

LIGHTRAG: SIMPLE AND FAST RETRIEVAL-AUGMENTED GENERATION

LIGHTRAG: 简单且快速的检索增强生成

Zirui Guo ^{1,2}, Lianghao Xia ², Yanhua Yu ^{1,*}, Tu Ao ¹, Chao Huang ^{2*}

郭子睿 ^{1,2}, 夏亮豪 ², 余艳华 ^{1,*}, 段涛 ¹, 黄超 ^{2*}

Beijing University of Posts and Telecommunications ¹

北京邮电大学 ¹

University of Hong Kong ²

香港大学 ²

zrguo101@hku.hk aka_xia@foxmail.com chaohuang75@gmail.com

zrguo101@hku.hk aka_xia@foxmail.com chaohuang75@gmail.com

ABSTRACT

摘要

Retrieval-Augmented Generation (RAG) systems enhance large language models (LLMs) by integrating external knowledge sources, enabling more accurate and contextually relevant responses tailored to user needs. However, existing RAG systems have significant limitations, including reliance on flat data representations and inadequate contextual awareness, which can lead to fragmented answers that fail to capture complex inter-dependencies. To address these challenges, we propose LightRAG, which incorporates graph structures into text indexing and retrieval processes. This innovative framework employs a dual-level retrieval system that enhances comprehensive information retrieval from both low-level and high-level knowledge discovery. Additionally, the integration of graph structures with vector representations facilitates efficient retrieval of related entities and their relationships, significantly improving response times while maintaining contextual relevance. This capability is further enhanced by an incremental update algorithm that ensures the timely integration of new data, allowing the system to remain effective and responsive in rapidly changing data environments. Extensive experimental validation demonstrates considerable improvements in retrieval accuracy and efficiency compared to existing approaches. We have made our LightRAG open-source and available at the link: <https://github.com/HKUDS/LightRAG>

检索增强生成 (RAG) 系统通过整合外部知识源增强大型语言模型 (LLMs), 使其能够生成更准确、更具上下文相关性的响应以满足用户需求。然而, 现有 RAG 系统存在显著局限, 包括依赖扁平数据表示和缺乏足够的上下文感知, 可能导致回答碎片化, 无法捕捉复杂的相互依赖关系。为了解决这些问题, 我们提出 LightRAG, 将图结构引入文本索引与检索流程。该创新框架采用双层检索系统, 增强了从低层次和高层次知识发现中获取全面信息的能力。此外, 将图结构与向量表示相结合便于高效检索相关实体及其关系, 显著提升响应速度同时保持上下文相关性。通过增量更新算法进一步提升了在不断变化的数据环境中及时整合新数据的能力, 使系统保持有效且响应迅速。大量实验证明与现有方法相比在检索准确性和效率上有显著提升。我们已将 LightRAG 开源, 地址为:<https://github.com/HKUDS/LightRAG>

1 INTRODUCTION

1 引言

Retrieval-Augmented Generation (RAG) systems have been developed to enhance large language models (LLMs) by integrating external knowledge sources Sudhi et al. (2024); Es et al. (2024); Salemi & Zamani (2024). This innovative integration allows LLMs to generate more accurate and contextually relevant responses, significantly improving their utility in real-world applications. By adapting to specific domain knowledge Tu et al. (2024), RAG systems ensure that the information provided is not only pertinent but also tailored to the user's needs. Furthermore, they offer access to up-to-date information Zhao et al. (2024), which is crucial in rapidly evolving fields. Chunking plays a vital role in facilitating the retrieval-augmented generation process Lyu et al. (2024). By breaking down a large external text corpus into smaller, more manageable segments, chunking significantly enhances the accuracy of information retrieval. This approach allows for more targeted similarity searches, ensuring that the retrieved content is directly relevant to user queries.

检索增强生成 (RAG) 系统旨在通过整合外部知识源来增强大型语言模型 (LLMs) Sudhi et al. (2024); Es et al. (2024); Salemi & Zamani (2024)。这种创新整合使 LLM 能生成更准确且上下文相关的响应, 大幅提升其在实际应用中的效用。通过适应特定领域知识 Tu et al. (2024), RAG 系统确保所提供信息既相关又贴合用户需求。此外, 它们还可获取最新信息 Zhao et al. (2024), 这在快速演进的领域至关重要。分块在促进检索增强生成过程中发挥关键作用 Lyu et al. (2024)。将大型外部文本语料拆分为更小、更易处理的片段, 可显著提高信息检索的准确性。这一方法允许更有针对性的相似性搜索, 确保检索到的内容与用户查询直接相关。

However, existing RAG systems have key limitations that hinder their performance. First, many methods rely on flat data representations, restricting their ability to understand and retrieve information based on intricate relationships between entities. Second, these systems often lack the contextual awareness needed to maintain coherence across various entities and their interrelations, resulting in responses that may not fully address user queries. For example, consider a user asking, "How does the rise of electric vehicles influence urban air quality and public transportation infrastructure?" Existing RAG methods might retrieve separate documents on electric vehicles, air pollution, and public transportation challenges but struggle to synthesize this information into a cohesive response. They may fail to explain how the adoption of electric vehicles can improve air quality, which in turn could affect public transportation planning. As a result, the user may receive a fragmented answer that does not adequately capture the complex inter-dependencies among these topics.

然而，现有 RAG 系统存在限制，影响其性能。首先，许多方法依赖扁平数据表示，限制了基于实体间复杂关系的理解与检索能力。其次，这些系统常缺乏维持各实体及其相互关系一致性的上下文感知，导致回答可能无法充分回应用户查询。例如，若用户询问：“电动汽车的兴起如何影响城市空气质量和公共交通基础设施？”现有 RAG 方法可能检索到关于电动汽车、空气污染和公共交通挑战的独立文档，但难以将这些信息综合为连贯的回答。它们可能无法解释电动汽车的采用如何改善空气质量，从而如何影响公共交通规划。因此，用户可能得到无法充分反映这些主题间复杂相互依赖的零散答案。

*Both Chao Huang and Yanhua Yu are corresponding authors.

* 黄超和余艳华为通讯作者。

To address these limitations, we propose incorporating graph structures into text indexing and relevant information retrieval. Graphs are particularly effective at representing the interdependencies among different entities Rampásek et al. (2022), which enables a more nuanced understanding of relationships. The integration of graph-based knowledge structures facilitates the synthesis of information from multiple sources into coherent and contextually rich responses. Despite these advantages, developing a fast and scalable graph-empowered RAG system that efficiently handles varying query volumes is crucial. In this work, we achieve an effective and efficient RAG system by addressing three key challenges: i) Comprehensive Information Retrieval. Ensuring comprehensive information retrieval that captures the full context of inter-dependent entities from all documents; ii) Enhanced Retrieval Efficiency. Improving retrieval efficiency over the graph-based knowledge structures to significantly reduce response times; iii) Rapid Adaptation to New Data. Enabling quick adaptation to new data updates, ensuring the system remains relevant in dynamic environments.

为应对这些限制，我们提出将图结构引入文本索引与相关信息检索。图在表示不同实体间相互依赖方面尤为有效 Rampásek et al. (2022)，从而实现关系的更细致理解。基于图的知识结构集成有助于来自多源的信息合成为连贯且上下文丰富的回答。尽管如此，开发一个快速且可扩展的图赋能 RAG 系统以高效应对不同查询量仍然至关重要。在本工作中，我们通过解决三大关键挑战实现了一个高效的 RAG 系统：i) 全面信息检索。确保从所有文档中捕获相互依赖实体的完整上下文；ii) 提升检索效率。在图式知识结构上改进检索效率以显著缩短响应时间；iii) 快速适应新数据。实现对新数据更新的快速适配，确保系统在动态环境中保持相关性。

In response to the outlined challenges, we propose LightRAG, a model that seamlessly integrates a graph-based text indexing paradigm with a dual-level retrieval framework. This innovative approach enhances the system’s capacity to capture complex inter-dependencies among entities, resulting in more coherent and contextually rich responses. LightRAG employs efficient dual-level retrieval strategies: low-level retrieval, which focuses on precise information about specific entities and their relationships, and high-level retrieval, which encompasses broader topics and themes. By combining both detailed and conceptual retrieval, LightRAG effectively accommodates a diverse range of queries, ensuring that users receive relevant and comprehensive responses tailored to their specific needs. Additionally, by integrating graph structures with vector representations, our framework facilitates efficient retrieval of related entities and relations while enhancing the comprehensiveness of results through relevant structural information from the constructed knowledge graph.

针对上述挑战，我们提出 LightRAG，一种将基于图的文本索引范式与双层检索框架无缝结合的模式。该创新方法增强了系统捕捉实体间复杂相互依赖的能力，从而生成更连贯且语境丰富的响应。LightRAG 采用高效的双层检索策略：低层检索聚焦于特定实体及其关系的精确信息，高层检索涵盖更广泛的主题与主题脉络。通过结合细粒度与概念性检索，LightRAG 有效适配多样化查询，确保用户获得与其需求相匹配的相关且全面的响应。此外，通过将图结构与向量表示相结合，我们的框架在便捷检索相关实体与关系的同时，借助构建的知识图谱中的结构信息提升结果的完整性。

In summary, the key contributions of this work are highlighted as follows:

总之，本工作的主要贡献如下：

- General Aspect. We emphasize the importance of developing a graph-empowered RAG system to overcome the limitations of existing methods. By integrating graph structures into text indexing, we can effectively represent complex interdependencies among entities, fostering a nuanced understanding of relationships and enabling coherent, contextually rich responses.

- 一般性观点。我们强调构建图增强 RAG 系统以克服现有方法局限的重要性。通过将图结构融入文本索引，可以有效表示实体间的复杂相互依赖，促成对关系的细致理解，并生成连贯且语境充实的回答。

- Methodologies. To enable an efficient and adaptive RAG system, we propose LightRAG, which integrates a dual-level retrieval paradigm with graph-enhanced text indexing. This approach captures both low-level and high-level information for comprehensive, cost-effective retrieval. By eliminating the need to rebuild the entire index, LightRAG reduces computational costs and accelerates adaptation, while its incremental update algorithm ensures timely integration of new data, maintaining effectiveness in dynamic environments.

- 方法论。为实现高效且自适应的 RAG 系统，我们提出 LightRAG，将双层检索范式与图增强文本索引相结合。该方法捕捉低层与高层信息以实现全面且成本效益高的检索。通过消除重建整个索引的需求，LightRAG 降低了计算成本并加速了适配，其增量更新算法确保新数据的及时融合，从而在动态环境中维持有效性。

- Experimental Findings. Extensive experiments were conducted to evaluate the effectiveness of LightRAG in comparison to existing RAG models. These assessments focused on several key dimensions, including retrieval accuracy, model ablation, response efficiency, and adaptability to new information. The results demonstrated significant improvements over baseline methods.

- 实验结果。我们进行了大量实验以评估 LightRAG 相较现有 RAG 模型的有效性。评估聚焦若干关键维度，包括检索准确性、模型消融、响应效率及对新信息的适应性。结果显示相较基线方法有显著提升。

2 RETRIEVAL-AUGMENTED GENERATION

2 检索增强生成

Retrieval-Augmented Generation (RAG) integrates user queries with a collection of pertinent documents sourced from an external knowledge database, incorporating two essential elements: the Retrieval Component and the Generation Component. 1) The retrieval component is responsible for fetching relevant documents or information from the external knowledge database. It identifies and retrieves the most pertinent data based on the input query. 2) After the retrieval process, the generation component takes the retrieved information and generates coherent, contextually relevant responses. It leverages the capabilities of the language model to produce meaningful outputs. Formally, this RAG framework, denoted as \mathcal{M} , can be defined as follows:

检索增强生成 (RAG) 将用户查询与来自外部知识库的一组相关文档整合, 包含两个要素: 检索组件与生成组件。1) 检索组件负责从外部知识库获取相关文档或信息, 基于输入查询识别并检索最相关的数据。2) 检索完成后, 生成组件利用检索到的信息生成连贯且语境相关的响应, 借助语言模型的能力产生有意义的输出。形式上, 该 RAG 框架记为 \mathcal{M} , 可定义如下:

$$\mathcal{M} = (\mathcal{G}, \mathcal{R} = (\varphi, \psi)), \mathcal{M}(q; \mathcal{D}) = \mathcal{G}(q, \psi(q; \hat{\mathcal{D}})), \hat{\mathcal{D}} = \varphi(\mathcal{D}) \quad (1)$$

In this framework, \mathcal{G} and \mathcal{R} represent the generation module and the retrieval module, respectively, while q denotes the input query and \mathcal{D} refers to the external database. The retrieval module \mathcal{R} includes two key functionalities: i) Data Indexer $\varphi(\cdot)$: which involves building a specific data structure $\hat{\mathcal{D}}$ based on the external database \mathcal{D} . ii) Data Retriever $\psi(\cdot)$: The relevant documents are obtained by comparing the query against the indexed data, also denoted as "relevant documents". By leveraging the information retrieved through $\psi(\cdot)$ along with the initial query q , the generative model $\mathcal{G}(\cdot)$ efficiently produces high-quality, contextually relevant responses.

在此框架中, \mathcal{G} 与 \mathcal{R} 分别表示生成模块与检索模块, q 表示输入查询, \mathcal{D} 指外部数据库。检索模块 \mathcal{R} 包含两项关键功能: i) 数据索引器 $\varphi(\cdot)$: 基于外部数据库 \mathcal{D} 构建特定数据结构 $\hat{\mathcal{D}}$ 。ii) 数据检索器 $\psi(\cdot)$: 通过将查询与已索引数据对比以获得相关文档, 亦称“相关文档”。通过利用由 $\psi(\cdot)$ 检索到的信息与初始查询 q , 生成模型 $\mathcal{G}(\cdot)$ 高效地产生高质量、语境相关的响应。

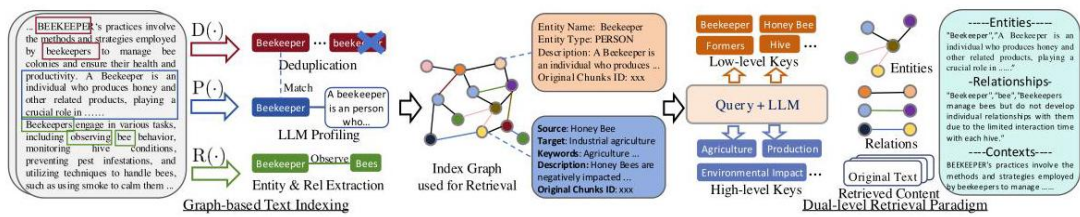


Figure 1: Overall architecture of the proposed LightRAG framework.

图 1: 所提 LightRAG 框架的总体架构。

In this work, we target several key points essential for an efficient and effective Retrieval-Augmented Generation (RAG) system which are elaborated below:

在本工作中, 我们针对若干构建高效且有效的检索增强生成 (RAG) 系统的要点进行了探讨, 具体如下:

- **Comprehensive Information Retrieval:** The indexing function $\varphi(\cdot)$ must be adept at extracting global information, as this is crucial for enhancing the model's ability to answer queries effectively.

- 全面信息检索: 索引功能 $\varphi(\cdot)$ 必须善于提取全局信息, 因为这对于增强模型回答查询的能力至关重要。

- **Efficient and Low-Cost Retrieval:** The indexed data structure $\hat{\mathcal{D}}$ must enable rapid and cost-efficient retrieval to effectively handle a high volume of queries.

- 高效低成本检索: 索引的数据结构 $\hat{\mathcal{D}}$ 必须支持快速且低成本的检索, 以有效应对大量查询。

- **Fast Adaptation to Data Changes:** The ability to swiftly and efficiently adjust the data structure to incorporate new information from the external knowledge base, is crucial for ensuring that the system remains current and relevant in an ever-changing information landscape.

- 快速适应数据变化: 能够迅速且高效地调整数据结构以纳入来自外部知识库的新信息, 对于确保系统在不断变化的信息环境中保持最新和相关至关重要。

3 THE LIGHTRAG ARCHITECTURE

3 LIGHTRAG 架构

3.1 GRAPH-BASED TEXT INDEXING

3.1 基于图的文本索引

Graph-Enhanced Entity and Relationship Extraction. Our LightRAG enhances the retrieval system by segmenting documents into smaller, more manageable pieces. This strategy allows for quick identification and access to relevant information without analyzing entire documents. Next, we leverage LLMs to identify and extract various entities (e.g., names, dates, locations, and events) along with the relationships between them. The information collected through this process will be used to create a comprehensive knowledge graph that highlights the connections and insights across the entire collection of documents. We formally represent this graph generation module as follows:

图增强的实体与关系抽取。我们的 LightRAG 通过将文档切分为更小、更易处理的片段来增强检索系统。这一策略使得可以快速定位并访问相关信息, 而无需分析完整文档。接着, 我们利用大模型识别并抽取各类实体(如姓名、日期、地点和事件)及其相互关系。通过此过程收集的信息将用于构建全面的知识图谱, 突显整个文档集合中的连接与洞见。我们将该图生成模块形式化表示如下:

$$\hat{\mathcal{D}} = (\hat{\mathcal{V}}, \hat{\mathcal{E}}) = \text{Dedupe} \circ \text{Prof}(\mathcal{V}, \mathcal{E}), \mathcal{V}, \mathcal{E} = \cup_{\mathcal{D}_i \in \mathcal{D}} \text{Recog}(\mathcal{D}_i) \quad (2)$$

where $\hat{\mathcal{D}}$ represents the resulting knowledge graphs. To generate this data, we apply three main processing steps to the raw text documents \mathcal{D}_i . These steps utilize a LLM for text analysis and processing. Details

about the prompt templates and specific settings for this part can be found in Appendix 7.3.2 The functions used in our graph-based text indexing paradigm are described as:

其中 $\hat{\mathcal{D}}$ 表示生成的知识图谱。为生成此数据，我们对原始文本文档 \mathcal{D}_i 应用三大处理步骤。这些步骤使用大模型进行文本分析与处理。关于提示模板和该部分具体设置的详情见附录 7.3.2。我们图基文本索引范式中使用的函数描述如下：

- Extracting Entities and Relationships. $R(\cdot)$: This function prompts a LLM to identify entities (nodes) and their relationships (edges) within the text data. For instance, it can extract entities like "Cardiologists" and "Heart Disease," and relationships such as "Cardiologists diagnose Heart Disease" from the text: "Cardiologists assess symptoms to identify potential heart issues." To improve efficiency, the raw text \mathcal{D} is segmented into multiple chunks \mathcal{D}_i .

- 抽取实体与关系。 $R(\cdot)$: 该函数提示大模型识别文本数据中的实体 (节点) 及其关系 (边)。例如，它可以从文本“心脏科医生评估症状以识别潜在心脏问题。”中抽取出“心脏科医生”和“心脏病”等实体，以及“心脏科医生诊断心脏病”之类的关系。为提高效率，原始文本 \mathcal{D} 被切分为多个片段 \mathcal{D}_i 。

- LLM Profiling for Key-Value Pair Generation. $P(\cdot)$: We employ a LLM-empowered profiling function, $P(\cdot)$, to generate a text key-value pair (K, V) for each entity node in \mathcal{V} and relation edge in \mathcal{E} . Each index key is a word or short phrase that enables efficient retrieval, while the corresponding value is a text paragraph summarizing relevant snippets from external data to aid in text generation. Entities use their names as the sole index key, whereas relations may have multiple index keys derived from LLM enhancements that include global themes from connected entities.

- 用于键值对生成的大模型画像。 $P(\cdot)$: 我们采用大模型支持的画像函数 $P(\cdot)$ ，为 \mathcal{V} 中的每个实体节点和 \mathcal{E} 中的关系边生成文本键值对 (K, V) 。每个索引键为单词或短语，以便高效检索，对应的值为总结自外部数据相关片段的文本段落以辅助文本生成。实体使用其名称作为唯一索引键，而关系可能具有多个由大模型增强生成的索引键，这些索引键包含来自连接实体的全局主题。

- Deduplication to Optimize Graph Operations. $D(\cdot)$: Finally, we implement a deduplication function, $D(\cdot)$, that identifies and merges identical entities and relations from different segments of the raw text \mathcal{D}_i . This process effectively reduces the overhead associated with graph operations on $\hat{\mathcal{D}}$ by minimizing the graph's size, leading to more efficient data processing.

- 去重以优化图操作。 $D(\cdot)$: 最后，我们实现了去重函数 $D(\cdot)$ ，该函数识别并合并来自原始文本不同片段 \mathcal{D}_i 的相同实体和关系。此过程通过最小化图的规模，有效降低对 $\hat{\mathcal{D}}$ 上图操作的开销，从而提升数据处理效率。

Our LightRAG offers two advantages through its graph-based text indexing paradigm. First, Comprehensive Information Understanding. The constructed graph structures enable the extraction of global information from multi-hop subgraphs, greatly enhancing LightRAG's ability to handle complex queries that span multiple document chunks. Second, Enhanced Retrieval Performance. the key-value data structures derived from the graph are optimized for rapid and precise retrieval. This provides a superior alternative to less accurate

embedding matching methods (Gao et al., 2023) and inefficient chunk traversal techniques (Edge et al. 2024) commonly used in existing approaches.

我们的 LightRAG 通过其基于图的文本索引范式提供两大优势。其一，全面信息理解。构建的图结构能够从多跳子图中提取全局信息，极大提升 LightRAG 处理跨多个文档片段的复杂查询的能力。其二，增强的检索性能。由图生成的键值数据结构针对快速且精确的检索进行了优化，优于现有方法中常用的较不精确的嵌入匹配方法 (Gao et al., 2023) 和低效的片段遍历技术 (Edge et al. 2024)。

Fast Adaptation to Incremental Knowledge Base. To efficiently adapt to evolving data changes while ensuring accurate and relevant responses, our LightRAG incrementally updates the knowledge base without the need for complete reprocessing of the entire external database. For a new document \mathcal{D}' , the incremental update algorithm processes it using the same graph-based indexing steps φ as before, resulting in $\hat{\mathcal{D}}' = (\hat{\mathcal{V}}', \hat{\mathcal{E}}')$. Subsequently, LightRAG combines the new graph data with the original by taking the union of the node sets $\hat{\mathcal{V}}$ and $\hat{\mathcal{V}}'$, as well as the edge sets $\hat{\mathcal{E}}$ and $\hat{\mathcal{E}}'$.

快速适应增量知识库。为在确保准确且相关响应的同时高效适应不断变化的数据，我们的 LightRAG 对知识库进行增量更新，无需对整个外部数据库进行完全重新处理。对于新文档 \mathcal{D}' ，增量更新算法使用与之前相同的基于图的索引步骤 φ 处理它，生成 $\hat{\mathcal{D}}' = (\hat{\mathcal{V}}', \hat{\mathcal{E}}')$ 。随后，LightRAG 通过对节点集 $\hat{\mathcal{V}}$ 与 $\hat{\mathcal{V}}'$ 以及边集 $\hat{\mathcal{E}}$ 与 $\hat{\mathcal{E}}'$ 取并集，将新图数据与原图合并。

Two key objectives guide our approach to fast adaptation for the incremental knowledge base: Seamless Integration of New Data. By applying a consistent methodology to new information, the incremental update module allows the LightRAG to integrate new external databases without disrupting the existing graph structure. This approach preserves the integrity of established connections, ensuring that historical data remains accessible while enriching the graph without conflicts or redundancies. **Reducing Computational Overhead.** By eliminating the need to rebuild the entire index graph, this method reduces computational overhead and facilitates the rapid assimilation of new data. Consequently, LightRAG maintains system accuracy, provides current information, and conserves resources, ensuring users receive timely updates and enhancing the overall RAG effectiveness.

两个关键目标指导我们对增量知识库的快速适应方法：无缝集成新数据。通过对新信息应用一致的方法，增量更新模块使 LightRAG 能在不破坏现有图结构的情况下整合新的外部数据库。这种方法保留已建立连接的完整性，确保历史数据可访问，同时在不产生冲突或冗余的前提下丰富图谱。降低计算开销。通过消除重建整个索引图的需求，该方法降低了计算开销并便于快速同化新数据。因此，LightRAG 保持系统准确性，提供最新信息并节约资源，确保用户及时获得更新，提升整体 RAG 效果。

3.2 DUAL-LEVEL RETRIEVAL PARADIGM

3.2 双层检索范式

To retrieve relevant information from both specific document chunks and their complex interdependencies, our LightRAG proposes generating query keys at both detailed and abstract levels.

为了从具体文档片段及其复杂相互依赖中检索相关信息，我们的 LightRAG 提出在细化与抽象两个层面生成查询键。

- **Specific Queries.** These queries are detail-oriented and typically reference specific entities within the graph, requiring precise retrieval of information associated with particular nodes or edges. For example, a specific query might be, "Who wrote 'Pride and Prejudice'?"

- **具体查询。**这类查询面向细节，通常指向图中具体实体，需精确检索与特定节点或边相关的信息。例如，一个具体查询可能是，“谁写了《傲慢与偏见》？”

- **Abstract Queries.** In contrast, abstract queries are more conceptual, encompassing broader topics, summaries, or overarching themes that are not directly tied to specific entities. An example of an abstract query is, "How does artificial intelligence influence modern education?"

- **抽象查询。**相反，抽象查询更具概念性，涵盖更广泛的话题、摘要或总体主题，不直接绑定于具体实体。一个抽象查询的例子是，“人工智能如何影响现代教育？”

To accommodate diverse query types, the LightRAG employs two distinct retrieval strategies within the dual-level retrieval paradigm. This ensures that both specific and abstract inquiries are addressed effectively, allowing the system to deliver relevant responses tailored to user needs.

为适应多样的查询类型，LightRAG 在双层检索范式中采用两种不同的检索策略。这确保具体与抽象查询都能得到有效响应，使系统能够根据用户需求提供相关答案。

- **Low-Level Retrieval.** This level is primarily focused on retrieving specific entities along with their associated attributes or relationships. Queries at this level are detail-oriented and aim to extract precise information about particular nodes or edges within the graph.

- **低层检索。**该层主要侧重检索具体实体及其关联属性或关系。此层的查询注重细节，旨在提取图中某些节点或边的精确信息。

- **High-Level Retrieval.** This level addresses broader topics and overarching themes. Queries at this level aggregate information across multiple related entities and relationships, providing insights into higher-level concepts and summaries rather than specific details.

- **高层检索。**该层处理更广泛的话题和总体主题。此层的查询在多个相关实体和关系间聚合信息，提供关于高层概念和摘要的见解，而非具体细节。

Integrating Graph and Vectors for Efficient Retrieval. By combining graph structures with vector representations, the model gains a deeper insight into the interrelationships among entities. This synergy enables the retrieval algorithm to effectively utilize both local and global keywords, streamlining the search process and improving the relevance of results.

整合图与向量以实现高效检索。通过将图结构与向量表示结合，模型能够更深入地理解实体间的相互关系。这种协同使检索算法能够有效利用局部与全局关键词，简化搜索过程并提高结果相关性。

- (i) Query Keyword Extraction. For a given query q , the retrieval algorithm of LightRAG begins by extracting both local query keywords $k^{(l)}$ and global query keywords $k^{(g)}$.

- (i) 查询关键词提取。对于给定查询 q ，LightRAG 的检索算法首先提取局部查询关键词 $k^{(l)}$ 和全局查询关键词 $k^{(g)}$ 。

- (ii) Keyword Matching. The algorithm uses an efficient vector database to match local query keywords with candidate entities and global query keywords with relations linked to global keys.

- (ii) 关键词匹配。该算法使用高效的向量数据库将局部查询关键词与候选实体匹配，将全局查询关键词与与全局键关联的关系匹配。

- (iii) Incorporating High-Order Relatedness. To enhance the query with higher-order relatedness, LightRAG further gathers neighboring nodes within the local subgraphs of the retrieved graph elements. This process involves the set $\{v_i \mid v_i \in \mathcal{V} \wedge (v_i \in \mathcal{N}_v \vee v_i \in \mathcal{N}_e)\}$, where \mathcal{N}_v and \mathcal{N}_e represent the one-hop neighboring nodes of the retrieved nodes v and edges e , respectively.

- (iii) 纳入高阶相关性。为了用更高阶的相关性增强查询，LightRAG 进一步收集被检索图元素的局部子图内的邻近节点。该过程涉及集合 $\{v_i \mid v_i \in \mathcal{V} \wedge (v_i \in \mathcal{N}_v \vee v_i \in \mathcal{N}_e)\}$ ，其中 \mathcal{N}_v 和 \mathcal{N}_e 分别表示被检索节点 v 和边 e 的一跳邻居节点。

This dual-level retrieval paradigm not only facilitates efficient retrieval of related entities and relations through keyword matching, but also enhances the comprehensiveness of results by integrating relevant structural information from the constructed knowledge graph.

该双层检索范式不仅通过关键词匹配促进相关实体与关系的高效检索，还通过整合构建的知识图中的相关结构信息提升结果的全面性。

3.3 RETRIEVAL-AUGMENTED ANSWER GENERATION

3.3 检索增强的答案生成

Utilization of Retrieved Information. Utilizing the retrieved information $\psi(q; \hat{\mathcal{D}})$, our LightRAG employs a general-purpose LLM to generate answers based on the collected data. This data comprises concatenated values V from relevant entities and relations, produced by the profiling function $P(\cdot)$. It includes names, descriptions of entities and relations, and excerpts from the original text.

利用检索到的信息。利用检索到的信息 $\psi(q; \hat{\mathcal{D}})$ ，我们的 LightRAG 使用通用 LLM 基于收集到的数据生成答案。这些数据由相关实体和关系的串联值 V 组成，由分析函数 $P(\cdot)$ 产生。它包括实体和关系的名称、描述以及原文摘录。

Context Integration and Answer Generation. By unifying the query with this multi-source text, the LLM generates informative answers tailored to the user's needs, ensuring alignment with the query's intent. This approach streamlines the answer generation process by integrating both context and query into the LLM model, as illustrated in detailed examples (Appendix 7.2).

上下文整合与答案生成。通过将查询与这些多源文本统一，LLM 生成符合用户需求的有信息量答案，确保与查询意图一致。这种方法通过将上下文和查询整合进 LLM 模型来简化答案生成过程，如详细示例所示 (附录 7.2)。

3.4 COMPLEXITY ANALYSIS OF THE LIGHTRAG FRAMEWORK

3.4 LIGHTRAG 框架的复杂性分析

In this section, we analyze the complexity of our proposed LightRAG framework, which can be divided into two main parts. The first part is the graph-based Index phase. During this phase, we use the large language model (LLM) to extract entities and relationships from each chunk of text. As a result, the LLM needs to be called chunk size times. Importantly, there is no additional overhead involved in this process, making our approach highly efficient in managing updates to new text.

在本节中，我们分析所提出的 LightRAG 框架的复杂性，可分为两大部分。第一部分是基于图的索引阶段。在此阶段，我们使用大型语言模型 (LLM) 从每个文本块中提取实体和关系。因此，LLM 需要被调用与块数相同的次数。重要的是，该过程不涉及额外开销，使我们的方法在管理新文本更新时极为高效。

The second part of the process involves the graph-based retrieval phase. For each query, we first utilize the large language model (LLM) to generate relevant keywords. Similar to current Retrieval-Augmented Generation (RAG) systems Gao et al. (2023; 2022); Chan et al. (2024), our retrieval mechanism relies on vector-based search. However, instead of retrieving chunks as in conventional RAG, we concentrate on retrieving entities and relationships. This approach markedly reduces retrieval overhead compared to the community-based traversal method used in GraphRAG.

第二部分是基于图的检索阶段。对于每个查询，我们首先利用大型语言模型 (LLM) 生成相关关键词。与现有的检索增强生成 (RAG) 系统 Gao et al. (2023; 2022); Chan et al. (2024) 类似，我们的检索机制依赖于基于向量的搜索。然而，与传统 RAG 检索文本块不同，我们侧重于检索实体和关系。与 GraphRAG 使用的基于社群的遍历方法相比，该方法显著降低了检索开销。

4 EVALUATION

4 评估

We conduct empirical evaluations on benchmark data to assess the effectiveness of the proposed LightRAG framework by addressing the following research questions: • (RQ1): How does LightRAG compare to existing RAG baseline methods in terms of generation performance? • (RQ2): How do dual-level retrieval and graph-based indexing enhance the generation quality of LightRAG? • (RQ3): What specific advantages does LightRAG demonstrate through case examples in various scenarios? (RQ4): What are the costs associated with LightRAG, as well as its adaptability to data changes?

我们在基准数据上进行实证评估，以通过以下研究问题评估所提出的 LightRAG 框架的有效性：• (RQ1): 在生成性能方面，LightRAG 与现有 RAG 基线方法相比如何？• (RQ2): 双层检索与基于图的索引如何提升 LightRAG 的生成质量？• (RQ3): 通过案例示例，LightRAG 在各种场景中表现出哪些具体优势？(RQ4): LightRAG 的成本如何，以及其对数据变化的适应性如何？

4.1 EXPERIMENTAL SETTINGS

4.1 实验设置

Evaluation Datasets. To conduct a comprehensive analysis of LightRAG, we selected four datasets from the UltraDomain benchmark (Qian et al. 2024). The UltraDomain data is sourced from 428 college textbooks and encompasses 18 distinct domains, including agriculture, social sciences, and humanities. From these, we chose the Agriculture, CS, Legal, and Mix datasets. Each dataset contains between 600,000 and 5,000,000 tokens, with detailed information provided in Table 4 Below is a specific introduction to the four domains utilized in our experiments:

评估数据集。为对 LightRAG 进行全面分析，我们从 UltraDomain 基准 (Qian et al. 2024) 中选取了四个数据集。UltraDomain 数据来自 428 本大学教材，涵盖农业、社会科学与人文学科等 18 个不同领域。我们从中选择了农业、计算机科学 (CS)、法律和混合数据集。每个数据集包含 600,000 到 5,000,000 个标记，详细信息见表 4。以下是我们实验中使用的四个领域的具体介绍：

- **Agriculture:** This domain focuses on agricultural practices, covering a range of topics including bee-keeping, hive management, crop production, and disease prevention.

• 农业: 该领域侧重于农业实践，涵盖养蜂、蜂群管理、作物生产与疾病防治等多方面内容。

- **CS:** This domain focuses on computer science and encompasses key areas of data science and software engineering. It particularly highlights machine learning and big data processing, featuring content on recommendation systems, classification algorithms, and real-time analytics using Spark.

• CS: 该领域侧重于计算机科学，涵盖数据科学与软件工程的关键领域，特别强调机器学习与大数据处理，包含推荐系统、分类算法和基于 Spark 的实时分析等内容。

- **Legal:** This domain centers on corporate legal practices, addressing corporate restructuring, legal agreements, regulatory compliance, and governance, with a focus on the legal and financial sectors.

• 法律: 该领域以公司法律实践为核心，涉及公司重组、法律协议、合规与治理，重点面向法律与金融部门。

- **Mixed:** This domain presents a rich variety of literary, biographical, and philosophical texts, spanning a broad spectrum of disciplines, including cultural, historical, and philosophical studies.

• 混合: 该领域呈现丰富的文学、传记与哲学文本，跨越包括文化、历史与哲学研究在内的广泛学科。

Question Generation. To evaluate the effectiveness of RAG systems for high-level sensemaking tasks, we consolidate all text content from each dataset as context and adopt the generation method outlined in Edge et al. (2024). Specifically, we instruct an LLM to generate five RAG users, along with five tasks for each user. Each generated user is accompanied by a textual description detailing their expertise and traits that motivate their question-raising activities. Each user task is also described, emphasizing one of the user’s potential intentions when interacting with RAG systems. For each user-task combination, the LLM generates five questions that require an understanding of the entire corpus. In total, this process results in 125 questions for each dataset.

问题生成。为评估 RAG 系统在高级理解任务中的有效性，我们将每个数据集的所有文本内容整合为上下文，并采用 Edge et al. (2024) 中的生成方法。具体而言，我们指示 LLM 生成五个 RAG 用户，以及每个用户的五个任务。每个生成的用户都附有描述其专长与促使其提问动机的特征文本。每个用户任务也有描述，强调用户与 RAG 系统交互时可能的意图之一。对于每个用户-任务组合，LLM 生成五个需要理解整个语料库的问题。总体而言，每个数据集产生 125 个问题。

Baselines. LightRAG is compared against the following state-of-the-art methods across all datasets:

基线。LightRAG 在所有数据集上与以下最先进方法进行比较:

- Naive RAG (Gao et al. 2023): This model serves as a standard baseline in existing RAG systems. It segments raw texts into chunks and stores them in a vector database using text embeddings. For queries, Naive RAG generates vectorized representations to directly retrieve text chunks based on the highest similarity in their representations, ensuring efficient and straightforward matching.

- 朴素 RAG (Gao et al. 2023): 该模型作为现有 RAG 系统的标准基线。它将原始文本切分为块并使用文本嵌入将其存储在向量数据库中。对于查询，朴素 RAG 生成向量表示以根据表示相似度直接检索文本块，确保高效且直接的匹配。

- RQ-RAG (Chan et al. 2024): This approach leverages the LLM to decompose the input query into multiple sub-queries. These sub-queries are designed to enhance search accuracy by utilizing explicit techniques such as rewriting, decomposition, and disambiguation.

- RQ-RAG (Chan et al. 2024): 该方法利用 LLM 将输入查询分解为多个子查询。通过重写、分解和消歧等显式技术，这些子查询旨在提高检索的准确性。

- HyDE (Gao et al. 2022): This method utilizes the LLM to generate a hypothetical document based on the input query. This generated document is then employed to retrieve relevant text chunks, which are subsequently used to formulate the final answer.

- HyDE (Gao et al. 2022): 该方法利用 LLM 根据输入查询生成假设性文档。然后使用该生成文档检索相关文本块，随后用这些文本块来构成最终答案。

- GraphRAG (Edge et al. 2024): This is a graph-enhanced RAG system that utilizes an LLM to extract entities and relationships from the text, representing them as nodes and edges. It generates corresponding descriptions for these elements, aggregates nodes into communities, and produces a community report to capture global information. When handling high-level queries, GraphRAG retrieves more comprehensive information by traversing these communities.

- GraphRAG (Edge et al. 2024): 这是一种图增强的 RAG 系统，利用 LLM 从文本中抽取实体与关系并将其表示为节点与边。它为这些元素生成相应描述，将节点聚合为社区，并产出社区报告以捕捉全局信息。处理高层次查询时，GraphRAG 通过遍历这些社区来检索更全面的信息。

Implementation and Evaluation Details. In our experiments, we utilize the nano vector database for vector data management and access. For all LLM-based operations in LightRAG, we default to using GPT-40-mini. To ensure consistency, the chunk size is set to 1200 across all datasets. Additionally, the gleaning parameter is fixed at 1 for both GraphRAG and LightRAG.

实现与评估细节。在我们的实验中，我们使用 nano 向量数据库进行向量数据管理与访问。在 LightRAG 中的所有基于 LLM 的操作默认使用 GPT-40-mini。为确保一致性，所有数据集的分块大小统一设置为 1200。此外，GraphRAG 与 LightRAG 的 gleaning 参数均固定为 1。

Defining ground truth for many RAG queries, particularly those involving complex high-level semantics, poses significant challenges. To address this, we build on existing work (Edge et al., 2024) and adopt an LLM-based multi-dimensional comparison method. We employ a robust LLM, specifically GPT-40-mini, to rank each baseline against our LightRAG. The evaluation prompt we used is detailed in Appendix 7.3.4 In total, we utilize four evaluation dimensions, including:

为许多 RAG 查询定义基准事实，尤其是涉及复杂高层语义的查询，具有显著挑战。为此，我们在现有工作 (Edge et al., 2024) 基础上采用基于 LLM 的多维比较方法。我们使用强健的 LLM，具体为 GPT-40-mini，对每个基线与我们的 LightRAG 进行排序。我们使用的评估提示详见附录 7.3.4。总共我们采用四个评估维度，包括：

- i) Comprehensiveness: How thoroughly does the answer address all aspects and details of the question?
- ii) Diversity: How varied and rich is the answer in offering different perspectives and insights related to the question?
- iii) Empowerment: How effectively does the answer enable the reader to understand the topic and make informed judgments?
- iv) Overall: This dimension assesses the cumulative performance across the three preceding criteria to identify the best overall answer.

i) 完整性: 回答在多大程度上全面涵盖问题的各个方面与细节? ii) 多样性: 回答在提供不同视角与见解方面有多丰富与多样? iii) 赋能性: 回答在多大程度上使读者理解主题并做出明智判断? iv) 综合: 该维度评估前三项标准的累积表现，以确定总体最佳答案。

The LLM directly compares two answers for each dimension and selects the superior response for each criterion. After identifying the winning answer for the three dimensions, the LLM combines the results to determine the overall better answer. To ensure a fair evaluation and mitigate the potential bias that could arise from the order in which the answers are presented in the prompt, we alternate the placement of each answer. We calculate win rates accordingly, ultimately leading to the final results.

LLM 对每个维度直接比较两个答案并为每一标准选择更优回应。在为三项维度确定获胜答案后，LLM 合并结果以判断总体更优答案。为确保公平评估并减少提示中答案呈现顺序可能带来的偏差，我们交替回答的摆放位置。据此计算胜率，最终得出结果。

4.2 COMPARISON OF LIGHTRAG WITH EXISTING RAG METHODS (RQ1)

4.2 LIGHTRAG 与现有 RAG 方法的比较 (RQ1)

We compare LightRAG against each baseline across various evaluation dimensions and datasets. The results are presented in Table 1 Based on these findings, we draw the following conclusions:

我们在不同评估维度与数据集上将 LightRAG 与各基线进行比较。结果呈现在表 1。基于这些发现，我们得出以下结论：

Table 1: Win rates (%) of baselines v.s. LightRAG across four datasets and four evaluation dimensions.

表 1: 基线对比 LightRAG 在四个数据集与四个评估维度上的胜率 (%)。

	Agriculture		CS		Legal		Mix	
	NaiveRAG	LightRAG	NaiveRAG	LightRAG	NaiveRAG	LightRAG	NaiveRAG	LightRAG
Comprehensiveness	32.4%	67.6%	38.4%	61.6%	16.4%	83.6%	38.8%	61.2%
Diversity	23.6%	76.4%	38.0%	62.0%	13.6%	86.4%	32.4%	67.6%
Empowerment	32.4%	67.6%	38.8%	61.2%	16.4%	83.6%	42.8%	57.2%
Overall	32.4%	67.6%	38.8%	61.2%	15.2%	84.8%	40.0%	60.0%
	RQ-RAG	LightRAG	RQ-RAG	LightRAG	RQ-RAG	LightRAG	RQ-RAG	LightRAG
Comprehensiveness	31.6%	68.4%	38.8%	61.2%	15.2%	84.8%	39.2%	60.8%
Diversity	29.2%	70.8%	39.2%	60.8%	11.6%	88.4%	30.8%	69.2%
Empowerment	31.6%	68.4%	36.4%	63.6%	15.2%	84.8%	42.4%	57.6%
Overall	32.4%	67.6%	38.0%	62.0%	14.4%	85.6%	40.0%	60.0%
	HyDE	LightRAG	HyDE	LightRAG	HyDE	LightRAG	HyDE	LightRAG
Comprehensiveness	26.0%	74.0%	41.6%	58.4%	26.8%	73.2%	40.4%	59.6%
Diversity	24.0%	76.0%	38.8%	61.2%	20.0%	80.0%	32.4%	67.6%
Empowerment	25.2%	74.8%	40.8%	59.2%	26.0%	74.0%	46.0%	54.0%
Overall	24.8%	75.2%	41.6%	58.4%	26.4%	73.6%	42.4%	57.6%
	GraphRAG	LightRAG	GraphRAG	LightRAG	GraphRAG	LightRAG	GraphRAG	LightRAG
Comprehensiveness	45.6%	54.4%	48.4%	51.6%	48.4%	51.6%	50.4%	49.6%
Diversity	22.8%	77.2%	40.8%	59.2%	26.4%	73.6%	36.0%	64.0%
Empowerment	41.2%	58.8%	45.2%	54.8%	43.6%	56.4%	50.8%	49.2%
Overall	45.2%	54.8%	48.0%	52.0%	47.2%	52.8%	50.4%	49.6%

	农业		计算机科学		法律		混合	
	NaiveRAG	LightRAG	NaiveRAG	LightRAG	NaiveRAG	LightRAG	NaiveRAG	LightRAG
全面性	32.4%	67.6%	38.4%	61.6%	16.4%	83.6%	38.8%	61.2%
多样性	23.6%	76.4%	38.0%	62.0%	13.6%	86.4%	32.4%	67.6%
赋能	32.4%	67.6%	38.8%	61.2%	16.4%	83.6%	42.8%	57.2%
总体	32.4%	67.6%	38.8%	61.2%	15.2%	84.8%	40.0%	60.0%
	RQ-RAG	LightRAG	RQ-RAG	LightRAG	RQ-RAG	LightRAG	RQ-RAG	LightRAG
全面性	31.6%	68.4%	38.8%	61.2%	15.2%	84.8%	39.2%	60.8%
多样性	29.2%	70.8%	39.2%	60.8%	11.6%	88.4%	30.8%	69.2%
赋能	31.6%	68.4%	36.4%	63.6%	15.2%	84.8%	42.4%	57.6%
总体	32.4%	67.6%	38.0%	62.0%	14.4%	85.6%	40.0%	60.0%
	HyDE	LightRAG	HyDE	LightRAG	HyDE	LightRAG	HyDE	LightRAG
全面性	26.0%	74.0%	41.6%	58.4%	26.8%	73.2%	40.4%	59.6%
多样性	24.0%	76.0%	38.8%	61.2%	20.0%	80.0%	32.4%	67.6%
赋能	25.2%	74.8%	40.8%	59.2%	26.0%	74.0%	46.0%	54.0%
总体	24.8%	75.2%	41.6%	58.4%	26.4%	73.6%	42.4%	57.6%
	GraphRAG	LightRAG	GraphRAG	LightRAG	GraphRAG	LightRAG	GraphRAG	LightRAG
全面性	45.6%	54.4%	48.4%	51.6%	48.4%	51.6%	50.4%	49.6%
多样性	22.8%	77.2%	40.8%	59.2%	26.4%	73.6%	36.0%	64.0%
赋能	41.2%	58.8%	45.2%	54.8%	43.6%	56.4%	50.8%	49.2%
总体	45.2%	54.8%	48.0%	52.0%	47.2%	52.8%	50.4%	49.6%

The Superiority of Graph-enhanced RAG Systems in Large-Scale Corpora When handling large token counts and complex queries that require a thorough understanding of the dataset’s context, graph-based RAG systems like LightRAG and GraphRAG consistently outperform purely chunk-based retrieval methods such as NaiveRAG, HyDE, and RQRAG. This performance gap becomes particularly pronounced as the dataset size increases. For instance, in the largest dataset (Legal), the disparity widens significantly, with baseline methods achieving only about 20% win rates compared to the dominance of LightRAG. This trend underscores the advantages of graph-enhanced RAG systems in capturing complex semantic dependencies within large-scale corpora, facilitating a more comprehensive understanding of knowledge and leading to improved generalization performance.

在处理大量 token 和需要全面理解数据集上下文的复杂查询时，基于图的 RAG 系统 (如 LightRAG 和 GraphRAG) 在大规模语料库中持续优于纯基于块的检索方法 (如 NaiveRAG、HyDE 和 RQRAG)。随着数据集规模增大，这种性能差距尤为明显。例如，在最大的数据集 (Legal) 中，差距显著扩大，基线方法的胜率仅约为 20%，而 LightRAG 占据主导。该趋势强调了图增强 RAG 系统在捕捉大规模语料库中复杂语义依赖方面的优势，有助于更全面地理解知识并提升泛化性能。

Enhancing Response Diversity with LightRAG: Compared to various baselines, LightRAG demonstrates a significant advantage in the Diversity metric, particularly within the larger Legal dataset. Its consistent lead in this area underscores LightRAG’s effectiveness in generating a wider range of responses, especially in scenarios where diverse content is essential. We attribute this advantage to LightRAG’s dual-level retrieval paradigm, which facilitates comprehensive information retrieval from both low-level and high-level dimensions. This approach effectively leverages graph-based text indexing to consistently capture the full context in response to queries.

提高响应多样性: 与多种基线相比, LightRAG 在 Diversity 指标上表现出显著优势, 尤其在较大的 Legal 数据集中其持续领先, 强调了 LightRAG 在生成更丰富响应方面的有效性, 尤其在需要内容多样性的场景中。我们将此优势归因于 LightRAG 的双层检索范式, 该范式有助于从低层次和高层次维度进行全面信息检索。该方法有效利用基于图的文本索引, 以在响应查询时稳定捕捉完整上下文。

LightRAG's Superiority over GraphRAG: While both LightRAG and GraphRAG use graph-based retrieval mechanisms, LightRAG consistently outperforms GraphRAG, particularly in larger datasets with complex language contexts. In the Agriculture, CS, and Legal datasets—each containing millions of tokens—LightRAG shows a clear advantage, significantly surpassing GraphRAG and highlighting its strength in comprehensive information understanding within diverse environments. **Enhanced Response Variety:** By integrating low-level retrieval of specific entities with high-level retrieval of broader topics, LightRAG boosts response diversity. This dual-level mechanism effectively addresses both detailed and abstract queries, ensuring a thorough grasp of information. **Complex Query Handling:** This approach is especially valuable in scenarios requiring diverse perspectives. By accessing both specific details and overarching themes, LightRAG adeptly responds to complex queries involving interconnected topics, providing contextually relevant answers.

LightRAG 优于 GraphRAG: 尽管 LightRAG 和 GraphRAG 都使用基于图的检索机制, 但 LightRAG 在更大且语言上下文复杂的数据集中持续优于 GraphRAG。在包含百万级 tokens 的 Agriculture、CS 和 Legal 数据集中, LightRAG 展现出明显优势, 显著超越 GraphRAG, 突出其在多样化环境中对信息的全面理解能力。增强的响应多样性: 通过将低层次的具体实体检索与高层次的主题检索相结合, LightRAG 提升了响应多样性。该双层机制有效应对细节型和抽象型查询, 确保对信息的全面把握。复杂查询处理: 在需要多角度视野的场景中, 该方法尤其有价值。通过同时获取具体细节和整体主题, LightRAG 能够灵活应对涉及相互关联主题的复杂查询, 提供语境相关的答案。

4.3 ABLATION STUDIES (RQ2)

4.3 消融研究 (RQ2)

We also conduct ablation studies to evaluate the impact of our dual-level retrieval paradigm and the effectiveness of our graph-based text indexing in LightRAG. The results are presented in Table 2

我们还进行了消融研究, 以评估双层检索范式和 LightRAG 中基于图的文本索引的有效性。结果呈现在表 2 中

Effectiveness of Dual-level Retrieval Paradigm. We begin by analyzing the effects of low-level and high-level retrieval paradigms. We compare two ablated models—each omitting one module—against LightRAG across four datasets. Here are our key observations for the different variants:

双层检索范式的有效性。我们首先分析低层次和高层次检索范式的影响。我们将两种各省略一个模块的消融模型与 LightRAG 在四个数据集上进行比较。以下是对不同变体的主要观察结果:

Table 2: Performance of ablated versions of LightRAG, using NaiveRAG as reference.

表 2:LightRAG 消融版本的性能，以 NaiveRAG 为参考。

	Agriculture		CS		Legal		Mix	
	NaiveRAG	LightRAG	NaiveRAG	LightRAG	NaiveRAG	LightRAG	NaiveRAG	LightRAG
Comprehensiveness	32.4%	67.6%	38.4%	61.6%	16.4%	83.6%	38.8%	61.2%
Diversity	23.6%	76.4%	38.0%	62.0%	13.6%	86.4%	32.4%	67.6%
Empowerment	32.4%	67.6%	38.8%	61.2%	16.4%	83.6%	42.8%	57.2%
Overall	32.4%	67.6%	38.8%	61.2%	15.2%	84.8%	40.0%	60.0%
	NaiveRAG	-High	NaiveRAG	-High	NaiveRAG	-High	NaiveRAG	-High
Comprehensiveness	34.8%	65.2%	42.8%	57.2%	23.6%	76.4%	40.4%	59.6%
Diversity	27.2%	72.8%	36.8%	63.2%	16.8%	83.2%	36.0%	64.0%
Empowerment	36.0%	64.0%	42.4%	57.6%	22.8%	77.2%	47.6%	52.4%
Overall	35.2%	64.8%	44.0%	56.0%	22.0%	78.0%	42.4%	57.6%
	NaiveRAG	-Low	NaiveRAG	-Low	NaiveRAG	-Low	NaiveRAG	-Low
Comprehensiveness	36.0%	64.0%	43.2%	56.8%	19.2%	80.8%	36.0%	64.0%
Diversity	28.0%	72.0%	39.6%	60.4%	13.6%	86.4%	33.2%	66.8%
Empowerment	34.8%	65.2%	42.8%	57.2%	16.4%	83.6%	35.2%	64.8%
Overall	34.8%	65.2%	43.6%	56.4%	18.8%	81.2%	35.2%	64.8%
	NaiveRAG	-Origin	NaiveRAG	-Origin	NaiveRAG	-Origin	NaiveRAG	-Origin
Comprehensiveness	24.8%	75.2%	39.2%	60.8%	16.4%	83.6%	44.4%	55.6%
Diversity	26.4%	73.6%	44.8%	55.2%	14.4%	85.6%	25.6%	74.4%
Empowerment	32.0%	68.0%	43.2%	56.8%	17.2%	82.8%	45.2%	54.8%
Overall	25.6%	74.4%	39.2%	60.8%	15.6%	84.4%	44.4%	55.6%

	农业		计算机科学		法律		混合	
	朴素检索增强生成	轻量检索增强生成	朴素检索增强生成	轻量检索增强生成	朴素检索增强生成	轻量检索增强生成	朴素检索增强生成	轻量检索增强生成
完整性	32.4%	67.6%	38.4%	61.6%	16.4%	83.6%	38.8%	61.2%
多样性	23.6%	76.4%	38.0%	62.0%	13.6%	86.4%	32.4%	67.6%
赋能	32.4%	67.6%	38.8%	61.2%	16.4%	83.6%	42.8%	57.2%
总体	32.4%	67.6%	38.8%	61.2%	15.2%	84.8%	40.0%	60.0%
	朴素检索增强生成	-高	朴素检索增强生成	-高	朴素检索增强生成	-高	朴素检索增强生成	-高
完整性	34.8%	65.2%	42.8%	57.2%	23.6%	76.4%	40.4%	59.6%
多样性	27.2%	72.8%	36.8%	63.2%	16.8%	83.2%	36.0%	64.0%
赋能	36.0%	64.0%	42.4%	57.6%	22.8%	77.2%	47.6%	52.4%
总体	35.2%	64.8%	44.0%	56.0%	22.0%	78.0%	42.4%	57.6%
	朴素检索增强生成	-低	朴素检索增强生成	-低	朴素检索增强生成	-低	朴素检索增强生成	-低
完整性	36.0%	64.0%	43.2%	56.8%	19.2%	80.8%	36.0%	64.0%
多样性	28.0%	72.0%	39.6%	60.4%	13.6%	86.4%	33.2%	66.8%
赋能	34.8%	65.2%	42.8%	57.2%	16.4%	83.6%	35.2%	64.8%
总体	34.8%	65.2%	43.6%	56.4%	18.8%	81.2%	35.2%	64.8%
	朴素检索增强生成	-来源	朴素检索增强生成	-来源	朴素检索增强生成	-来源	朴素检索增强生成	-来源
完整性	24.8%	75.2%	39.2%	60.8%	16.4%	83.6%	44.4%	55.6%
多样性	26.4%	73.6%	44.8%	55.2%	14.4%	85.6%	25.6%	74.4%
赋能	32.0%	68.0%	43.2%	56.8%	17.2%	82.8%	45.2%	54.8%
总体	25.6%	74.4%	39.2%	60.8%	15.6%	84.4%	44.4%	55.6%

- Low-level-only Retrieval: The -High variant removes high-order retrieval, leading to a significant performance decline across nearly all datasets and metrics. This drop is mainly due to its emphasis on the specific information, which focuses excessively on entities and their immediate neighbors. While this approach enables deeper exploration of directly related entities, it struggles to gather information for complex queries that demand comprehensive insights.

- 仅低层检索:-High 变体移除了高阶检索，导致几乎在所有数据集和评估指标上性能显著下降。此下降主要源于其对特定信息的偏重，过分关注实体及其直接邻域。尽管这种方法能更深入地挖掘直接相关的实体，但在需要全面洞察的复杂查询中难以收集足够信息。

- High-level-only Retrieval: The -Low variant prioritizes capturing a broader range of content by leveraging entity-wise relationships rather than focusing on specific entities. This approach offers a significant

advantage in comprehensiveness, allowing it to gather more extensive and varied information. However, the trade-off is a reduced depth in examining specific entities, which can limit its ability to provide highly detailed insights. Consequently, this high-level-only retrieval method may struggle with tasks that require precise, detailed answers.

- 仅高层检索:-Low 变体倾向于通过利用按实体的关系来捕捉更广泛的内容，而不是聚焦于具体实体。该方法在全面性上具有明显优势，能够收集更广泛且多样的信息。但代价是对具体实体的深度检视减少，限制了其提供高度详尽见解的能力。因此，这种仅高层的检索方法在需要精确、细致答案的任务上可能表现不足。

- Hybrid Mode: The hybrid mode, or the full version of LightRAG, combines the strengths of both low-level and high-level retrieval methods. It retrieves a broader set of relationships while simultaneously conducting an in-depth exploration of specific entities. This dual-level approach ensures both breadth in the retrieval process and depth in the analysis, providing a comprehensive view of the data. As a result, LightRAG achieves balanced performance across multiple dimensions.

- 混合模式: 混合模式，即 LightRAG 的完整版本，结合了低层和高层检索方法的优势。它在检索更广泛的关系集的同时，对特定实体进行深入探查。该双层方法既保证了检索过程的广度，又保证了分析的深度，提供了对数据的全面视角。因此，LightRAG 在多个维度上实现了均衡的性能。

Semantic Graph Excels in RAG. We eliminated the use of original text in our retrieval process. Surprisingly, the resulting variant, -Origin, does not exhibit significant performance declines across all four datasets. In some cases, this variant even shows improvements (e.g. in Agriculture and Mix). We attribute this phenomenon to the effective extraction of key information during the graph-based indexing process, which provides sufficient context for answering queries. Additionally, the original text often contains irrelevant information that can introduce noise in the response.

语义图在 RAG 中表现出色。我们在检索过程中去除了原始文本。出人意料的是，得到的变体 -Origin 在四个数据集上并未表现出显著的性能下降。在某些情况下(如 Agriculture 和 Mix)，该变体甚至有所提升。我们将此归因于图形化索引过程中对关键信息的有效抽取，这为回答查询提供了充分的上下文。此外，原始文本中常包含与问题无关的信息，可能为响应引入噪声。

4.4 CASE STUDY (RQ3)

4.4 案例研究 (RQ3)

To provide a clear comparison between baseline methods and our LightRAG, we present specific case examples in Table 3, which includes responses to a machine learning question from both the competitive baseline, GraphRAG, and our LightRAG framework. In this instance, LightRAG outperforms in all evaluation dimensions assessed by the LLM judge, including comprehensiveness, diversity, empowerment, and overall quality. Our key observations are as follows:

为清晰比较基线方法与我们的 LightRAG，我们在表 3 中给出具体案例示例，展示来自竞争性基线 GraphRAG 和我们 LightRAG 框架对同一机器学习问题的回答。在该实例中，LightRAG 在由 LLM 评审的所有评估维度上均优于对手，包括全面性、多样性、赋能性和整体质量。我们主要观察如下：

i) Comprehensiveness. Notably, LightRAG covers a broader range of machine learning metrics, showcasing its comprehensiveness and ability to effectively discover relevant information. This highlights the strength of our graph-based indexing paradigm, which excels in precise entity and relation extraction as well as LLM profiling. ii) Both Diversity and Empowerment. Furthermore, LightRAG not only offers a more diverse array of information but also delivers more empowering content. This success is due to LightRAG’s hierarchical retrieval paradigm, which combines in-depth explorations of related entities through low-level retrieval to enhance empowerment with broader explorations via high-level retrieval to improve answer diversity. Together, these approaches capture a comprehensive global perspective of the knowledge domain, contributing to better RAG performance.

i) 全面性。值得注意的是，LightRAG 覆盖了更广范围的机器学习指标，展现出其全面性和有效发现相关信息的能力。这凸显了我们基于图的索引范式在精确实体与关系抽取及 LLM 画像方面的优势。ii) 多样性与赋能性。此外，LightRAG 不仅提供更为多样的信息，也带来了更具赋能性的内容。该成功归因于 LightRAG 的分层检索范式：通过低层检索对相关实体进行深入探索以增强赋能性，并通过高层检索进行更广泛的探索以提升答案多样性。两者结合捕捉了知识领域的全面全局视角，有助于提升 RAG 性能。

Table 3: Case Study: Comparison Between LightRAG and the Baseline Method GraphRAG.

表 3: 案例研究:LightRAG 与基线方法 GraphRAG 的比较。

Model	Response
GraphRAG	GraphRAG response text...
LightRAG	LightRAG response text...

4.5 MODEL COST AND ADAPTABILITY ANALYSIS (RQ4)

4.5 模型成本与适应性分析 (RQ4)

We compare the cost of our LightRAG with that of the top-performing baseline, GraphRAG, from two key perspectives. First, we examine the number of tokens and API calls during the indexing and retrieval processes. Second, we analyze these metrics in relation to handling data changes in dynamic environments. The results of this evaluation on the legal dataset are presented in Table 2. In this context, T_{extract} represents the token overhead for entity and relationship extraction, C_{max} denotes the maximum number of tokens allowed per API call, and C_{extract} indicates the number of API calls required for extraction.

我们从两个关键角度将 LightRAG 的成本与表现最优的基线 GraphRAG 进行比较。首先，考察索引与检索过程中令牌数和 API 调用次数；其次，分析这些指标在动态环境中应对数据变更时的表现。针对法律数据集的评估结果见表 2。其中， T_{extract} 表示实体与关系抽取的令牌开销， C_{max} 表示每次 API 调用允许的最大令牌数， C_{extract} 表示抽取所需的 API 调用次数。

Figure 2: Comparison of Cost in Terms of Tokens and API Calls for GraphRAG and LightRAG on the Legal Dataset.

图 2: 在法律数据集上，GraphRAG 与 LightRAG 在令牌数与 API 调用次数方面的成本比较。

Phase	Retrieval Phase		Incremental Text Update	
Model	GraphRAG	Ours	GraphRAG	Ours
Tokens	610×1,000	< 100	$1,399 \times 2 \times 5,000 + T_{\text{extract}}$	T_{extract}
API Calls	610×1,000 C_{max}	1	$1,399 \times 2 + C_{\text{extract}}$	C_{extract}

阶段	检索阶段		增量文本更新	
模型	GraphRAG	我们的	GraphRAG	我们的
令牌	610×1,000	< 100	$1,399 \times 2 \times 5,000 + T_{\text{extract}}$	T_{extract}
API 调用	610×1,000 C_{max}	1	$1,399 \times 2 + C_{\text{extract}}$	C_{extract}

In the retrieval phase, GraphRAG generates 1,399 communities, with 610 level-2 communities actively utilized for retrieval in this experiment. Each community report averages 1,000 tokens, resulting in a total token consumption of 610,000 tokens (610 communities × 1,000 tokens per community). Additionally, GraphRAG’s requirement to traverse each community individually leads to hundreds of API calls, significantly increasing retrieval overhead. In contrast, LightRAG optimizes this process by using fewer than 100 tokens for keyword generation and retrieval, requiring only a single API call for the entire process. This efficiency is achieved through our retrieval mechanism, which seamlessly integrates graph structures and vectorized representations for information retrieval, thereby eliminating the need to process large volumes of information upfront.

在检索阶段，GraphRAG 生成了 1,399 个社区，本次实验中实际用于检索的有 610 个二级社区。每个社区报告平均为 1,000 个标记，导致总标记消耗为 610,000 标记（每个社区 × 1,000 标记，共 610 个社区）。此外，GraphRAG 需要逐一遍历每个社区，导致数百次 API 调用，显著增加检索开销。相比之下，LightRAG 优化了该过程：用于关键词生成和检索的标记少于 100 个，并且仅需一次 API 调用即可完成整个流程。该效率通过我们的检索机制实现，该机制将图结构与向量化表示无缝结合用于信息检索，从而消除了预先处理大量信息的必要。

In the incremental data update phase, designed to address changes in dynamic real-world scenarios, both models exhibit similar overhead for entity and relationship extraction. However, GraphRAG shows significant inefficiency in managing newly added data. When a new dataset of the same size as the legal dataset is introduced, GraphRAG must dismantle its existing community structure to incorporate new entities and relationships, followed by complete regeneration. This process incurs a substantial token cost of approximately 5,000 tokens per community report. Given 1,399 communities, GraphRAG would require around $1,399 \times 2 \times 5,000$

tokens to reconstruct both the original and new community reports—an exorbitant expense that underscores its inefficiency. In contrast, LightRAG seamlessly integrates newly extracted entities and relationships into the existing graph without the need for full reconstruction. This approach results in significantly lower overhead during incremental updates, demonstrating its superior efficiency and cost-effectiveness.

在为应对动态真实场景变化而设计的增量数据更新阶段，两种模型在实体与关系抽取方面的开销相近。然而，GraphRAG 在管理新增数据时表现出明显低效。当引入与法律数据集相同规模的新数据集时，GraphRAG 必须拆解现有社区结构以纳入新实体与关系，然后重新完整生成。该过程为每个社区报告带来约 5,000 个标记的高昂代价。鉴于有 1,399 个社区，GraphRAG 重建原有与新社区报告将需要约 $1,399 \times 2 \times 5,000$ 标记——这一天文数字凸显了其低效性。相比之下，LightRAG 能将新提取的实体与关系无缝整合进现有图结构，无需完全重构，从而在增量更新时显著降低开销，展示出更高的效率与成本效益。

5 RELATED WORK

5 相关工作

5.1 RETRIEVAL-AUGMENTED GENERATION WITH LLMs

5.1 与大模型结合的检索增强生成

Retrieval-Augmented Generation (RAG) systems enhance LLM inputs by retrieving relevant information from external sources, grounding responses in factual, domain-specific knowledge Ram et al. (2023); Fan et al. (2024). Current RAG approaches Gao et al. (2022); 2023); Chan et al. (2024); Yu et al. (2024) typically embed queries in a vector space to find the nearest context vectors. However, many of these methods rely on fragmented text chunks and only retrieve the top-k contexts, limiting their ability to capture comprehensive global information needed for effective responses.

检索增强生成 (RAG) 系统通过从外部来源检索相关信息来强化大模型输入，使回答以事实性、领域特定知识为依据 Ram et al. (2023); Fan et al. (2024)。当前的 RAG 方法 Gao et al. (2022); 2023); Chan et al. (2024); Yu et al. (2024) 通常将查询嵌入向量空间以查找最近的上下文向量。然而，许多方法依赖于碎片化的文本块且仅检索 top-k 上下文，限制了其捕捉生成有效回答所需的全面全局信息的能力。

Although recent studies Edge et al. (2024) have explored using graph structures for knowledge representation, two key limitations persist. First, these approaches often lack the capability for dynamic updates and expansions of the knowledge graph, making it difficult to incorporate new information effectively. In contrast, our proposed model, LightRAG, addresses these challenges by enabling the RAG system to quickly adapt to new information, ensuring the model’s timeliness and accuracy. Additionally, existing methods often rely on brute-force searches for each generated community, which are inefficient for large-scale queries. Our LightRAG framework overcomes this limitation by facilitating rapid retrieval of relevant information from the graph through our proposed dual-level retrieval paradigm, significantly enhancing both retrieval efficiency and response speed.

尽管近期研究 Edge et al. (2024) 探讨了使用图结构进行知识表示, 但仍存在两大关键局限。首先, 这些方法往往缺乏对知识图进行动态更新与扩展的能力, 难以有效纳入新信息。与之相对, 我们提出的 LightRAG 通过使 RAG 系统能够快速适应新信息来解决这些挑战, 确保模型的时效性与准确性。此外, 现有方法常对每个生成的社区依赖暴力搜索, 这对于大规模查询效率低下。我们的 LightRAG 框架通过提出的双层检索范式, 使得从图中快速检索相关信息成为可能, 显著提升了检索效率与响应速度。

5.2 LARGE LANGUAGE MODEL FOR GRAPHS

5.2 面向图的巨型语言模型

Graphs are a powerful framework for representing complex relationships and find applications in numerous fields. As Large Language Models (LLMs) continue to evolve, researchers have increasingly focused on enhancing their capability to interpret graph-structured data. This body of work can be divided into three primary categories: i) GNNs as Prefix where Graph Neural Networks (GNNs) are utilized as the initial processing layer for graph data, generating structure-aware tokens that LLMs can use during inference. Notable examples include GraphGPT Tang et al. (2024) and LLaGA Chen et al. (2024). ii) LLMs as Prefix involves LLMs processing graph data enriched with textual information to produce node embeddings or labels, ultimately refining the training process for GNNs, as demonstrated in systems like GALM Xie et al. (2023) and OFA Liu et al. (2024). iii) LLMs-Graphs Integration focuses on achieving a seamless interaction between LLMs and graph data, employing techniques such as fusion training and GNN alignment, and developing LLM-based agents capable of engaging with graph information directly Li et al. (2023); Brannon et al. (2023).

图是表示复杂关系的强大框架, 并在众多领域中得到应用。随着巨型语言模型 (LLMs) 的不断发展, 研究者越来越关注增强其解读图结构数据的能力。这类工作可分为三类: i) 将图神经网络 (GNNs) 作为前缀, 即将 GNN 作为图数据的初始处理层, 生成结构感知的标记供 LLM 在推理时使用, 代表性例子包括 GraphGPT Tang et al. (2024) 与 LLaGA Chen et al. (2024)。ii) 将 LLMs 作为前缀, 指 LLM 处理富含文本信息的图数据以生成节点嵌入或标签, 进而优化 GNN 的训练流程, 如 GALM Xie et al. (2023) 与 OFA Liu et al. (2024) 等系统所示。iii) LLMs-图集成, 侧重实现 LLM 与图数据的无缝交互, 采用融合训练与 GNN 对齐等技术, 并开发能够直接与图信息交互的基于 LLM 的智能体 Li et al. (2023); Brannon et al. (2023)。

6 CONCLUSION

6 结论

This work introduces an advancement in Retrieval-Augmented Generation (RAG) through the integration of a graph-based indexing approach that enhances both efficiency and comprehension in information retrieval. LightRAG utilizes a comprehensive knowledge graph to facilitate rapid and relevant document retrieval, enabling a deeper understanding of complex queries. Its dual-level retrieval paradigm allows for the extraction of both specific and abstract information, catering to diverse user needs. Furthermore, LightRAG's seamless incremental update capability ensures that the system remains current and responsive to new information, thereby maintaining its effectiveness over time. Overall, LightRAG excels in both efficiency and

effectiveness, significantly improving the speed and quality of information retrieval and generation while reducing costs for LLM inference.

本工作提出通过引入基于图的索引方法来改进检索增强生成 (RAG), 从而提升信息检索的效率与理解能力。LightRAG 利用全面的知识图谱实现快速且相关的文档检索, 促进对复杂查询的更深刻理解。其双层检索范式兼顾具体与抽象信息的提取, 以满足多样化用户需求。此外, LightRAG 无缝的增量更新能力确保系统对新信息保持同步与响应, 从而随时间维持其有效性。总体而言, LightRAG 在效率与效果上表现卓越, 显著提升信息检索与生成的速度与质量, 同时降低 LLM 推理成本。

REFERENCES

REFERENCES

William Brannon, Suyash Fulay, Hang Jiang, Wonjune Kang, Brandon Roy, Jad Kabbara, and Deb Roy. Congrat: Self-supervised contrastive pretraining for joint graph and text embeddings. arXiv preprint arXiv:2305.14321, 2023.

William Brannon, Suyash Fulay, Hang Jiang, Wonjune Kang, Brandon Roy, Jad Kabbara, and Deb Roy. Congrat: Self-supervised contrastive pretraining for joint graph and text embeddings. arXiv preprint arXiv:2305.14321, 2023.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. Rq-rag: Learning to refine queries for retrieval augmented generation. arXiv preprint arXiv:2404.00610, 2024.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. Rq-rag: Learning to refine queries for retrieval augmented generation. arXiv preprint arXiv:2404.00610, 2024.

Runjin Chen, Tong Zhao, AJAY KUMAR JAISWAL, Neil Shah, and Zhangyang Wang. Llaga: Large language and graph assistant. In International Conference on Machine Learning (ICML), 2024.

Runjin Chen, Tong Zhao, AJAY KUMAR JAISWAL, Neil Shah, and Zhangyang Wang. Llaga: Large language and graph assistant. In International Conference on Machine Learning (ICML), 2024.

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.

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.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In International Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 150-158, 2024.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In International Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 150-158, 2024.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In International Conference on Knowledge Discovery and Data Mining (KDD), pp. 6491-6501, 2024.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In International Conference on Knowledge Discovery and Data Mining (KDD), pp. 6491-6501, 2024.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. arXiv preprint arXiv:2212.10496, 2022.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. arXiv preprint arXiv:2212.10496, 2022.

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.

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.

Yichuan Li, Kaize Ding, and Kyumin Lee. Grenade: Graph-centric language model for self-supervised representation learning on text-attributed graphs. In International Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2745-2757, 2023.

Yichuan Li, Kaize Ding, and Kyumin Lee. Grenade: Graph-centric language model for self-supervised representation learning on text-attributed graphs. In International Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2745-2757, 2023.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks. In International Conference on Learning Representations (ICLR), 2024.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks. In International Conference on Learning Representations (ICLR), 2024.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. arXiv preprint arXiv:2401.17043, 2024.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. arXiv preprint arXiv:2401.17043, 2024.

Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery, 2024. URL <https://arxiv.org/abs/2409.05591>

Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery, 2024. URL <https://arxiv.org/abs/2409.05591>

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics (TACL), 11:1316-1331, 2023.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics (TACL), 11:1316-1331, 2023.

Ladislav Rampásek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. International Conference on Neural Information Processing Systems (NeurIPS), 35:14501-14515, 2022.

Ladislav Rampásek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. 通用、强大且可扩展的图变换器配方。国际神经信息处理系统大会 (NeurIPS), 35:14501-14515, 2022。

Alireza Salemi and Hamed Zamani. Evaluating retrieval quality in retrieval-augmented generation. In ACM International Conference on Research and Development in Information Retrieval (SIGIR), pp. 2395-2400, 2024.

Alireza Salemi and Hamed Zamami. 在检索增强生成中评估检索质量。ACM 信息检索研究与发展国际会议 (SIGIR), 第 2395-2400 页, 2024。

Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. Rag-ex: A generic framework for explaining retrieval augmented generation. In ACM International Conference on Research and Development in Information Retrieval (SIGIR), pp. 2776-2780, 2024.

Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. Rag-ex: 一个用于解释检索增强生成的通用框架。ACM 信息检索研究与发展国际会议 (SIGIR), 第 2776-2780 页, 2024。

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. In ACM International Conference on Research and Development in Information Retrieval (SIGIR), pp. 491-500, 2024.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. GraphGPT: 面向大型语言模型的图指令微调。ACM 信息检索研究与发展国际会议 (SIGIR), 第 491-500 页, 2024。

Shangqing Tu, Yuanchun Wang, Jifan Yu, Yuyang Xie, Yaran Shi, Xiaozhi Wang, Jing Zhang, Lei Hou, and Juanzi Li. R-eval: A unified toolkit for evaluating domain knowledge of retrieval augmented large language models. In International Conference on Knowledge Discovery and Data Mining (KDD), pp. 5813-5824, 2024.

Shangqing Tu, Yuanchun Wang, Jifan Yu, Yuyang Xie, Yaran Shi, Xiaozhi Wang, Jing Zhang, Lei Hou, and Juanzi Li. R-eval: 一个用于评估检索增强大型语言模型领域知识的统一工具包。知识发现与数据挖掘国际会议 (KDD), 第 5813-5824 页, 2024。

Han Xie, Da Zheng, Jun Ma, Houyu Zhang, Vassilis N Ioannidis, Xiang Song, Qing Ping, Sheng Wang, Carl Yang, Yi Xu, et al. Graph-aware language model pre-training on a large graph corpus can help multiple graph applications. In International Conference on Knowledge Discovery and Data Mining (KDD), pp. 5270-5281, 2023.

Han Xie, Da Zheng, Jun Ma, Houyu Zhang, Vassilis N Ioannidis, Xiang Song, Qing Ping, Sheng Wang, Carl Yang, Yi Xu, et al. 在大规模图语料上进行图感知语言模型预训练可助力多种图应用。知识发现与数据挖掘国际会议 (KDD), 第 5270-5281 页, 2023。

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. arXiv preprint arXiv:2407.02485, 2024.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. RankRAG: 在大型语言模型中统一上下文排序与检索增强生成。arXiv 预印本 arXiv:2407.02485, 2024。

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. arXiv preprint arXiv:2402.19473, 2024.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 针对 AI 生成内容的检索增强生成综述。arXiv 预印本 arXiv:2402.19473, 2024。

7 APPENDIX

7 附录

In this section, we elaborate on the methodologies and experimental settings used in the LightRAG framework. It describes the specific steps for extracting entities and relationships from documents, detailing how large language models (LLMs) are utilized for this purpose. The section also specifies the prompt templates

and configurations used in LLM operations, ensuring clarity in the experimental setup. Additionally, it outlines the evaluation criteria and dimensions used to assess the performance of LightRAG against baselines from various dimensions.

本节详述 LightRAG 框架中的方法与实验设置，描述从文档中抽取实体与关系的具体步骤，说明如何利用大型语言模型 (LLMs) 完成该任务；并给出 LLM 操作中使用的提示模板与配置，确保实验设置清晰；同时列出用于从多个维度评估 LightRAG 与基线性能的评估标准与维度。

7.1 EXPERIMENTAL DATA DETAILS

7.1 实验数据详情

Table 4: Statistical information of the datasets.

表 4: 数据集的统计信息。

Statistics	Agriculture	CS	Legal	Mix
Total Documents	12	10	94	61
Total Tokens	2,017,886	2,306,535	5,081,069	619,009

统计	农业	计算机科学	法律	混合
文档总数	12	10	94	61
总标记数	2,017,886	2,306,535	5,081,069	619,009

Table 4 presents statistical information for four datasets: Agriculture, CS, Legal, and Mix. The Agriculture dataset consists of 12 documents totaling 2,017,886 tokens, while the CS dataset contains 10 documents with 2,306,535 tokens. The Legal dataset is the largest, comprising 94 documents and 5,081,069 tokens. Lastly, the Mix dataset includes 61 documents with a total of 619,009 tokens.

表 4 列出了四个数据集的统计信息:Agriculture、CS、Legal 和 Mix。Agriculture 数据集包含 12 篇文档，合计 2,017,886 个标记；CS 数据集包含 10 篇文档，合计 2,306,535 个标记。Legal 数据集最大，包含 94 篇文档，合计 5,081,069 个标记。最后，Mix 数据集包含 61 篇文档，合计 619,009 个标记。

7.2 CASE EXAMPLE OF RETRIEVAL-AUGMENTED GENERATION IN LIGHTRAG.

7.2 LIGHTRAG 中检索增强生成的案例示例。

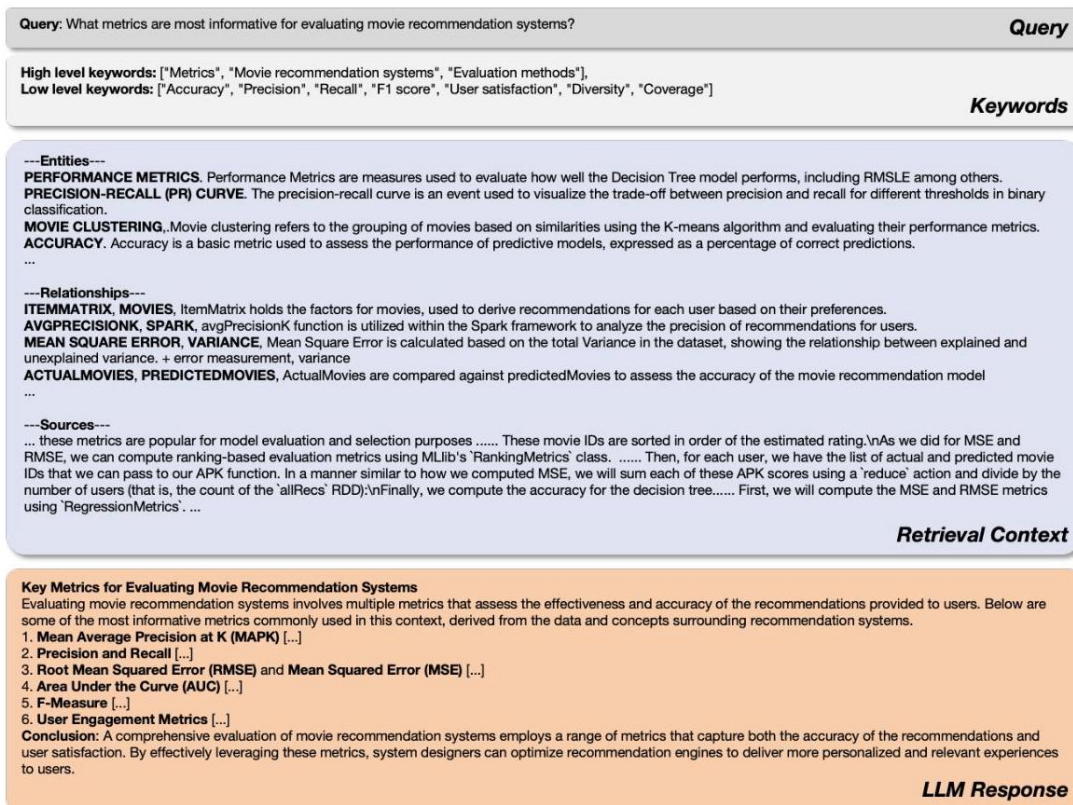


Figure 3: A retrieval and generation example.

图 3: 一个检索与生成的示例。

In Figure 3 we illustrate the retrieve-and-generate process. When presented with the query, "What metrics are most informative for evaluating movie recommendation systems?", the LLM first extracts both low-level and high-level keywords. These keywords guide the dual-level retrieval process on the generated knowledge graph, targeting relevant entities and relationships. The retrieved information is organized into three components: entities, relationships, and corresponding text chunks. This structured data is then fed into the LLM, enabling it to generate a comprehensive answer to the query.

在图 3 中我们展示了检索-生成流程。面对查询“评估电影推荐系统最有信息量的指标有哪些？”时，LLM 首先提取低层次和高层次关键词。这些关键词引导在生成的知识图上进行双层检索，定位相关实体与关系。检索到的信息被组织为三部分：实体、关系及相应的文本片段。然后将这些结构化数据输入 LLM，使其生成对该查询的全面回答。

7.3 OVERVIEW OF THE PROMPTS USED IN LIGHTRAG

7.3 LIGHTRAG 中使用提示的概览

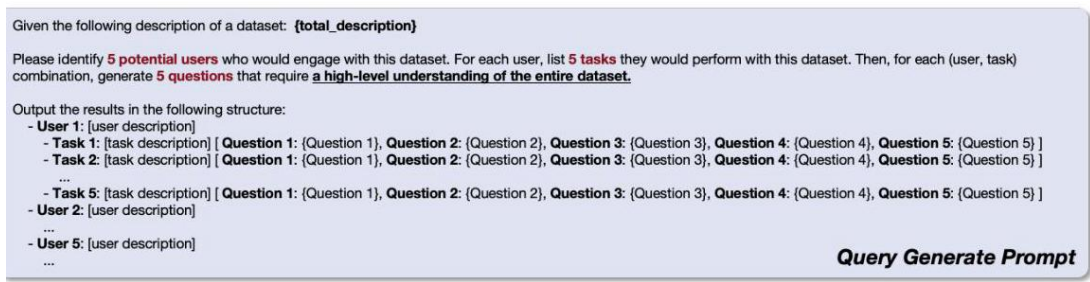


Figure 5: Prompts for Query Generation

图 5: 查询生成的提示

In Figure 5, the query generation prompt outlines a framework for identifying potential user roles (e.g., data scientist, finance analyst, and product manager) and their objectives for generating queries based on a specified dataset description. The prompt explains how to define five distinct users who would benefit from interacting with the dataset. For each user, it specifies five key tasks they would perform while working with the dataset. Additionally, for each (user, task) combination, five high-level questions are posed to ensure a thorough understanding of the dataset.

在图 5 中，查询生成提示概述了一个框架，用于识别潜在用户角色（例如数据科学家、金融分析师和产品经理）及其基于指定数据集描述生成查询的目标。提示说明如何定义五类会从与该数据集交互中受益的用户。对每位用户，指定他们在使用数据集时会执行的五项关键任务。此外，对每个（用户，任务）组合，提出五个高层次问题以确保对数据集的全面理解。

7.3.3 PROMPTS FOR KEYWORD EXTRACTION

7.3.3 用于关键词提取的提示

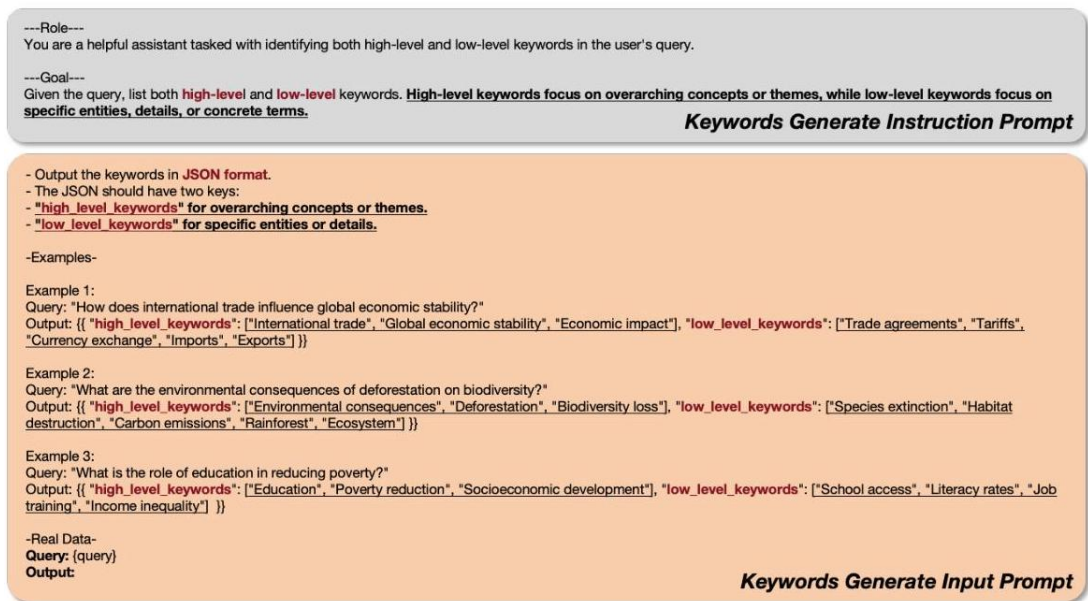


Figure 6: Prompts for Keyword Extraction

图 6: 关键词提取的提示

In Figure 6, the prompt describes a method for extracting keywords from a user's query, distinguishing between high-level and low-level keywords. High-level keywords represent broad concepts or themes, while low-level keywords focus on specific entities and details. The extracted keywords are returned in JSON format, organized into two fields: "high_level_keywords" for overarching ideas and "low_level_keywords" for specific details.

在图 6 中，提示描述了从用户查询中提取关键词的方法，区分高层次和低层次关键词。高层次关键词代表宽泛的概念或主题，低层次关键词着重具体实体和细节。提取的关键词以 JSON 格式返回，组织为两个字段：“high_level_keywords”用于总体概念，“low_level_keywords”用于具体细节。

7.3.4 Prompts for RAG EVALUATION

7.3.4 用于 RAG 评估的提示

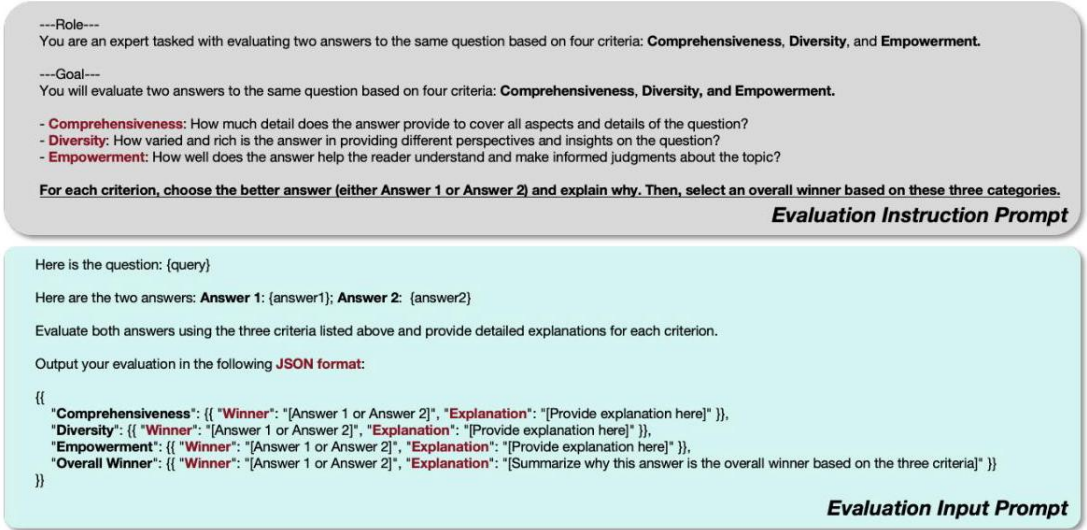


Figure 7: Prompts for RAG Evaluation

图 7:RAG 评估的提示

The evaluation prompt is illustrated in Figure 7. It introduces a comprehensive evaluation framework for comparing two answers to the same question based on three key criteria: Comprehensiveness, Diversity, and Empowerment. Its purpose is to guide the LLM through the process of selecting the better answer for each criterion, followed by an overall assessment. For each of the three criteria, the

评估提示如图 7 所示。它提出了一个用于比较针对同一问题的两个答案的综合评估框架，基于三个关键标准:全面性、多样性和赋能。其目的是引导大型模型按每个标准选择更佳答案，随后给出总体评估。对于每个标准，

Table 5: Case Study: Comparison Between LightRAG and the Baseline NaiveRAG.

表 5: 案例研究:LightRAG 与基线 NaiveRAG 的比较。

Query	
NaiveRAG	How do indigenous perspectives on ownership and collaboration influence corporate mergers in countries like Canada and Australia? Indigenous perspectives on ownership and collaboration play a significant role in shaping corporate mergers in countries like Canada and Australia, particularly where resources and land rights intersect with Indigenous sovereignty and cultural values. These perspectives emphasize community engagement, respect for traditional land use, and a collaborative approach to resource management, influencing corporate strategies and outcomes.
LightRAG	Both entities illustrate that prioritizing Indigenous perspectives leads to more sustainable and equitable outcomes. Acknowledging Indigenous rights and fostering collaborative relationships can reduce conflicts and enhance corporate reputations. Ultimately, integrating Indigenous perspectives is a pathway to better business practices and community relations. Indigenous Perspectives on Ownership and Collaboration in Corporate Mergers Indigenous perspective significantly influence corporate mergers in countries like Canada and Australia. These views emphasize communal rights to land and resources, spiritual connections to the environment, and the importance of collaboration over competition. Cultural Significance of Land Ownership [...] The Role of Collaboration [...] Legal and Regulatory Frameworks [...] Conclusion Indigenous perspectives shape corporate mergers by emphasizing communal ownership and collaborative partnerships. Corporations must adopt respectful, inclusive practices that honor Indigenous cultural values and rights, paving the way for sustainable, mutually beneficial outcomes.
LLM Decision	Comprehensiveness: "Winner": "Answer 2 (LightRAG)" Diversity: "Winner": "Answer 2 (LightRAG)" Empowerment: "Winner": "Answer 2 (LightRAG)" Overall Winner: "Winner": "Answer 2 (LightRAG)" "Explanation": "Answer 2 provides a thorough exploration of the Indigenous perspectives by discussing cultural significance, collaboration, and legal frameworks, including specific examples and detailed insights. In contrast, Answer 1, while informative, lacks the same depth in analyzing the various dimensions of the Indigenous perspectives on ownership and collaboration." "Explanation": "Answer 2 presents a wider array of perspectives, focusing on the communal aspect of land ownership, spiritual connections, and practical examples of collaboration. It covers more ground by contrasting Indigenous views with Western notions, whereas Answer 1 primarily focuses on corporate strategies and outcomes with limited perspectives." "Explanation": "Answer 2 equips the reader with nuanced understanding and actionable insights by highlighting the importance of collaboration and the legal frameworks involved. It empowers corporations by illustrating how adopting an inclusive approach can facilitate better outcomes. Answer 1, while practical, does not emphasize the moral or ethical implications as strongly." "Explanation": "Answer 2 stands out due to its comprehensive exploration, diversity of perspectives, and empowerment of the reader with actionable insights about Indigenous perspectives and collaboration in corporate mergers. Although Answer 1 is more direct, the depth and breadth of Answer 2 make it the stronger response."

查询	
NaiveRAG	在加拿大和澳大利亚等国，原住民关于所有权与协作的观点如何影响公司合并？ 在加拿大和澳大利亚等地，原住民关于所有权与协作的观点在公司合并中具有重要影响，特别是在资源与土地权利与原住民主权和文化价值交汇的情况下。这些观点强调社区参与、尊重传统土地使用以及协同管理资源的方法，从而影响公司战略和结果。 两国都表明，将原住民观点置于优先地位会带来更可持续和更公平的结果。承认原住民权利并促进合作关系可以减少冲突并提升企业声誉。最终，整合原住民观点是通往更好商业实践和社区关系的一条途径。
LightRAG	公司合并中的原住民关于所有权与协作的观点 在加拿大和澳大利亚等国，原住民观点显著影响公司合并。这些观点强调对土地和资源的集体权利、与环境的精神联系以及协作优于竞争的重要性。 土地所有权的文化意义 [...] 协作的作用 [...] 法律与监管框架 [...] 结论 原住民观点通过强调集体所有制和合作伙伴关系来塑造公司合并。公司必须采取尊重且包容的做法，尊重原住民的文化价值和权利，为可持续且互利的结果铺平道路。
LLM Decision	全面性: "获胜者": "答案 2 (LightRAG)", 多样性: "获胜者": "答案 2 (LightRAG)", 赋能: "获胜者": "答案 2 (LightRAG)", 总体获胜者: "获胜者": "答案 2 (LightRAG)" "解释": "答案 2 通过讨论文化意义、协作和法律框架并包含具体示例和详细见解，对原住民观点进行了深入探讨。相比之下，答案 1 虽有信息，但在分析原住民关于所有权与协作的各个维度方面缺乏同等深度。" "解释": "答案 2 提供了更广泛的观点，侧重于土地所有的集体性、与环境的精神联系以及协作优于竞争的重要性。它通过展示原住民观点与西方观念对比覆盖了更多内容，而答案 1 主要集中于公司策略和结果，视角有限。" "解释": "答案 2 通过强调协作和相关法律框架，提供了细致的理解和可行见解，从而赋能读者。它通过展示采用包容性方法如何促进更好结果来帮助企业。答案 1 虽实用，但未能如此强烈地强调道德或伦理含义。" "解释": "答案 2 在全面性、观点多样性及通过可行见解赋能读者方面表现出色。关于原住民观点与公司合并中的协作问题提供了更强的回应。尽管答案 1 更直接，但答案 2 的深度与广度使其成为更强的回复。"

LLM must identify which answer performs better and provide a rationale for its choice. Ultimately, an overall winner is determined based on performance across all three dimensions, accompanied by a detailed summary that justifies the decision. The evaluation is structured in JSON format, ensuring clarity and consistency, and facilitating a systematic comparison between the two answers.

LLM 必须识别哪个答案表现更好并给出理由。最终根据在三项维度上的表现确定总体胜者，并附上说明决策的详细总结。评估以 JSON 格式构建，确保清晰与一致，并便于对两种答案进行系统比较。

7.4 CASE STUDY: COMPARISON BETWEEN LIGHTRAG AND THE BASELINE NAIVERAG.

7.4 案例研究:LightRAG 与基线 NaiveRAG 的比较。

To further illustrate LightRAG's superiority over baseline models in terms of comprehensiveness, empowerment, and diversity, we present a case study comparing LightRAG and NaiveRAG in Table 5. This study addresses a question regarding indigenous perspectives in the context of corporate mergers. Notably, LightRAG offers a more in-depth exploration of key themes related to indigenous perspectives, such as cultural significance, collaboration, and legal frameworks, supported by specific and illustrative examples. In contrast, while NaiveRAG provides informative responses, it lacks the depth needed to thoroughly examine the various dimensions of indigenous ownership and collaboration. The dual-level retrieval process employed by LightRAG enables a more comprehensive investigation of specific entities and their interrelationships, facilitating extensive searches that effectively capture overarching themes and complexities within the topic.

为进一步说明 LightRAG 在全面性、赋能性与多样性方面优于基线模型，我们在表 5 中呈现了 LightRAG 与 NaiveRAG 的比较案例研究。本研究针对公司合并背景下的原住民视角问题。值得注意的是，LightRAG 就与原住民视角相关的关键主题（如文化意义、合作与法律框架）提供了更深入的探讨，并辅以具体且具说明性的示例。相比之下，尽管 NaiveRAG 提供了有信息量的回答，但缺乏深入性，无法彻底审视原住民所有权与合作的各个维度。LightRAG 采用的双层检索流程能够更全面地调查具体实体及其相互关系，促进广泛检索，有效捕捉主题中的总体性主题与复杂性。