tokens, for instance. The selection process aims to ensure the data set covers multiple minerals domains and represents the complexity of the report set. A subset of 15 reports was selected, totalling approximately 2.7 million tokens – comparable in scale to the datasets used in the original Microsoft GraphRAG paper [7]. We believe this scale increases both the reliability of our results and the usefulness of the resources. The selection process is described in detail in the GitHub repository³. The selected PDF reports were converted to individual “.txt” files using the PyPDF2 [9] Python package for use in subsequent steps.

³ https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/blob/main/supplementary_materials/mriwa_report_sample_selection.md

4.2 Knowledge Graph Schema Development

The development of a minerals-domain KG schema began with a review of the concepts in MRIWA report executive summaries. These executive summaries (300-1000 words) provide an overview of topics and domain concepts covered in the full report content. Initially, we explored annotation classes (**Activity**, **Physical Object**, **Process**, **Property**, and **State**) used in our previous industrial annotation work⁴. This resulted in almost every item in the text having an annotation, and much of what was being annotated was irrelevant to addressing the questions being asked. We then explored discipline-specific (e.g., geology and mining) entities, but these were difficult to annotate consistently due to the spans. Following this experience, and with the assistance of MRIWA subject matter experts, we landed on more abstract concepts like **Naturally Occurring Object** to describe geological formations (present in many of the reports) and separated these from **Processed Object**, which are produced in some engineered process. We developed separate classes for objects, processes, location-specific concepts and for identifying organisations. We called this the Minerals Domain Schema (MDS). Early tests showed GraphRAG performed well with these aggregated concepts.

We then extended this schema to include some additional subclasses, splitting **Processed Object** into **Processed Material** and **Manufactured Product** (mirroring the mineral processing and manufactured output in mining), and sub-classing **Process** into **Natural Process**, **Lab testing Process** and **Industrial Process**. This was called the Expanded Minerals Domain Schema (EMDS). For a summary of these two schemas, see Table 4. We used our QuickGraph annotation tool⁵ [2] throughout this iterative process. Figure 2 provides a detailed example of entity typing under the EMDS from our manual annotation sessions. If the MDS were applied instead, the same entities would be labelled as follows:

- Entities labelled as **Natural_Process**, **Lab_Process**, or **Industrial_Process** would all be labelled as **Process**.
- Entities labelled as **Processed_Material** or **Manufactured_Product** would both be labelled as **Processed_Object**.

In considering the entity types in the MDS and EMDS, we also considered the possibility that future work might wish to align the schema classes to the Basic Formal Ontology (BFO) [14] or another top level ontology. Thus, the entity types **Naturally_Occurring_Object**, **Processed_Object** and **Manufactured_Product** can be mapped to the BFO class **bfo:MaterialEntity**, **Process** to **bfo:Process**, **Organisation** to **bfo:ObjectAggregate**, and **Site_Location_Boundary** to **bfo:ImmaterialEntity**.

Whilst defining relationships between entities would render the schemas more expressive, we do not, as GraphRAG [7] supports only the pre-definition of entity types. Relations are generated dynamically by the LLM during *Indexing* (see Section 4.3.1). Results of GraphRAG entity extraction using the schemas, including quantities of entities extracted, are discussed in Section 5.1.1.

4.3 GraphRAG

Microsoft’s GraphRAG approach [7] is designed to answer complex semantic queries and aggregation queries. Figure 3 shows the two GraphRAG phases: *Indexing* and *Query*.

⁴ <https://github.com/nlp-tlp/maintie>

⁵ <https://quickgraph.tech/>

■ **Table 4** Minerals Domain Knowledge Graph Schema and Expanded Minerals Domain Knowledge Graph Schema.

Minerals Domain Knowledge Graph Schema (MDS)		
Entity Type	Description	Superclass
Naturally Occurring Object	Objects formed naturally.	Object
Processed Object	Objects formed through human intervention.	Object
Process	Activities in which one or more objects participate.	N/A
Site Location Boundary	Immaterial spatial regions.	N/A
Organisation	Person/s organised together for a purpose.	N/A
Expanded Minerals Domain Knowledge Graph Schema (EMDS)		
Entity Type	Description	Superclass
Naturally Occurring Material	Objects formed naturally.	Object
Processed Material	Objects formed and materials used in an industrial process.	Object
Manufactured Product	Objects that are the product of a manufacturing process.	Object
Natural Process	Naturally occurring processes.	Process
Lab testing Process	Processes involving chemical and other analyses.	Process
Industrial Process	Processes that produce a product.	Process
Site Location Boundary	Immaterial spatial regions.	N/A
Organisation	Person/s organised together for a purpose.	N/A

4.3.1 GraphRAG Indexing

The *Indexing* stage takes our multi-domain subset of MRIWA reports and splits them into 300-token chunks with 100-token overlap. Utilising the specified schema, the LLM extracts both entities and relations from each chunk. Entities and relations are in the form `{name, type, description}` and `{source entity, target entity, description}` respectively. Entities with the same **name** and **type** are merged by the LLM to summarise all the descriptions into one summary per entity. Relations with the same **source entity** and **target entity** are merged in the same manner. This process yields a KG of MRIWA reports. An example of an entity and relation relating the concept *Rutile* is displayed in Table 5, depicting the detail that is captured through *Indexing*.

Community detection is conducted on the KG using the Leiden algorithm [34]. The algorithm is recursively applied to generate a hierarchical cluster of the graph. The LLM summarises each community into a community report. Each community report provides a summary of key points and outlines the important entities and relations present within the community. These community reports vary in their level of abstraction, depending on the level of the community within the hierarchy. The rationale, here, is to enable the LLM to answer high-level aggregation queries, hence addressing a baseline RAG weakness. The text chunks, entities, relations, and community reports are all vectorised and stored for retrieval during the *Query* stage.

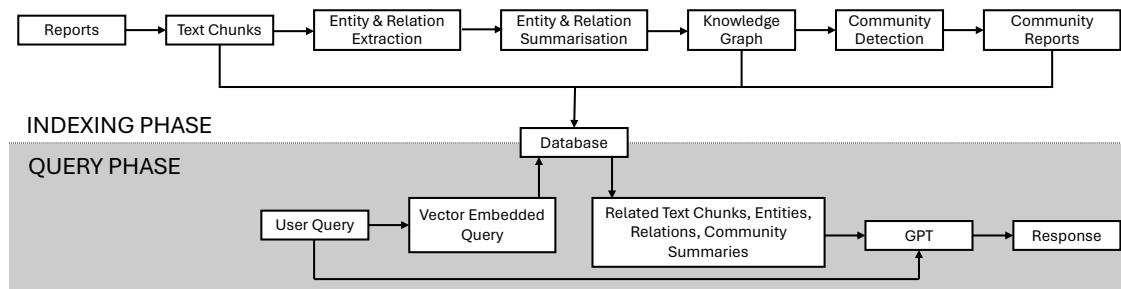
4.3.2 GraphRAG Query: Local and Global Search

The *Query* stage of GraphRAG comprises two search options: **local search** and **global search**. Local search vectorises the query and identifies a set of semantically related entities within the KG. Based on connectedness to these entities, additional candidate entities, relations, community

3:10 GraphRAG on Technical Documents – Impact of Knowledge Graph Schema

rutile (tio2) is a physically and chemically robust heavy mineral which inherit a geochemical fingerprint during formation that may unambiguously characterise its formation environment, in particular with respect to origin in a mineralise vs barren host rock. this project utilises the combination of these feature to assess the viability of use detrital rutile as a new tool for mineral exploration, focuss on gold and base metal. through a phd project fund by this project (see appendix), we has establish and fine - tune sample processing protocol to extract rutile grain from rock, unconsolidate sediment and polish thin section, identify rutile from other tio2 polymorph use ebsd - sem technique, and perform in-situ chemical and isotopic analysis on rutile use la-icp- ms and shrimp method. the in-situ analytical technique include establish glass and natural rutile standard to produce optimal geochemical datum. a geochemical database of rutile analysis from known formation environment has be establish and interrogate, the trace element fingerprint of rutile form in kalgoorlie - big bell - type orogenic gold ore system may be unambiguously distinguish from the chemical signature of other gold system and all other rutile - form environment. further, this chemical signature appear to survive regolith formation and later overprint event, except for extreme metasomatism, and is preserve in detrital grain derive from weather - erosion, transport, deposition and diagenesis. as such, use of detrital rutile in gold exploration may be comparable to diamond indicator mineral for diamond exploration, with similar transport scale from their source (i . e . 100 km). publish literature suggest that rutile from base metal ore may be similarly endow with a distinctive geochemical fingerprint. however, study of a limite number of western australian example suggest the validity of this or other geochemical fingerprint unique to base metal ore has yet to be confirm.

■ **Figure 2** Example of a manually annotated pre-processed executive summary from MRIWA Report 448 using the Expanded Minerals Domain Schema.



■ **Figure 3** Two phases of the Microsoft GraphRAG pipeline (derived from [19]).

reports, and text chunks are identified. These information sources are ranked and filtered, then passed to the context window of the LLM to produce a response. This search method is suited to answer targeted queries about specific entities [19].

In contrast, global search is designed to answer high-level aggregation queries about the dataset. First, all community reports at a specific level of the hierarchy are retrieved and chunked. In parallel, each chunk is used by the LLM to produce an intermediate response to the query. Each response is ranked by importance, where the most important are passed to the context window for the LLM to produce a final response.

In addition to local and global search, the GraphRAG framework includes other query methods such as **basic search** and **DRIFT search**. The basic search option is an implementation of baseline vector RAG and was used in our experiments for comparison. The DRIFT search method was released after our experiments were conducted and is therefore not evaluated.

4.4 Experimental Setup

The experiment utilises GraphRAG Version 0.1.1 [20] in its default configuration with GPT-4o-Mini (temperature: 0.0). We run GraphRAG *Indexing* in four separate pipelines, each using one of the following KG schemas as the `entity_types` parameter for entity and relation extraction.

■ **Table 5** Example GraphRAG entity and relation of rutile and gold exploration.

Entity Name	Entity Type	Entity Description
Rutile	Naturally Occurring Object	Rutile is a naturally occurring mineral primarily composed of titanium dioxide (TiO2) and is significant in various geological contexts. It is commonly found in metamorphic and igneous rocks, including sedimentary deposits, and is characterized by its high refractive index, strong dispersion, and distinct chemical properties. Rutile often occurs...
Relation Source Entity	Relation Target Entity	Relation Description
Rutile	Gold Exploration	Rutile is a mineral that plays a significant role in gold exploration. It is studied for its trace element signatures, which provide insights into the conditions under which it formed. As an indicator mineral, rutile assists in identifying potential gold deposits , making it a valuable tool in the gold exploration process . Its properties allow it to serve as a geochemical indicator, helping geologists locate areas that may contain gold. Overall, rutile's...

1. **Minerals Domain Schema (MDS):**

See Table 4. The LLM is not told the descriptions of the schema’s entity types, only the names of the entity types. Pipeline configuration⁶.

2. **Expanded Minerals Domain Schema (EMDS):**

See Table 4. The LLM is not told the descriptions of the schema’s entity types, only the names of the entity types. Pipeline configuration⁷.

3. **Auto-Generated Schema (AGS):**

This schema is generated by an LLM through the GraphRAG Prompt Tuner. We utilise the default LLM prompts, not the auto-tuned LLM prompts, to ensure a fair comparison between the other schema pipelines. We only utilise the entity type names generated by the Prompt Tuner, as follows.

{chemical process, mineral, geological survey, geochemistry, exploration technique, sample, formation, project, research study}

Pipeline configuration⁸.

4. **Schema-less (SL):**

This pipeline leaves the KG schema unspecified, and instead prompts the LLM to “identify all entities needed from the text in order to capture the information and ideas in the text”. Pipeline configuration⁹.

⁶ https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/src/minerals_domain_schema

⁷ https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/src/expanded_minerals_domain_schema

⁸ https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/src/auto_generated_schema

⁹ https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/src/schema_less

Additionally, we included **Baseline RAG** for comparison purposes. For this, we utilise a basic search pipeline from GraphRAG Version 2.3.0 [21] in its default configuration with GPT-4o-Mini (temperature: 0.0, chunk size: 300 tokens, chunk overlap: 100 tokens). The pipeline configuration is available¹⁰.

Our code is available at <https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/src>.

4.5 Competency Question Evaluation

Each of the four pipelines, through both local and global search, as well as the basic search pipeline, are asked 15 competency questions defined by MRIWA (see Table 3). Results can be accessed at <https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/results>.

Evaluation of RAG systems focuses on two key aspects: retrieval quality, which refers to the usefulness of retrieved information for answering a query, and generation quality, which measures the ability to generate accurate and relevant responses from the retrieved information [11]. GraphRAG uses our minerals reports and summarises them into a KG. It is not plausible to define a correct set of nodes/summaries to retrieve for a certain query, given that the number of nodes in the generated KG exceeds 195,000 for each of our trials. This precludes the use of many traditional metrics of retrieval quality, including Hit Rate and Mean Reciprocal Rank, which focus on the frequency at which the system retrieves the “correct” piece of information. Studies that have existing SQL or Graph databases are able to define concrete queries or answers for which the LLM’s output can be compared, enabling the use of metrics including Average Overall Execution Accuracy [31]. This technique is also not possible with the size of the GraphRAG-generated database of our technical reports. Our evaluation of generation quality is constrained to the use of our MRIWA-defined ideal answers, which contain key information expected in a completely correct response, as a basis of comparison. Given the technical nature of MRIWA’s reports, however, this process of comparison requires subject matter expertise which further precludes using automated LLM-based evaluation approaches, like Retrieval Augmented Generation Assessment (RAGAS) [8] and Automated RAG Evaluation System (ARES) [29], which use non-expert general-purpose LLMs to conduct the evaluation.

For example, the answer to our simplest question “*identify which MRIWA reports reference MERIWA or MRIWA*” should be a simple list of report identifiers (MRIWA reports 432, 448, etc.). Instead, when using GraphRAG with the auto-generated schema, the response (shown below) is very verbose, containing a subset of the correct report numbers as well as information about the reports that was not requested and in some cases not completely correct.

“The following MRIWA reports reference either MERIWA or MRIWA:

1. ***MRIWA REPORT M532***: This report focuses on the geology, mineralogy, and geometallurgy of electric vehicle (EV) materials deposits in Western Australia. It highlights the significance of lithium production and provides insights into local resources associated with these deposits. The report is a comprehensive study that contributes to understanding the characteristics and potential of EV materials in the region [Data: Reports (12493); Entities (290, 10929, 1445, 23142, +more)].
2. ***MRIWA REPORT NO. 488***: This report details research findings on fluid-rock interactions and lixiviant/oxidant behavior for metal recovery from deep ore bodies. It is part of a broader initiative to understand the processes involved in mineral extraction and recovery [Data: Sources (9809, 9748, 9861); Entities (290, 10929, 1445, 23142, +more)].

¹⁰https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/supplementary_materials/basic_search

3. ****FINAL REPORT TO MRIWA****: *This document encapsulates the findings and analyses from the By-products (LEAF) Project M451, detailing the project’s outcomes, methodologies, and implications regarding soil amendments and their effects on leachate. It serves as a vital resource for understanding the impacts of soil amendments on leachate and the broader environmental context [Data: Reports (11245); Entities (290, 10929, 1445, 23142, +more)]. These reports illustrate the collaborative efforts and research initiatives overseen by MRIWA, contributing to advancements in mineral research and environmental management in Western Australia. The references to MRIWA in these reports underscore its role in funding and overseeing significant research projects in the region."*

Evaluation of long-form answers, like the above, in technical domains is in its infancy, and traditional benchmarking methods that rely on standard train/test datasets have shown limitations in assessing LLM-generated responses [1, 4]. The process still depends, in large part, on evaluation against domain-expert provided model answers.

To evaluate these RAG pipelines on the competency questions we used a human-in-the-loop approach, incorporating subject matter evaluation and reference to the ideal answers.

Our classification approach is informed by evaluation measures for 1) Quality (Context Relevance, Answer Faithfulness, Answer Relevance) and 2) Required Abilities (Noise Robustness, Negative Rejection, Information Integration, and Counterfactual Robustness) [11]. Rather than provide an assessment against each of these seven dimensions, which was very time-consuming and difficult to agree on, we created a five-level classification approach shown in Table 6. We found that the wording enabled evaluators to provide a yes/no answer for each level. While based on the aforementioned dimensions, our classification scheme is coherent with the metrics proposed in RAGEval [43]: Completeness (B1 and B2), Irrelevance (B3 and B4), and Hallucination (B5).

■ **Table 6** Evaluation of GraphRAG pipeline performance.

Code	Name	Criteria
B1	High Answer Relevance	All <i>ideal answer</i> information is present in the response.
B2	Medium Answer Relevance	Some <i>ideal answer</i> information is present in the response.
B3	Medium Noise Robustness	Some factually correct information that is closely related to the question, but that does not directly answer the question, is present in the response.
B4	Low Noise Robustness	Some factually correct information that is irrelevant to the question is present in the response.
B5	Hallucination	A hallucination is present in the response.

Two independent raters classified all 135 responses to the competency questions using the classification scheme. Each question’s ideal answer outlined the specific information required for a response to be considered fully correct. Prior to rating, the raters reviewed each ideal answer and agreed on how the classification scheme would apply to that particular question, ensuring consistent interpretation. For example, the question “Identify which MRIWA reports reference MERIWA or MRIWA” has the following ideal answer and agreed upon evaluation scheme:

Final-Report_448_MRIWA_M0448.txt: MRIWA
Final-Report_278_MERIWA_M0384.txt: MERIWA

Final-Report_289_MERIWA_M0401.txt: MERIWA
 Final-Report_532-MRIWA_M0532.txt: MRIWA
 Final-Report_479_MRIWA_M0479.txt: MRIWA
 Final-Report_459-489_MRIWA_M0459-M0484.txt: MRIWA
 Final-Report_458_MRIWA_M0458.txt: MRIWA
 Final-Report_476_MRIWA_M0476.txt: MRIWA
 Final-Report_432_MRIWA_M0432.txt: MRIWA
 Final-Report_470_MRIWA_M0470.txt: MRIWA
 Final-Report_295_MERIWA_M0416.txt: MERIWA
 Final-Report_494_MRIWA_M0494.txt: MRIWA, MERIWA
 Final-Report_488_MRIWA_M0488.txt: MRIWA
 Final-Report_451_MRIWA_M0451.txt: MRIWA
 Final-Report_505_MRIWA_M0505.txt: MRIWA

B1 – All above reports noted as referencing MRIWA/MERIWA.

B2 – Some of above reports noted as referencing MRIWA/MERIWA.

B3 – Discussions about MRIWA/MERIWA in MRIWA’s reports generally, but not specifying or referring to the specific MRIWA report/s that mention MRIWA/MERIWA.

B4 – Discussions not about or unrelated to MRIWA/MERIWA in MRIWA’s reports (e.g., MRIWA appearing in non-MRIWA reports).

B5 – False statement.

Multiple performance classifications (B1 – B5) may be attributed to a single response. For example, a response may contain “all ideal answer information” (B1) from our MRIWA-defined ideal answer but also include “some factually correct information that is irrelevant to the question” (B4).

5 Results and Discussion

5.1 KG Schema and Local-Global Search Performance

The evaluation results (averaged across the raters) using the five-level classification system for each GraphRAG pipeline (with four KG schema options, and local/global search) and the Baseline RAG pipeline are shown in Table 7. All pipeline responses are available¹¹. As described in Section 4.5, the GraphRAG responses contain multiple paragraphs, relevant and irrelevant answers and facts, and in some cases reference the retrieved information. The average length of responses is 302.40 words (396.97 tokens).

We make the following observations about the schema pipeline data in Table 7 before presenting in-depth discussions on the performance of specific KG schemas in the following subsections. In discussing the performance classifications, we consider a good performance as having a high total for B1 and B2 (relevance) and a low total for B5 (hallucinations). B3 and B4 (robustness) are informative in that a high total indicates that additional peripheral information is being returned that may or may not be relevant to the question.

Relevance (B1 and B2):

- All schemas exhibit similar levels of relevance (B1 and B2).
- 57% of responses exhibit B2 (contain some ideal answer information) indicating majority of responses are partially correct.

¹¹ https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/supplementary_materials/performance_analysis

■ **Table 7** Average total number of responses per pipeline that exhibit each performance classification (B1–B5).

	Local Search					Global Search				
	B1	B2	B3	B4	B5	B1	B2	B3	B4	B5
Minerals Domain Schema	2.0	9.0	11.0	4.0	2.0	0.0	8.5	9.0	8.5	1.5
Expanded Minerals Domain Schema	0.5	9.5	8.0	3.0	2.5	1.0	8.0	7.0	9.5	1.5
Auto-Generated Schema	2.0	9.0	6.5	1.0	3.0	0.5	8.0	5.5	9.5	4.0
Schema-less	1.0	8.0	8.0	3.5	3.5	0.5	8.0	9.5	9.5	2.0
	Basic Search									
	B1	B2	B3	B4	B5	–	–	–	–	–
Baseline RAG	1.0	8.5	6.0	4.0	3.5	–	–	–	–	–

- Only approximately 7 of the 120 responses received a B1 score (all ideal answer information present in the response).
- Local search slightly outperforms global search on relevance (B1 and B2).

Robustness (B3 and B4):

- Global search pipelines produced more irrelevant information (B4) than local search pipelines.
- The MDS pipeline with local search produced fewer examples of peripheral information (B3 and B4) than MDS using global search.
- The MDS pipeline produced the second highest amount of peripheral information (B3 and B4) and produced the most relevant peripheral information (B3).

Hallucination (B5):

- The MDS pipeline produced hallucinations (B5) at a lower rate (though comparable with EMDS) than the other pipelines, only hallucinating on the following four competency questions: Two *Simple* queries “Which MRIWA reports has Commonwealth Scientific Industrial Research Organisation been involved with in any capacity (including being listed in references)?”, “Which MRIWA report is related to the East Kimberley region?”, one *Simple with Condition* query “Which MRIWA reports has Commonwealth Scientific Industrial Research Organisation been involved with as a sponsor?”, and one *Aggregation* query “What is the average number of references in each MRIWA report?”
- All local search variants hallucinated on the *Aggregation* query “What is the average number of references in each MRIWA report?”, while none of the global search pipelines hallucinated.
- The highest number of hallucinations is from the AGS pipeline using global search, hallucinating on four out of the 15 competency questions.

General Observations:

- The MDS and EMDS pipelines using local search had the most desirable set of scores.
- Global search results of all pipelines are not clear to rank in terms of performance.

5.1.1 Comparison of KG schema selection

These results suggest that GraphRAG on MRIWA’s reports performs the best when the MDS or EMDS is used with local search. To elicit possible reasons for this we examined the number of entities in the KG generated by each pipeline as shown in Table 8. We make the following observations.

The MDS pipeline extracts approximately 10% more entities from MRIWA’s reports than the other three pipelines. We hypothesise that with an additional 20,000 entities, the MDS pipeline extracts and retrieves more relevant information from MRIWA’s reports. As a result, its context

■ **Table 8** Statistics of each pipeline’s generated knowledge graph.

	Minerals Do- main Schema	Expanded Minerals Do- main Schema	Auto- Generated Schema	Schema-less
# Entities	218274	202878	198090	195930
# Relations	59087	57580	60602	67791
# Communities	6056	5867	5807	6421

window includes more high-value content, which contributes to the observed improvement in answer quality compared to the AGS and SL pipelines. This also explains the additional relevant peripheral information present in MDS responses compared to EMDS.

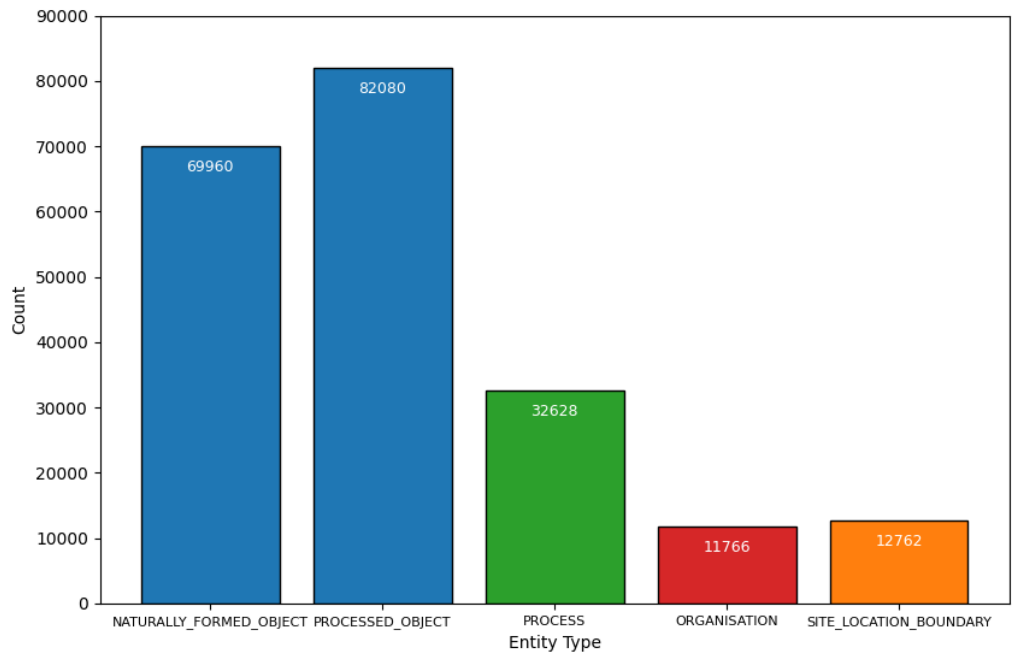
In contrast, pipelines with fewer entities in the KG retrieve less valuable information, leaving more room for irrelevant content in the context window. This increases the risk of distracting the LLM, a phenomenon known as the Distraction Problem [5], and would explain the worse response quality to competency questions.

The five entities in the MDS represent the abstract concepts in the MRIWA texts. The MDS pipeline extracted the highest number of entities from MRIWA’s reports compared to the other pipelines. This observation raises an important question: why does specifying the MDS lead to more entities – but not more relations or communities – being extracted into the KG compared to other schema options? In the MDS pipeline, we only provided GraphRAG with five abstract entity types: `Naturally_Occurring_Object`, `Processed_Object`, `Process`, `Site_Location_Boundary`, and `Organisation`, without any examples to define these terms. We hypothesise that the MDS may be easier for a general-purpose LLM like GPT-4o-Mini to apply because 1) limiting the number of classes reduces complexity for the LLM, and 2) the classes are broad and require less domain-specific expertise to interpret. This may explain why the EMDS schema – which adds domain-specific subclasses to these same five types – resulted in fewer entities being extracted. However, we found no prior literature that explains this phenomenon.

These results suggest there may be an optimal level of abstraction when defining entity types for GraphRAG pipelines. A schema that is too specific may constrain the LLM and reduce recall, while one that is too generic may overlook important distinctions. Striking the right balance appears to be a critical consideration in KG schema design for GraphRAG, and we identify this as an area for future research.

Figures 4 and 5 show that object-related entity types dominate the KGs in both the MDS and EMDS pipelines, which is expected given the nature of MRIWA’s reports. However, we did not observe a clear difference in how well the pipelines answered object-related versus process-related questions, so the impact of this imbalance is unclear.

Another finding is that the LLM occasionally labelled entities with types that were not part of the specified schema. For example, in the MDS pipeline, approximately 3% of the 218,000 extracted entities were assigned types including, but not limited to, `Person`, `Technology`, and `Drillhole`, despite not being among the five MDS-defined entity types. This demonstrates the impact of the non-deterministic nature of LLM outputs, even with temperature set to 0.0, and highlights a limitation of using prompts alone to constrain entity extraction. This limitation may be mitigated by incorporating a post-processing validation step to enforce schema compliance and by tailoring the GraphRAG few-shot examples in the entity extraction prompt to the domain of interest.



■ **Figure 4** Distribution of MDS entity types extracted by the MDS GraphRAG pipeline.

5.1.2 Comparison of Local and Global Search

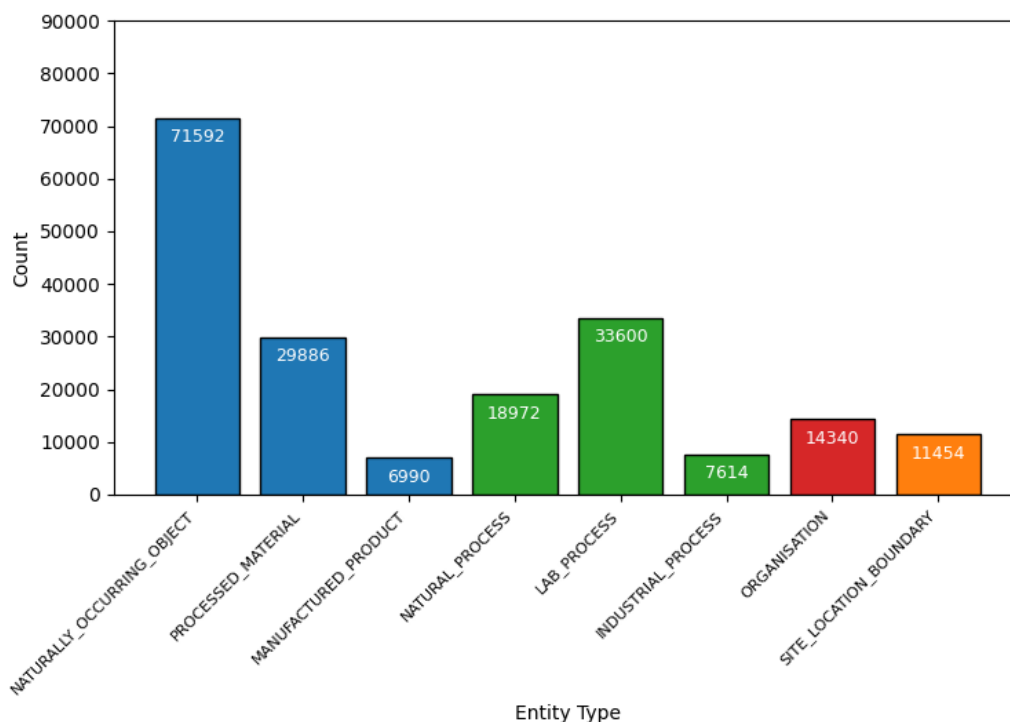
The results in Table 7 suggest global search responses tend to contain more peripheral information (higher B3 and B4 counts) than local search responses.

This result is expected as global search is designed to answer queries at a higher level of abstraction. In doing so, the approach includes a larger breadth of information during the generation procedure (approximately 15 to 20 times the token usage of local search), which may explain the inclusion of additional peripheral information in the final response.

This finding implies that global search is better suited for high-level queries requiring broad information coverage. We observe this to be true for our dataset and competency questions. For example, global search responses to “Which elements are considered in the MRIWA reports?” and “Which regions of Western Australia are referenced in the MRIWA reports?” typically list more elements and regions than local search, and offer more detailed answers. On average, global search responses are 15% longer. However, neither method fully answered these queries, each capturing only a subset of the *ideal answer*.

5.1.3 Comparison of GraphRAG and Baseline RAG

Compared to the GraphRAG pipelines, baseline RAG shows comparable B1 and B2 scores, indicating similar answer relevance. However, the baseline RAG answers are significantly shorter and narrower in scope. On average, GraphRAG responses are 2,273 characters long, while baseline RAG responses average 1,628 characters – a 39.64% difference. This brevity often corresponds to a reduced subset of the ideal answer information. For example, in response to the question “Which regions of Western Australia are referenced in the MRIWA reports?”, baseline RAG returned just two regions, whereas the MDS global pipeline listed eight. Baseline RAG also includes slightly less relevant peripheral information (B3) and slightly more irrelevant information (B4) than the MDS and EMDS local pipelines. Baseline RAG’s B5 score is roughly double that of the MDS and EMDS local pipelines, indicating a higher rate of hallucinations.



■ **Figure 5** Distribution of EMDS entity types extracted by the EMDS GraphRAG pipeline.

5.1.4 Reflection on Performance Classification Scheme

The scores discussed in Table 7 were assigned by two people using the ideal answers provided by the MRIWA experts. The inter-rater agreement was high, with a Cohen’s Kappa of 0.804 (out of 1) and percentage agreement of 91.56%. These values were calculated across all 675 binary ratings, comparing agreement on the presence or absence of each performance criterion. The agreement scores indicate strong consistency in how the classification scheme was applied. The main area of difference was a debate around the distinction between B3 (Medium Noise Robustness) and B4 (Low Noise Robustness). The second marker suggests that B3 (Medium Noise Robustness) and B4 (Low Noise Robustness) could be combined, as having information that does not directly answer the question, whether on-topic or not, is equally unhelpful to the user. However, the MRIWA team argued that providing a wide range of information could help users make new connections between different findings. Ultimately, we keep B3 (Medium Noise Robustness) and B4 (Low Noise Robustness) separate, but we note the value of each performance classification may depend on the user’s goal and the specific question being asked.

We recognise a limitation of the B2 classification: it applies equally to answers containing nearly all ideal facts (e.g., 14/15) and those containing very few (e.g., 1/15). This lack of granularity means that poor responses can still receive a medium relevance score. The impact of this depends on the scale of the dataset and the length of the ideal answer. When ideal answers are short (e.g., 5 key facts), B2 may offer sufficient resolution. However, for more complex questions requiring many facts, the classification becomes less representative of actual answer quality, and an additional rating criterion may be needed to capture this distinction.

5.2 Performance of the MDS pipeline based on type of competency question

In this section we explore the impact of the type of question on the performance of the MDS GraphRAG pipeline by looking at some specific examples in more detail. We chose the MDS pipeline as it has strong performance compared to the other pipelines, as discussed in the preceding sections.

5.2.1 Simple Query: Identification of MRIWA Organisation in Reports

The question “*Identify which MRIWA reports reference MERIWA or MRIWA.*” requires keyword search for the organisation “MRIWA/MERIWA”. A correct response would list all 15 MRIWA report identifiers because each report contains “MRIWA” and/or “MERIWA”.

We compare the MDS local search response with the MDS global search response for this question. With local search, the response correctly identifies three of the 15 reports available. The response also identifies an additional MRIWA report not provided in our sample but that is mentioned in one of our 15 reports. In contrast, global search identifies eight MRIWA reports correctly.

Generally, the ability to partially answer this question with no hallucinations is likely due to “MRIWA” and “MERIWA” being tagged with our MDS entity type **Organisation** during *Indexing*, allowing the related information to be retrieved during the *Query* stage. The global search identifies more correct reports because it incorporates a broader range of information during generation, using approximately 15 to 20 times more tokens than local search.

5.2.2 Set Query: Identification of Chemical Elements

The question, “*Which elements are considered in the MRIWA reports?*” requires the system to collate the set of all chemical elements across documents. A good response might identify, for each report, the elements of focus and the context in which they are mentioned. Elements frequently mentioned in our report subset include, but are not limited to, Gold, Carbon, Copper, Sulfur, Iron, Potassium, Phosphorus and Lithium.

We compare the MDS local search response with the MDS global search response for this question. With local search, the response identifies five chemical elements, which are a small subset of all the elements mentioned in the reports. Two include Rubidium and Manganese, neither of which rank in the top 10 most frequently mentioned elements in the documents. With only four MRIWA reports being explicitly mentioned in the response, the result demonstrates an inability of local search to effectively aggregate information about chemical elements across reports.

With global search, the response identifies seven elements, a small improvement over local search. However, no specific MRIWA reports are mentioned in the response. This limitation indicates a disadvantage of the global search technique, as it only utilises community reports when generating a response. By utilising these high-level summaries of the document repository, the names of the underlying information sources may not be available to the pipeline during final response generation. Additionally, the global search response notes several minerals in MRIWA’s reports, which is irrelevant information.

5.2.3 Aggregation Query: Mathematical Aggregation of Report Citations

The question, “*What is the average number of references in each MRIWA report?*”, requires the system to both retrieve all references and calculate the average quantity of references in each report. None of our pipelines answered this question successfully.

Local search responses often contain an abundance of hallucinations regarding how many references are in each document. This search strategy first identifies relevant entities in the KG – for a question of this nature, it is not obvious what entities would be relevant given the schema. Nonetheless, if relevant entities are identified, then all 300-token chunks of reference lists need to be retrieved and pieced together – this does not occur and explains the failure to answer the question.

Global search responses typically contain statements to the effect of *“I am sorry but I am unable to answer this question given the provided data.”* Given that global search only retrieves community reports of summarised information, full reference lists are not included in the final context window, resulting in the response failure.

5.2.4 Post-Processing-Heavy Query: Identification of Mining Extraction in Reports

The question *“Which MRIWA reports relate to mining extraction?”*, focuses on a specific process, rather than an object or organisation. A good response might identify a set of reports related to mining extraction and describe each report’s contribution to the topic.

We compare the MDS local search response with the MDS global search response for this question. With local search, the response successfully identifies MRIWA Report 488, which is the only report in our subset that relates to mining extraction. However, the response contains four more MRIWA reports, which are not related to mining extraction.

With global search, the response does not successfully identify Report 488, instead identifying four unrelated MRIWA reports. The response also identifies an additional report “Laboratory Flotation Test”, which is a community report, not a MRIWA report. This is a common issue across all global search responses. Since we use GraphRAG’s default LLM prompts, the LLM is not aware of the surrounding context of the system and, therefore, fails to understand the difference between a MRIWA report and a generated community report.

These complex semantic queries, generally, are not answered well. The LLM, being a non-expert in minerals research, often fails to correctly distinguish between concepts such as “mineral processing”, “mining extraction”, and “exploration”. This issue may extend to other technical domains, too. A solution may be to fine-tune the LLM on the text being retrieved to improve response quality.

5.2.5 Identification of Organisational Involvement in Projects

The questions *“Which MRIWA reports has Commonwealth Scientific Industrial Research Organisation been involved with as a researcher?”*, and *“Which MRIWA report author has been involved in more than one report/project?”* both require the system to identify involvement of entities in projects. The key difference is the former question names a specific organisation to be identified, whilst the latter question specifies “authors” to be identified.

We compare the MDS local search response to each question. The response to the CSIRO question successfully identifies three of the six MRIWA reports in which CSIRO is involved in as a researcher. In contrast, the response to the authors question is incorrect. Specifically, the five authors identified are valid authors, but none have been involved in multiple projects. This question is not answered successfully by any of the pipelines tested.

The explanation for this result may lie in the schema, as we specify **Organisation** as an entity type but not **Person** or **Author**. Further, the authors of reports are typically contained within document metadata which GraphRAG does not capture during the *Indexing* stage. This information may assist the LLM in answering metadata-related queries.

5.3 Phrase-by-Phrase Analysis

To investigate the depth of answers, we do a phrase-by-phrase analysis of the responses to the question, “Which MRIWA report is related to the East Kimberley region?”. We classify phrases into five categories, as is displayed in Figure 6. These figures aim to give the reader a visual sense of the complexity of evaluating RAG responses. We want to see answers with cool colours (green and blue) and without hot colours (red and pink).

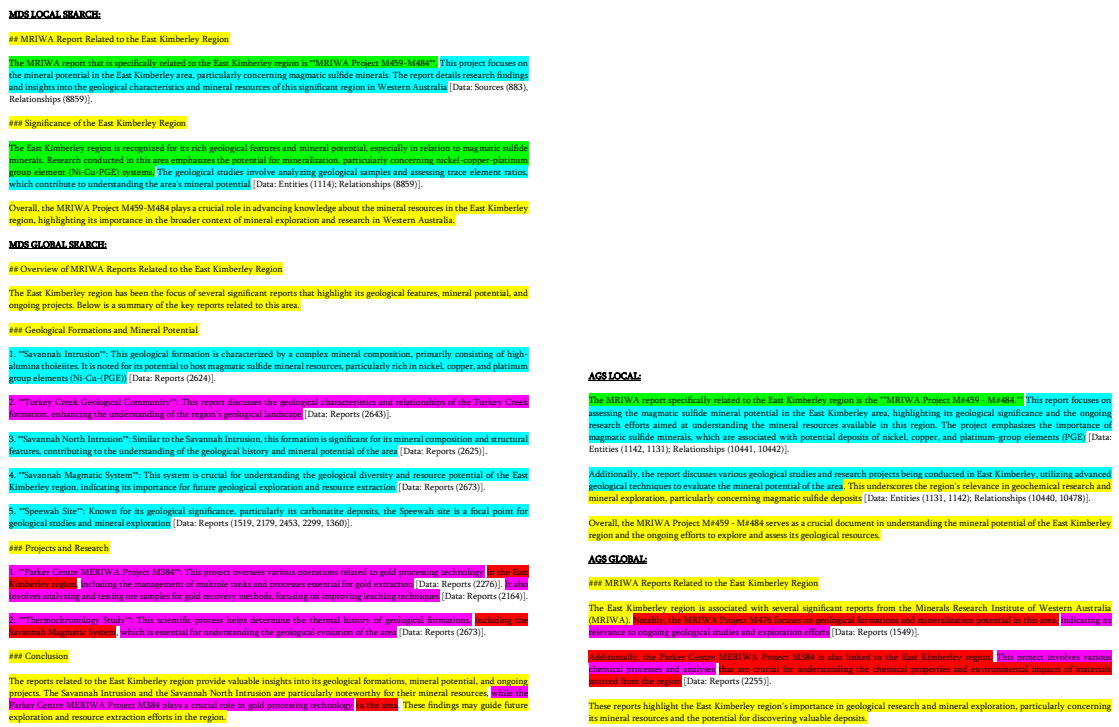


Figure 6 Minerals Domain Schema pipeline responses (left) and Auto-Generated Schema pipeline responses (right) to East Kimberley question. Each schema has two responses, Local Search (upper) and Global Search (lower).

In general, pipelines using a MRIWA-specific schema produce responses with a larger quantity of correct and relevant statements, compared to the auto-generated schema and schema-less pipelines. This finding is consistent with the results of our performance analysis.

The Local Search responses tend to be more concise and contain more correct and relevant statements (blue and green) than global search responses. This result is consistent with the fact that low-level targeted queries, such as “Which MRIWA report is related to the East Kimberley region?” are answered better with Local Search.

Default GraphRAG usually provides an unasked-for final paragraph starting with the words “overall”, “in conclusion”, “in summary”, or “this report highlights”. These paragraphs usually contain no new information. This behaviour may be altered by changing the default GraphRAG LLM prompts and parameters. For example, the `response_type` parameter of the `LocalSearch` and `GlobalSearch` classes allow a description of the response format to be specified – the parameter defaults to “Multiple Paragraphs”.

None of the pipelines identify MRIWA Report 448 as being related to the East Kimberley region. While the region is not a primary focus of the report, mentions of the Kimberley appear in several data tables and references. However, these mentions do not occur near the term “Report 448” (or equivalent). Since the report is initially split into 300-token chunks, “Report 448” and “Kimberley” do not appear in the same chunk. As a result, the LLM fails to associate Report 448 with the Kimberley, leading to its omission in responses. Capturing document structure and metadata may help trace the source report of specific entities.

Among the pipelines that correctly identify MRIWA Report 459-484 as primarily focused on the East Kimberley, all fail to capture key details from the report’s abstract when generating a summary. This result suggests GraphRAG’s summarisation process during *Indexing* may lose important depth typically found in a technical report abstract.

5.4 Limitations

GPT-4o-Mini is a non-deterministic LLM. Therefore, repetitions of the experiment are likely to yield different results [33]. The impact of non-determinism may be prevalent during entity and relation extraction. Given the same KG schema and set of text chunks, the entities and relations identified may not be identical between trials, resulting in variation of the final KG, impacting GraphRAG responses. As Edge et al. [7] note, this issue may be solved by conducting multiple gleanings, where the process of entity extraction is repeated to detect additional entities missed by the LLM.

A risk of using OpenAI’s GPT is its closed-source nature, which can lead to vendor lock-in and raise privacy and confidentiality concerns. Additionally, changes to the underlying model may impact repeatability. While using an open-source LLM like Llama can provide more control, it does not eliminate the possibility of updates affecting results and may introduce performance differences.

Finally, our experiments utilise a subset of 15 MRIWA reports and do not include tables or figures from the reports. These may provide valuable information for the construction of a more comprehensive KG. However, all the MRIWA reports are publicly available, and for the 15 reports we have used, we provide the questions, our ideal answers, schemas, and details on how we performed each evaluation. This will allow others to replicate and extend our work.

6 Conclusions

This paper explores the application of Microsoft’s GraphRAG pipeline to technical reports about geology, chemistry and mineral processing published by MRIWA. The project assesses how 1) domain-specific knowledge graph schema, and 2) the selection of local or global GraphRAG search options, impact the quality of GraphRAG responses to MRIWA-defined queries. Pipeline performance evaluation is based on domain-expert curated answers, and original scoring systems considering relevance, robustness and hallucinations.

GraphRAG with a simple five-class minerals-domain schema (MDS) extracts approximately 10% more entities from MRIWA’s reports compared to pipelines using a complex eight-class minerals-domain schema (EMDS), an auto-generated schema (AGS), and a schema-less (SL)

GraphRAG. While MDS responses contain more relevant peripheral information, both MDS and EMDS produce the most factually correct answers with the fewest hallucinations. These results suggest that a domain-specific schema can assist GraphRAG with entity extraction on technical texts, leading to better response quality.

References

- Bradley P Allen, Lise Stork, and Paul Groth. Knowledge engineering using large language models. *arXiv preprint arXiv:2310.00637*, 2023. doi: 10.48550/arXiv.2310.00637.
- Tyler Bikaun, Michael Stewart, and Wei Liu. Quickgraph: A rapid annotation tool for knowledge graph extraction from technical text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, 2022. doi: 10.18653/v1/2022.acl-demo.27.
- Kurt Bollacker, Robert Cook, and Patrick Tufts. Freebase: A shared database of structured general human knowledge. In *AAAI*, volume 7, pages 1962–1963, 2007. URL: <http://www.aaai.org/Library/AAAI/2007/aaai07-355.php>.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024. doi:10.1145/3641289.
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729, 2024. doi:10.1145/3626772.3657834.
- Armita Davarpanah, Hassan A Babaie, and W Crawford Elliott. Knowledge-based query system for the critical minerals. *Applied Computing and Geosciences*, 22:100167, 2024. doi:10.1016/j.acags.2024.100167.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A Graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. doi:10.48550/arXiv.2404.16130.
- Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of Retrieval Augmented Generation. *arXiv preprint arXiv:2309.15217*, 2023. doi:10.48550/arXiv.2309.15217.
- Mathieu Fenniak and PyPDF2 Contributors. PyPDF2: A pure-Python PDF library, version 3.x. GitHub, 2022. Accessed: 17-Jul-2025. URL: <https://github.com/py-pdf/pypdf>.
- E. Friis. Langchain. GitHub, 2024. Accessed: 09-May-2024. URL: <https://github.com/langchain-ai>.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-Augmented Generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. doi: 10.48550/arXiv.2312.10997.
- Luan Fonseca Garcia, Mara Abel, Michel Perrin, and Renata dos Santos Alvarenga. The geocore ontology: a core ontology for general use in geology. *Computers & Geosciences*, 135:104387, 2020. doi:10.1016/j.cageo.2019.104387.
- Tania Hidalgo, Laura Kuhar, Andreas Beinlich, and Andrew Putnis. Understanding fluid-rock interactions and lixiviant/oxidant behaviour for the in-situ recovery of metals from deep ore bodies. MRIWA Report M0488, Minerals Research Institute of Western Australia, 2020. Report No. 488.
- International Organization for Standardization. ISO/IEC 21838-1:2021 Information technology — Top-level ontologies (TLO) - Part 1: Requirements. International standard, International Organization for Standardization, 2021.
- Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, et al. Think and retrieval: A hypothesis knowledge graph enhanced medical large language models. *arXiv preprint arXiv:2312.15883*, 2023. doi:10.48550/arXiv.2312.15883.
- Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. Knowledge graph-augmented language models for knowledge-grounded dialogue generation. *arXiv preprint arXiv:2305.18846*, 2023. doi:10.48550/arXiv.2305.18846.
- Jerry Liu. Llamaindex. GitHub, 2022. Accessed: 09-May-2024. URL: https://github.com/jerryliu/llama_index.
- Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*, 2023. doi:10.48550/arXiv.2310.01061.
- Microsoft. GraphRAG. <https://microsoft.github.io/graphrag/>. Accessed: 2024-09-30.
- Microsoft. GraphRAG. <https://github.com/microsoft/graphrag/tree/v0.1.1>, 2024. Accessed: 2024-09-30.
- Microsoft. GraphRAG. <https://github.com/microsoft/graphrag/tree/v2.3.0>, 2025. Accessed: 2025-07-25.
- Mineral Research Institute of Western Australia. Project Portfolio. <https://www.mriwa.wa.gov.au/research-projects/project-portfolio/>, 2024. [Online; accessed 9-May-2024].
- Pin Ni, Ramin Okhrati, Steven Guan, and Victor Chang. Knowledge graph and deep learning-based text-to-graphql model for intelligent medical consultation chatbot. *Information Systems*

- Frontiers*, 26(1):137–156, 2024. doi:10.1007/s10796-022-10295-0.
- 24 Daniela Oliveira, Ratnesh Sahay, and Mathieu d'Aquin. Leveraging ontologies for knowledge graph schemas. In *KGB@ ESWC*, pages 24–36, 2019. URL: <https://ceur-ws.org/Vol-2489/paper3.pdf>.
 - 25 Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36:3580–3599, 2024. doi:10.1109/tkde.2024.3352100.
 - 26 Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph Retrieval-Augmented Generation: A survey. *arXiv preprint arXiv:2408.08921*, 2024. doi:10.48550/arXiv.2408.08921.
 - 27 Tyler Thomas Procko and Omar Ochoa. Graph Retrieval-Augmented Generation for large language models: A survey. In *2024 Conference on AI, Science, Engineering, and Technology (AIx-SET)*, pages 166–169. IEEE, 2024.
 - 28 David Robinson. Improving solvent extraction technology. MRIWA Report M0401, Minerals Research Institute of Western Australia, 2010. Report No. 289.
 - 29 Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for Retrieval-Augmented Generation systems. *arXiv preprint arXiv:2311.09476*, 2023. doi:10.48550/arXiv.2311.09476.
 - 30 Henri Scaffidi, Melinda Hodkiewicz, Caitlin Woods, and Nicole Rooke. nlp-tlp/GraphRAG-on-Minerals-Domain. InteractiveResource (visited on 2025-10-07). URL: <https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain>, doi:10.4230/artifacts.24886.
 - 31 Juan Sequeda, Dean Allemang, and Bryon Jacob. A benchmark to understand the role of knowledge graphs on large language model's accuracy for question answering on enterprise SQL databases. In *Proceedings of the 7th Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, pages 1–12, 2024. doi:10.1145/3661304.3661901.
 - 32 Tanya Serry, Tonya Stebbins, Andrew Martchenko, Natalie Araujo, and Brigid McCarthy. Improving access to COVID-19 information by ensuring the readability of government websites. *Health Promotion Journal of Australia*, 34(2):595–602, 2023. doi:10.1002/hpja.610.
 - 33 Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*, 2024. doi:10.48550/arXiv.2407.10457.
 - 34 Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019. doi:10.1038/s41598-019-41695-z.
 - 35 Denny Vrandečić. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st international conference on World Wide Web*, pages 1063–1064, 2012. doi:10.1145/2187980.2188242.
 - 36 Chengbin Wang, Xiaogang Ma, Jianguo Chen, and Jingwen Chen. Information extraction and knowledge graph construction from geoscience literature. *Computers & Geosciences*, 112:112–120, 2018. doi:10.1016/j.cageo.2017.12.007.
 - 37 Xintao Wang, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao, and Wei Wang. Knowledgept: Enhancing large language models with retrieval and storage access on knowledge bases. *arXiv preprint arXiv:2308.11761*, 2023. doi:10.48550/arXiv.2308.11761.
 - 38 M. Wells, M. Aylmore, and B. McInnes. The geology, mineralogy and geometallurgy of EV materials deposits in Western Australia. MRIWA Report M532, Geological Survey of Western Australia, 2022. Report No. 228.
 - 39 Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. Retrieval-Augmented Generation with knowledge graphs for customer service question answering. *arXiv preprint arXiv:2404.17723*, 2024. doi:10.1145/3626772.3661370.
 - 40 Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Gui, Ziran Jiang, Ziyu Jiang, et al. CRAG-comprehensive RAG benchmark. *Advances in Neural Information Processing Systems*, 37:10470–10490, 2024. doi:10.48550/arXiv.2406.04744.
 - 41 Huaxuan Zhao, Yueling Pan, and Feng Yang. Research on information extraction of technical documents and construction of domain knowledge graph. *IEEE Access*, 8:168087–168098, 2020. doi:10.1109/access.2020.3024070.
 - 42 Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. A comprehensive survey on automatic knowledge graph construction. *ACM Computing Surveys*, 56(4):1–62, 2023. doi:10.1145/3618295.
 - 43 Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, et al. Rageval: Scenario specific RAG evaluation dataset generation framework. *arXiv preprint arXiv:2408.01262*, 2024. doi:10.48550/arXiv.2408.01262.