
PIKE-RAG: sPecIalized KnowledgE and Rationale Augmented Generation

Jinyu Wang* Jingjing Fu* Rui Wang Lei Song Jiang Bian

Microsoft Research Asia

{jinywan, jifu, ruiwa, lesong, jiabia}@microsoft.com

Abstract

Despite notable advancements in Retrieval-Augmented Generation (RAG) systems that expand large language model (LLM) capabilities through external retrieval, these systems often struggle to meet the complex and diverse needs of real-world industrial applications. The reliance on retrieval alone proves insufficient for extracting deep, domain-specific knowledge performing in logical reasoning from specialized corpora. To address this, we introduce sPecIalized KnowledgE and Rationale Augmentation Generation (PIKE-RAG), focusing on extracting, understanding, and applying specialized knowledge, while constructing coherent rationale to incrementally steer LLMs toward accurate responses. Recognizing the diverse challenges of industrial tasks, we introduce a new paradigm that classifies tasks based on their complexity in knowledge extraction and application, allowing for a systematic evaluation of RAG systems' problem-solving capabilities. This strategic approach offers a roadmap for the phased development and enhancement of RAG systems, tailored to meet the evolving demands of industrial applications. Additionally, we propose knowledge atomizing and knowledge-aware task decomposition to effectively extract multifaceted knowledge from the data chunks and iteratively construct the rationale based on original query and the accumulated knowledge, respectively, showcasing exceptional performance across various benchmarks. Furthermore, we introduce a trainable knowledge-aware decomposer that incorporates domain-specific rationale into the task decomposition and result-seeking process. The code is publicly available at <https://github.com/microsoft/PIKE-RAG>.

1 Introduction

Large Language Models (LLMs) have revolutionized the field of natural language processing by demonstrating the capability to generate coherent and contextually relevant text. These advanced models are trained on expansive corpora, equipping them with the versatility to execute a diverse spectrum of linguistic tasks, ranging from text completion to translation and summarization [5, 10, 53, 7]. Despite their broad capabilities, LLMs exhibit pronounced limitations when tasked with specialized queries in professional domains [39, 58], a demand that is particularly acute in industrial applications. This primarily stems from the scarcity of domain-specific training material and a limited grasp of specialized knowledge and rationale within these domains. As a result, LLMs may produce responses that are not only potentially erroneous but also lack the detail and precision required for expert-level engagement [12]. Besides the limitations in the domain-specific tasks, another striking issue with LLMs is the phenomena known as "hallucination", where the model generates information that is not grounded in reality or factual data [11, 61]. Moreover, the knowledge base of LLMs, being static and crystallized at the point of their last update, introduces temporal stasis [14]. Further compounding these challenges is the issue of long-context comprehension [38]. Existing LLMs

*Equal Contribution.

struggle to maintain an understanding of task definitions across long context, and their performance tends to deteriorate significantly when confronted with more complex and demanding tasks.

To address the inherent limitations of LLMs, Retrieval-Augmented Generation (RAG) [36] has been proposed, which merges the generative capabilities of LLMs with a retrieval mechanism, allowing the incorporation of relevant external information to anchor the generated text in factual data. This integrated strategy improves both the accuracy and reliability of the generated content, providing a promising pathway for the practical deployment of LLMs in industrial applications. However, current RAG methods remain heavily reliant on text retrieval and the comprehension capabilities of LLMs, with a lack of attention to extracting, understanding, and utilizing knowledge from the diverse source data. In industrial applications requiring expertise, such as specialized knowledge and problem-solving rationale, existing RAG approaches primarily designed for research benchmarks demonstrate significant limitations. There is a lack of clarity regarding the challenges that RAG encounters in industrial applications. Gaining a comprehensive insight into these challenges is crucial for the development of RAG algorithms. Therefore, we summarize the main challenges as follows.

- **Knowledge source diversity:** RAG systems are constructed upon a diverse corpus of source documents collected over many years from various domains, encompassing a wide range of file formats like scanned images, digital text files, and web data, sometimes accompanied by specialized databases. In contrast, widely-used datasets [29, 64, 54] typically feature pre-segmented, simplified corpora that do not capture the complexity of real-world data. Existing methods designed for these benchmarks struggle to efficiently extract specialized knowledge and uncover underlying rationales from diverse sources, particularly in industrial applications. *For example*, an LED product datasheet typically comprises specifications such as performance characteristics presented in complex tables, electrical properties depicted in charts, and installation instructions illustrated with figures. Addressing queries related to the non-textual knowledge presents significant challenges for existing RAG approaches.
- **Domain specialization deficit:** In industrial applications, RAG are expected to leverage the specialized knowledge and rationale in professional fields. However, these specialized knowledge are characterized by domain-specific terminologies, expertise, and distinctive logical frameworks that are integral to their functioning. RAG approaches built on common knowledge-centric datasets demonstrate unsatisfactory performance when applied to professional fields, as LLMs exhibit deficiencies in extracting, understanding, and organizing domain specific knowledge and rationale [39]. *For example*, in the field of semiconductor design, research relies heavily on a deep understanding of underlying physical properties. When LLMs are utilized to extract and organize the specialized knowledge and rationale from the research documents, they often fail to properly capture essential physical principles and achieve a comprehensive understanding due to their inherent limitations. Consequently, RAG systems frequently produce incomplete or inaccurate interpretations of critical problem elements and generate responses that lack proper rationale grounded in physical principles. Moreover, assessing the quality of professional content generation poses a significant challenge. This issue not only impedes the development and optimization of RAG algorithms but also complicates their practical deployment across various industrial applications.
- **One-size-fits-all:** Various RAG application scenarios, although based on a similar framework, present different challenges that require diverse capabilities, particularly for extracting, understanding, and organizing domain-specific knowledge and rationale. The complexity and focus of questions vary across these scenarios, and within a single scenario, the difficulty can also differ. *For example*, in rule-based query scenarios, such as determining the legal conditions for mailing items, RAG systems primarily focus on retrieving relevant factual rules by bridging the semantic gap between the query and the rules. In multihop query scenarios, such as comparing products across multiple aspects, RAG systems emphasize extracting information from diverse sources and performing multihop reasoning to arrive at accurate answers. Most existing RAG approaches [66] adopt a one-size-fits-all strategy, failing to account for the varying complexities and specific demands both within and across scenarios. This results in solutions that do not meet the comprehensive accuracy standards required for practical applications, thereby limiting the development and integration of RAG systems in real-world environments.

We believe that the key to addressing these challenges lies in advancing beyond traditional retrieval augmentation, by effectively extracting, understanding, and applying specialized knowledge, and

developing appropriate reasoning logic tailored to the specific tasks and the knowledge involved. We refer to this approach as *sPecialized Knowledge and Rationale Augmentation*. Given that various tasks require diverse capabilities, particularly for extracting, understanding, and organizing domain-specific knowledge and rationale, we summarize and categorize the questions commonly encountered into four types with respect to their difficulty: factual questions, linkable-reasoning questions, predictive questions, and creative questions. Accordingly, we propose a classification of RAG system capability levels, aligned with the system’s ability to solve these different types of problems. This classification serves as a guideline for systematically advancing the system’s capabilities in a controllable and measurable manner.

Furthermore, we propose **sPecialized Knowledge and Rationale Augmented Generation (PIKE-RAG)** framework, which not only support phased system development and deployment, demonstrating excellent versatility, but also enhances capabilities by effectively leveraging specialized knowledge and rationale. Within this framework, knowledge extraction components are employed to extract specialized knowledge from diverse source data, laying a robust foundation for knowledge-based retrieval and reasoning. Additionally, a task decomposer is utilized to dynamically manage the routing of retrieval and reasoning operations, creating specialized rationale based on available knowledge. PIKE-RAG enables a phased exploration of RAG capabilities, which facilitates the progressive refinement of RAG algorithms and the staged implementation of RAG applications. For each developing phase, the RAG framework and its modules are tailored to address specific challenges. For example, in the knowledge base construction phase, a multi-layer heterogeneous graph is employed to effectively represent relationship between various components of the data, enhancing knowledge organization and integration. The RAG system, designed for factual questions, introduces multi-granularity retrieval, allowing for multi-layer, multi-granularity retrieval across a heterogeneous knowledge graph to improve factual retrieval accuracy. In the advanced RAG system, aiming at addressing complex queries, knowledge atomizing is introduced to fully explore the intrinsic knowledge from data chunks, while knowledge-aware task decomposition manages the retrieval and organization of multiple pieces of atomic knowledge to construct a coherent rationale.

Extensive experiments are conducted to evaluate the performance of the proposed PIKE-RAG framework on both open-domain and legal benchmarks, and experimental results demonstrate the effectiveness of PIKE-RAG. Our framework and staged development strategy could further advance the current research and application of RAG in industrial contexts. In summary, the contributions of this work are as follows:

- We propose that specialized knowledge and rationale should serve as the core foundation for augmentation, empowering the resolution of tasks that current retrieval-augmented frameworks are unable to solve effectively. Therefore, we introduce a new paradigm that classifies tasks to distinct types based on their difficulty in the knowledge extraction, comprehension and utilization, offering a novel conceptual framework for system design and evaluation. By applying this paradigm, RAG systems’ capabilities are stratified to support phased development, particularly enhancing their application in industrial settings.
- We introduce specialized Knowledge and Rationale Augmented Generation (PIKE-RAG) framework, which is primarily designed with a focus on specialized knowledge extraction and rationale construction. PIKE-RAG enhances the system capabilities by effectively extracting, comprehending, and organizing specialized knowledge and rationale. Additionally, it can customize the system framework to meet varying levels of capability requirements, demonstrating exceptional versatility.
- We propose knowledge atomizing and knowledge-aware task decomposition to tackle the complex questions, such as multihop queries, achieving significant performance improvements on the multihop benchmarks, particularly in scenarios involving more than two hops. This demonstrates that task decomposition effectively breaks down complex questions into atomic questions, enabling efficient retrieval and organization of atomic knowledge and constructing a coherent rationale to arrive at accurate answers.
- We introduce a knowledge-aware task decomposer training strategy that involves collecting rationale-driven data by sampling context and creating diverse interaction. This approach trains the decomposer to incorporate domain-specific rationale into both task decomposition and result-seeking trajectories.

2 Related work

2.1 RAG

Retrieval-Augmented Generation (RAG) has emerged as a promising solution that effectively incorporates external knowledge to enhance response generation. Initially, retrieval-augmented techniques were introduced to improve the performance of pre-trained language models on knowledge-intensive tasks [36, 30, 13]. With the booming of Large Language Models [5, 10, 53, 7], most research in the RAG paradigm has shifted towards a framework that initially retrieves pertinent information from external data sources and subsequently integrates it into the context of the query prompt as supplementing knowledge for contextually relevant generation [47]. Following this framework, naive RAG research paradigm [26] converts raw data into uniform plain text and segment it into smaller chunks, which are encoded into vector space for query-based retrieval. The top k relevant chunks are used to expand the context of the prompt for generation. To enhance the retrieval quality of the naive RAG, advanced RAG approaches implement specific enhancements across the pre-retrieval, retrieval, and post-retrieval processes, including query optimization [40, 67], multi-granularity chunking [17, 69], mixed retrieval and chunk re-ranking.

Beyond the aforementioned RAG paradigms, numerous sophisticated enhancements in RAG pipelines and system modules are introduced within modular RAG systems [27], aiming to improve system capability and versatility. These advancements have enabled the processing of a wider variety of source data, facilitating the transformation of raw information into structured data and, ultimately, into valuable knowledge [60, 21]. Furthermore, the indexing and retrieval modules have been refined with multi-granularity and multi-architecture approaches [62, 69]. Various pre-retrieval [25, 68] and post-retrieval [19, 31] functions are proposed to enhance both the retrieval effectiveness and the quality of sequential generation. It has been recognized that naïve RAG systems are insufficient to tackle complex tasks such as summarization [28] and multi-hop reasoning [54, 29]. Consequently, most recent research focuses on developing advanced coordination schemes that leverage existing modules to collaboratively address these challenges. ITERETGEN [49] and DSP [34] employ retrieve-read iteration to leverage generation response as the context for next round retrieval. FLARE [32] proposes a confidence-based active retrieval mechanism that dynamically adjusts query with respect to the low-confidence tokens in the regenerated sentences. These loop-based RAG pipelines progressively converge towards the correct answer and provide enhanced flexibility to RAG systems in addressing diverse requirements.

2.2 Knowledge bases for RAG

In naïve RAG approaches, source data is converted to plain text and chunked for retrieval. However, as RAG applications expand and demand for diversity grows, plain text-based retrieval becomes insufficient for several reasons: (1) textual information is generally redundant and noisy, leading to decreased retrieval quality; (2) complex problems require the integration of multiple data sources, and plain text alone cannot adequately represent the intricate relationships between objects. As a result, researchers are exploring diverse data sources to enrich the corpus, incorporating search engines [63, 57], databases [59, 42, 48], knowledge graphs [50, 60], and multimodal corpora [18, 16]. Concurrently, there is an emphasis on developing efficient knowledge representations for corpus to enhance knowledge retrieval. A graph is regarded as a powerful knowledge representation because of its capacity to intuitively model complex relationships. GraphRAG [21] combines knowledge graph generation and query-focused summarization with RAG to address both local and global questions. HOLMES [43] construct hyper-relational KGs and prune them to distilled graphs, which serve as an input to LLMs for multihop question answering. However, the construction of knowledge graphs is extremely resource-intensive, and the associated costs scale up with the size of the corpus.

2.3 Multi-hop QA

Multi-hop Question Answering (MHQA) [64] involves answering questions that require reasoning over multiple pieces of information, often scattered across different documents or paragraphs. This task presents unique challenges as it necessitates not only retrieving relevant information but also effectively combining and reasoning over the retrieved pieces to arrive at a correct answer. The traditional graph-based methods in MHQA solve the problem by building graphs and inferring on graph neural networks(GNN) to predict answers [45, 22]. With the advent of LLMs, recent

graph-based methods [37, 43] have evolved to construct knowledge graphs for retrieval and generate response through LLMs. Another branch of methods dynamically convert multi-hop questions into a series of sub-queries by generating subsequent questions based on the answers to previous ones [55, 34, 24]. The subqueries guides the sequential retrieval and the retrieved results in turn are used to improve reasoning. Treating MHQA as a supervised problem, Self-RAG [65] trains an LM to learn to retrieve, generate, and critique text passages, and beam-retrieval [8] models the multi-hop retrieval process in an end-to-end manner by jointly optimizing an encoder and classification heads across all hops. Self-Ask [44] improves CoT by explicitly asking itself follow-up questions before answering the initial question. This method enables the automatic decomposition of questions and can be seamlessly integrated with retrieval mechanisms to tackle Multi-hop Question Answering.

3 Problem formulation

Existing research mainly concentrates on algorithmic enhancements to improve the performance of RAG systems. However, there is limited effort in providing a comprehensive and systematic discussion of the RAG framework. In this work, we conceptualize the RAG framework from three key perspectives: knowledge base, task classification, and system development. We assert that the knowledge base serves as the fundamental cornerstone of RAG, underpinning all retrieval and generation processes. Furthermore, we recognize that RAG tasks can vary significantly in complexity and difficulty, depending on the required generation capabilities and the availability of supporting corpora. By categorizing tasks according to their difficulty levels, we classify RAG systems into distinct levels based on their problem-solving capabilities across the different types of questions.

3.1 Knowledge base

In industrial applications, specialized knowledge primarily originates from years of accumulated data within specific fields such as manufacturing, energy, and logistics. For example, in the pharmaceutical industry, data sources include extensive research and development documentation, as well as drug application files amassed over many years. These sources are not only diverse in file formats, but also encompass a significant amount of multi-modal contents such as tables, charts, and figures, which are also crucial for problem-solving. Furthermore, there are often functional connections between files within a specialized domain, such as hyperlinks, references, and relational database links, which explicitly or implicitly reflect the logical organization of knowledge within the professional field. Currently, existing datasets provide pre-segmented corpora and do not account for the complexities encountered in real-world applications, such as the integration of multi-format data and the maintenance of referential relationships between documents. Therefore, the construction of a comprehensive knowledge base is foundational for Retrieval-Augmented Generation (RAG) in the industrial field. As the architecture and quality of the knowledge base directly influence the retrieval methods and their performance, we propose structuring the knowledge base as a multi-layer heterogeneous graph, denoted as G , with corresponding nodes and edges represented by (V, E) . The graph nodes can include documents, sections, chunks, figures, tables, and customized nodes from distilled knowledge. The edges signify the relationships among these nodes, encapsulating the interconnections and dependencies within the graph. This multi-layer heterogeneous graph encompasses three distinct layers: the information resource layer G_i , the corpus layer G_c and the distilled knowledge layer G_{dk} . Each layer corresponds to different stages of information processing, representing varying levels of granularity and abstraction in knowledge.

3.2 Task classification

Contemporary RAG frameworks frequently overlook the intricate difficulty and logistical demands inherent to diverse tasks, typically employing a one-size-fits-all methodology. However, even with comprehensive knowledge retrieval, current RAG systems are insufficient to handle tasks of varying difficulty with equal effectiveness. Therefore, it is essential to categorize tasks and analyze the typical strategies for overcoming the challenges inherent to each category. The difficulty of a task is closely associated with several critical factors.

- *Relevance and Completeness of Knowledge:* The extent to which the necessary information is present within the knowledge base and how comprehensively it covers the topic.

Factual Questions

<p>Q: Is Abrilada an interchangeable biosimilar product?</p> <p>S: (link) “the prefilled syringe presentation of Abrilada (adalimumab-afzb) injection for subcutaneous use, 10 mg/0.2 mL and 20 mg/0.4 mL, as interchangeable with the corresponding presentation of US-licensed Humira (adalimumab) injection for subcutaneous use, 10 mg/0.2 mL and 20 mg/0.4 mL respectively.”</p>	<p>Q: How many biosimilar products have been approved by the FDA? (Q1)</p> <p>S: (link) </p>
---	--

Linkable & Reasoning Questions

Bridging Questions	Quantitative Questions
<p>Q: What are the BLA IDs of the biosimilar product applications for HUMIRA that have been successfully approved?</p> <p>S: (analytical reasoning) Decompose the task. (link) Retrieve biosimilar products with HUMIRA as reference. (link) Retrieve BLA ID of these products.</p>	<p>Q: How many interchangeable biosimilar products have been approved by the FDA? (Q2)</p> <p>S: (analytical reasoning) Decompose the task. (link) Retrieve the interchangeable biosimilar products. (statistics) Count the total number of the products.</p>
Comparative Questions	Summarizing Questions
<p>Q: Please provide the product name of the latest biosimilar product for HUMIRA that was successfully approved.</p> <p>S: (analytical reasoning) Decompose the task. (link) Retrieve biosimilar products with HUMIRA as reference and their approval dates. (compare) Compare approval dates and get the product name.</p>	<p>Q: Please analyze the labels and see if you can find major differences in the clinical study approaches between Idacio, Cyltezo, and Amjevitा.</p> <p>S: (analytical reasoning) Decompose the task. (link) Retrieve the contents related to clinical study in label files. (organize) Organize the information into categories. (compare) Compare on clinical study contents of each category. (summarize) Summarize the major differences</p>

Predictive Questions

Creative Questions

<p>Q: Please provide an approximate prediction: How many biosimilar products are expected to be approved in 2024? (Q3)</p> <p>S: (analytical reasoning) Decompose the task. (link) Retrieve the approval date of each biosimilar product. (statistics) Count the total number of approval for each year. (organize) Collect total numbers over years. (analytical reasoning) Analyze the variation trend and range of total numbers and conduct prediction.</p>	<p>Q: Assuming a biosimilar candidate with Humira as the reference, how can one prepare for the BLA to minimize the application time?</p> <p>S: (analytical reasoning) Decompose the task. (link) Retrieve BLA materials of the biosimilar products with Humira as reference. (organize) Extract application stages with their respective durations. (analytical reasoning) Identify and distill the common patterns, rules, and influential factors affecting the time consumption of these stages. (plan) Develop solutions based the time cost evalution.</p>
---	--



Answers **explicitly** present in corpus



Answers **implicitly** exist in corpus



Answers **not** present in corpus

Figure 1: Illustrative examples of distinct question types

- *Complexity of Knowledge Extraction*: The difficulty in accurately identifying and retrieving all relevant pieces of knowledge, especially when scattered across multiple sources or implicitly embedded in the text.
- *Depth of Understanding and Reasoning*: The level of cognitive and inferential processing required to comprehend the retrieved information, establish connections, and perform multi-step reasoning.
- *Effectiveness of Knowledge Utilization*: The sophistication involved in applying the extracted knowledge to formulate responses, including synthesizing, organizing, and generating insights or predictions.

In categorizing real-world RAG tasks within industries, we focus on the processes of knowledge extraction, understanding, organization, and utilization to provide structured and insightful responses. Taking the aforementioned factors into account, we identify four distinct classes of questions that address a broad spectrum of demands. The first type, *Factual Questions*, involves extracting specific, explicit information directly from the corpus, relying on retrieval mechanisms to identify the relevant facts. *Linkable-Reasoning Questions* demand a deeper level of knowledge integration, often requiring multi-step reasoning and linking across multiple sources. *Predictive Questions* extend beyond the available data, requiring inductive reasoning and structuring of retrieved facts into analyzable forms,

such as time series, for future-oriented predictions. Finally, *Creative Questions* engage domain-specific logic and creative problem-solving, encouraging the generation of innovative solutions by synthesizing knowledge and identifying patterns or influencing factors. This categorization, driven by varying levels of reasoning and knowledge management, ensures a comprehensive approach to addressing industry-specific queries.

The criteria defining each category are elaborated in the following sections, with representative examples for each provided in Figure 1. For each question type, we also present the associated support data and the expected reasoning processes to illustrate the differences between these categories. These inquiries are formulated by experts in pharmaceutical applications, based on the data released by the FDA.²

Factual Questions These questions seek specific, concrete pieces of information explicitly presented in the original corpus. The referenced text can be processed within the context of a conversation in LLMs. As shown in Figure 1, this class of questions can be effectively answered if the relevant fact is successfully retrieved.

Linkable-Reasoning Questions Answering these questions necessitates gathering pertinent information from diverse sources and/or executing multi-step reasoning. The answers may be implicitly distributed across multiple texts. Due to variations in the linking and reasoning processes, we further divide this category into four subcategories: bridging questions, comparative questions, quantitative questions, and summarizing questions. Examples of each subcategory are illustrated in Figure 1. Specifically, bridging questions involve sequentially bridging multiple entities to derive the answer. Quantitative questions require statistical analysis based on the retrieved data. Comparative questions focus on comparing specified attributes of two entities. Summarizing questions require condensing or synthesizing information from multiple sources or large volumes of text into a concise, coherent summary, and they often involve integrating key points, identifying main themes, or drawing conclusions based on the aggregated content. Summarizing questions may combine elements of other question types, such as bridging, comparative, or quantitative questions, as they frequently require the extraction and integration of diverse pieces of information to generate a comprehensive and meaningful summary. Given these questions require multi-step retrieval and reasoning, it is crucial to establish a reasonable operation route for answer-seeking in interaction with the knowledge base.

Predictive Questions For this type of questions, the answers are not directly available in the original text and may not be purely factual, necessitating inductive reasoning and prediction based on existing facts. To harness the predictive capabilities of LLMs or other external prediction tools, it is essential to gather and organize relevant knowledge to generate structured data for further analysis. For instance, as illustrated in Figure 1, all biosimilar products with the approval dates are retrieved, and the total number of approvals for each year is calculated and organized to year-indexed time series data for prediction purposes. Furthermore, it is important to note that the correct answer to predictive questions may not be unique, reflecting the inherent uncertainty and variability in predictive tasks.

Creative Questions One significant demand of RAG is to mine valuable domain-specific logic from professional knowledge bases and introduce novel perspectives that can innovate and advance existing solutions. Addressing creative questions necessitates creative thinking based on the availability of factual information and an understanding of the underlying principles and rules. As illustrated in the example, it is essential to organize the extracted information to highlight key stages and their duration, and then identify common patterns and influential factors. Subsequently, solutions are developed with the objective of evaluating potential outcomes and stimulating fresh ideas. The goal of these responses is to inspire experts to generate innovative ideas, rather than to provide ready-to-implement solutions.

It is crucial to recognize that the classification of a question may shift with changes in the knowledge base. Questions Q1, Q2, and Q3 in Figure 1, although seemingly similar, are categorized differently depending on the availability of information and the logical steps required to derive an answer. For instance, Q1 is classified as a factual question because it can be directly answered using a table that concisely lists all biosimilar products along with their respective approval dates, providing sufficient explicit information. In contrast, Q2, which inquires about the total count of interchangeable

²<https://www.fda.gov/>

Table 1: Level definition based on RAG system’s capability

Level	System capability description
L1	The L1 system is designed to provide accurate and reliable answers to factual questions , ensuring a solid foundation for basic information retrieval.
L2	The L2 system extends its functionality to include accurate and reliable responses to both factual questions and linkable-reasoning questions , enabling more complex multi-step retrieval and reasoning tasks.
L3	The L3 system further enhances its capabilities by incorporating the ability to deliver reasonable predictions for predictive questions , while maintaining accuracy and reliability in answering both factual questions and linkable-reasoning questions .
L4	The L4 system is capable of proposing well-reasoned plans or solutions to creative questions . In addition, it retains the ability to provide reasonable predictions for predictive questions , alongside accurate and reliable answers to factual questions and linkable-reasoning questions .

biosimilar products, cannot be resolved by directly referencing a single explicit source. To answer Q2, one must identify all the products meeting the specified criteria and subsequently calculate the total, necessitating an additional step of statistical aggregation. Therefore, Q2 is categorized as a linkable-reasoning question due to the need for an intermediate processing. Finally, Q3 poses a challenge because the answer does not explicitly exist within the knowledge base. Addressing Q3 requires gathering relevant data, organizing it to infer hidden patterns, and making predictions based on these inferred rules. As a result, Q3 is categorized as a predictive question, indicating the requirement to extrapolate beyond the existing data to forecast potential outcomes or trends.

3.3 RAG system level

In industrial RAG systems, inquiries encompass a broad spectrum of difficulties and are approached from diverse perspectives. Although RAG systems can leverage the general question-answering(QA) abilities of LLMs, their limited comprehension of expert-level knowledge often leads to inconsistent response quality across questions of varying complexities. In response to this status quo, we propose categorizing RAG systems into four distinct levels based on their problem-solving capabilities across the four classes of questions outlined in the previous subsection. This stratified approach facilitates the phased development of RAG systems, allowing capabilities to be incrementally enhanced through iterative module refinement and algorithmic optimization. Our framework is strategically designed to provide a standardized, objective methodology for developing RAG systems that effectively meet the specialized needs of various industry scenarios. The definition of RAG systems in different level is presented in Table 1. It highlights the systems’ capabilities to handle increasingly complex queries, demonstrating the evolution from simple information retrieval to advanced predictive and creative problem-solving. Each level represents a step towards more sophisticated interactions with knowledge bases, requiring the RAG systems to demonstrate higher levels of understanding, reasoning, and innovation.

More specifically, at the foundational level, RAG systems respond to factual questions with answers that are directly extractable from provided texts. Advancing to the second level, RAG systems are equipped to handle complex questions involving linkage and reasoning. These queries necessitate the synthesis of information from disparate sources or multi-step reasoning processes. The RAG could address a variety of composite questions, includes bridging questions that necessitate a sequence of logical reasoning, comparative questions demanding parallel analysis, and summarizing questions that involve condensing information into comprehensive responses. At the third level, the systems are intricately designed to tackle predictive questions where answers are not immediately discernible from the original text. Finally, RAG systems at the forth level demonstrate the capacity for creative problem-solving, utilizing a solid factual base to foster novel concepts or strategies. While these systems may not offer ready-to-implement solutions, they play a crucial role in stimulating expert creativity to advance fields such as analytics or treatment design.

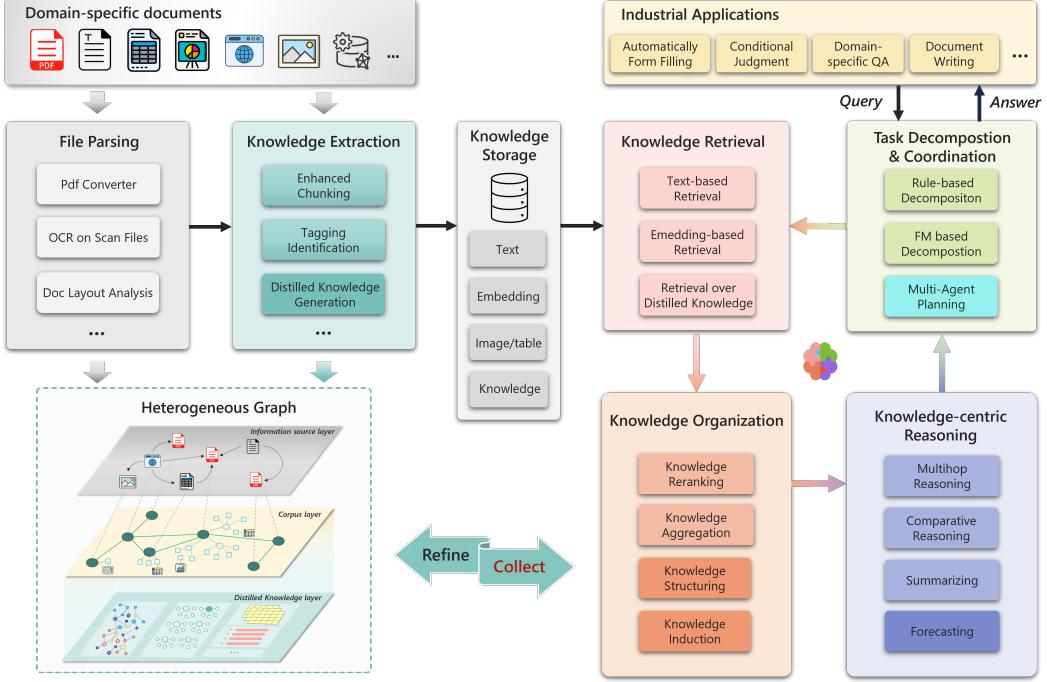


Figure 2: Overview of the PIKE-RAG framework, comprising several key components: file parsing, knowledge extraction, knowledge storage, knowledge retrieval, knowledge organization, task decomposition and coordination, and knowledge-centric reasoning. Each component can be tailored to meet the evolving demands of system capability.

4 Methodology

4.1 Framework

Based on the formulation of RAG systems in terms of knowledge base, task classification, and system-level division, we propose a versatile and expandable RAG framework. Within this framework, the progression in levels of RAG systems can be achieved by adjusting submodules within the main modules. The overview of our framework is depicted in Figure 2. The framework primarily consists of several fundamental modules, including file parsing, knowledge extraction, knowledge storage, knowledge retrieval, knowledge organization, knowledge-centric reasoning, and task decomposition and coordination. In this framework, domain-specific documents of diverse formats are processed by file parsing module to convert the file to machine-readable formats, and file units are generated to build up graph in information source layer. The knowledge extraction module chunks the text and generates corpus and knowledge units to construct graph in corpus layer and distilled knowledge layer. The heterogeneous graph established is utilized as the knowledge base for retrieval. Extracted knowledge is stored in multiple structured formats, and the knowledge retrieval module employs hybrid retrieval strategy to access relevant information. Note that the knowledge base not only serves as the source of knowledge gathering but also benefits from a feedback loop, where the organized and verified knowledge is regarded as feedback to refine and improve the knowledge base.

As highlighted in the task classification examples, questions of different classes require distinct rationale routing for answer-seeking, influenced by multiple factors such as the availability of relevant information, the complexity of knowledge extraction, and the sophistication of reasoning. It is challenging to address these questions in a single retrieval and generation pass. To tackle this, we propose an iterative retrieval-generation mechanism supervised by task decomposition and coordination. This iterative mechanism enables the gradual collection of relevant information and progressive reasoning over incremental context, ensuring a more accurate and comprehensive response. More specifically, the questions in industrial applications are fed into task decomposition module to produce preliminary decomposition scheme. This scheme outlines the retrieval steps, reasoning steps, and other necessary operations. Following these instructions, the knowledge retrieval

Table 2: Proposed frameworks for different system levels. To address the challenges facing at each level, we propose customized frameworks based on the framework illustrated in Figure 2. The following abbreviations are used: "PA" for file parsing, "KE" for knowledge extraction, "RT" for knowledge retrieval, "KO" for knowledge organization, and "KR" for knowledge-centric reasoning.

Level	Challenges	Proposed Framework
L0	<ul style="list-style-type: none"> Challenges arise in knowledge extraction due to the diverse formats of source documents, requiring sophisticated file parsing techniques. The construction of a high-quality knowledge base from raw, heterogeneous data introduces significant complexity in knowledge organization and integration. 	
L1	<ul style="list-style-type: none"> The understanding and extraction of knowledge are hindered by improper chunking, which disrupts semantic coherence, complicating accurate retrieval. Knowledge retrieval is impacted by the limitations of embedding models in aligning professional terminologies and aliases, reducing the system's precision. 	
L2	<ul style="list-style-type: none"> Effective knowledge extraction and utilization are critical, as chunked text often contains both relevant and irrelevant information. Ensuring the retrieval of high-quality data is essential for accurate generation. The understanding and decomposition of tasks and rationale behind them often overlook the availability of supporting data, relying heavily on LLM capabilities. 	
L3	<ul style="list-style-type: none"> The challenges at this level center on knowledge collection and organization, which are vital for supporting predictive reasoning. LLMs have limitations in applying specialized reasoning logic, restricting their effectiveness in predictive tasks. 	
L4	<ul style="list-style-type: none"> The difficulty lies in extracting coherent logical rationales from complex knowledge bases, where interdependencies between multiple factors can result in non-unique solutions. The open-ended nature of creative questions complicates the evaluation of the reasoning and knowledge synthesis process, making it difficult to quantitatively assess answer quality. 	

module retrieves relevant information, which is then passed to the knowledge organization module for processing and organization. The organized knowledge is used to perform knowledge-centric reasoning, yielding an intermediate answer. With the updated relevant information and intermediate answer, the task decomposition module regenerates an updated scheme for the next iteration. This design boasts excellent adaptability, allowing us to tackle problems of varying difficulties and perspectives by adjusting the modules and iterative mechanisms.

4.2 Phased system development

We have categorized RAG systems into four distinct levels based on their problem-solving capabilities across the four classes of questions, as outlined in Table 1. Recognizing the pivotal role of knowledge base generation in RAG systems, we designate the construction of the knowledge base as the L0 stage of system development. The challenges faced by RAG systems vary across different levels. We analyze these challenges for each level and propose corresponding frameworks in Table 2. This stratified approach facilitates the phased development of RAG systems, enabling incremental enhancement of capabilities through iterative module refinement and algorithmic optimization.

We observe that from L0 to L4, higher-level systems can inherit modules from lower levels and add new modules to enhance system capabilities. For instance, compared to an L1 system, an L2 system not only introduces a task decomposition and coordination module to leverage iterative retrieval-generation routing but also incorporates more advanced knowledge extraction modules, such as distilled knowledge generation, indicated in dark green in Figure 2. In the L3 system, the growing emphasis on predictive questioning necessitates enhanced requirements for knowledge organization and reasoning. Consequently, the knowledge organization module introduces additional submodules for knowledge structuring and knowledge induction, indicated in dark orange. Similarly, the knowledge-centric reasoning module has been expanded to include a forecasting submodule,

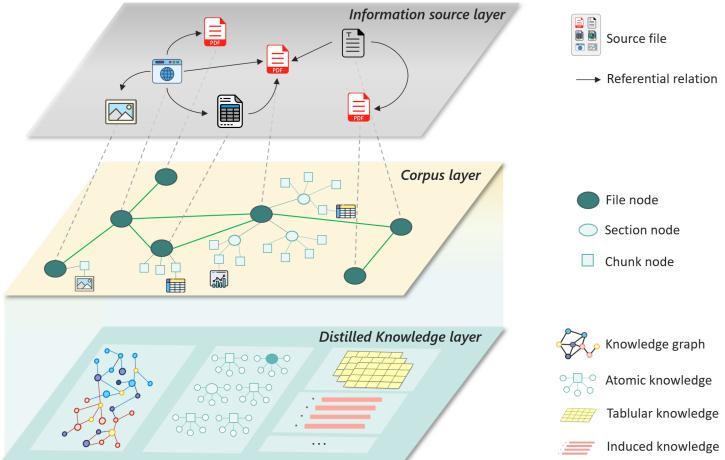


Figure 3: Multi-layer heterogeneous graph as the knowledge base. The graph comprises three distinct layers: information resource layer, corpus layer and distilled knowledge layer.

highlighted in dark purple. In the L4 system, extracting complex rationale from an established knowledge base is highly challenging. To address this, we introduce multi-agent planning module to activate reasoning from diverse perspectives.

5 Detailed Implementation

In this section, we delve into the implementation specifics of each module within our proposed versatile and expandable RAG framework. By elucidating the details at each level, we aim to provide a comprehensive understanding of how the framework operates and how its modularity and expandability are achieved. The subsections that follow will cover the file parsing, knowledge extraction, knowledge storage, knowledge-centric reasoning, and task decomposition and coordination modules, providing insights into their individual functionalities and interactions.

5.1 Level-0: Knowledge Base Construction

The foundational stage of the proposed RAG systems is designated as the L0 system, focuses on the construction of a robust and comprehensive knowledge base. This stage is critical for enabling effective knowledge retrieval in subsequent levels. The primary objective of the L0 system is to process and structure domain-specific documents, transforming them into a machine-readable format and organizing the extracted knowledge into a heterogeneous graph. This graph serves as the backbone for all higher-level reasoning and retrieval tasks. The L0 system encompasses several key modules: file parsing, knowledge extraction, and knowledge storage. Each of these modules plays a crucial role in ensuring that the knowledge base is both extensive and accurately reflects the underlying information contained within the source documents.

5.1.1 File parsing

The ability to effectively parse and read various types of files is a critical component in the development of RAG systems that rely on diverse data sources. Frameworks such as LangChain³ provide a comprehensive suite of tools for natural language processing (NLP), including modules for parsing and extracting information from unstructured text documents. Its file reader capabilities are designed to handle a wide range of file formats, ensuring that data from heterogeneous sources can be seamlessly integrated into the system. Additionally, several deep learning-based tools [2, 3] and commercial cloud APIs [1, 4] have been developed to conduct robust Optical Character Recognition (OCR) and accurate table extraction, enabling the conversion of scanned documents and images into structured, machine-readable text. Given that domain-specific files often encompass sophisticated

³<https://www.langchain.com>

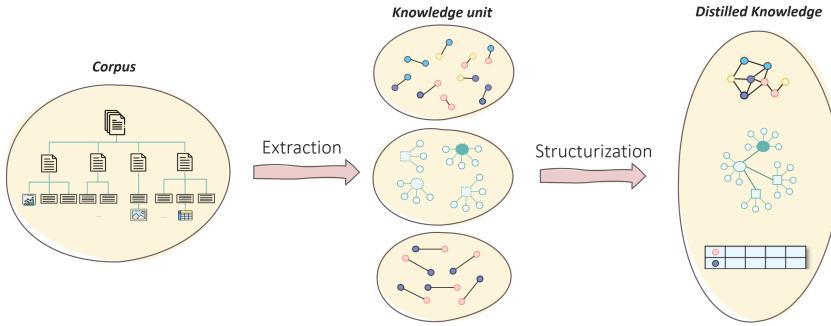


Figure 4: The process of distilling knowledge from corpus text. The corpus text are processed to extract knowledge units following customized extraction patterns. These knowledge units are then organized to structured knowledge in the distilled knowledge layer, which may take the form of knowledge graphs, atomic knowledge, tabular knowledge, and other induced knowledge.

tables, charts, and figures, text-based conversion may lead to information loss and disrupt the inherent logical structure. Therefore, we propose conducting layout analysis for these files and preserving multi-modal elements such as charts and figures. The layout information can aid the chunking operation, maintaining the completeness of chunked text, while figures and charts can be described by Vision-Language Models (VLMs) to assist in knowledge retrieval. This approach ensures that the integrity and richness of the original documents are retained, enhancing the efficacy of RAG systems.

5.1.2 Knowledge Organization

The proposed knowledge base is structured as a multi-layer heterogeneous graph, representing different levels of information granularity and abstraction. The graph captures relationships between various components of the data (e.g., documents, sections, chunks, figures, and tables) and organizes them into nodes and edges, reflecting their interconnections and dependencies. As depicted in Figure 3, this multi-layer structure, encompassing the information resource layer, corpus layer, and distilled knowledge layer, enables both semantic understanding and rationale-based retrieval for downstream tasks.

Information Resource Layer: This layer captures the diverse information sources, treating them as source nodes with edges that denote referential relationships among them. This structure aids in cross-referencing and contextualizing the knowledge, establishing a foundation for reasoning that depends on multiple sources.

Corpus Layer: This layer organizes the parsed information into sections and chunks while preserving the document’s original hierarchical structure. Multi-modal content such as tables and figures is summarized by LLMs and integrated as chunk nodes, ensuring that multi-modal knowledge is available for retrieval. This layer enables knowledge extraction with varying levels of granularity, allowing for accurate semantic chunking and retrieval across diverse content types.

Distilled Knowledge Layer: The corpus is further distilled into structured forms of knowledge (e.g., knowledge graphs, atomic knowledge, and tabular knowledge). This process, driven by techniques like Named Entity Recognition (NER) [20] and relationship extraction [41], ensures that the distilled knowledge captures key logical relationships and entities, supporting advanced reasoning processes. By organizing this structured knowledge in a distilled layer, we enhance the system’s ability to reason and synthesize based on deeper domain-specific knowledge. The knowledge distillation process is depicted in Figure 4. Below are the detailed distillation processes for typical knowledge forms.

- *Knowledge graph:* Entities and their relationships are extracted from the corpus text using LLMs, generating knowledge units in form of “node-edge-node” structure, where nodes represent entities and edges represent the relationships between them. All knowledge units are then integrated to construct a graph.

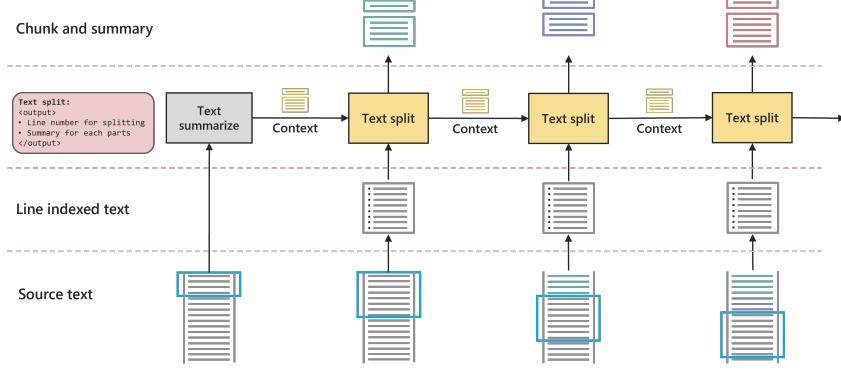


Figure 5: Illustration of enhanced chunking with recurrent text splitting.

- *Atomic knowledge*: The corpus text is partitioned into a set of atomic statements, which are considered as knowledge units. By combining these atomic statements with the relationships between corpus nodes, atomic knowledge is generated.
- *Tabular knowledge*: Entity pairs with specified types and relationships are extracted from corpus text. These entity pairs are treated as knowledge units and can be combined to construct tabular knowledge.

5.2 Level-1: Factual Question focused RAG System

Building upon the L0 system, the L1 system introduces knowledge retrieval and knowledge organization to realize its retrieval and generation capabilities. The primary challenges at this level are semantic alignment and chunking. The abundance of professional terminology and aliases can affect the accuracy of chunk retrieval, and unreasonable chunking can disrupt semantic coherence and introduce noise interference. To mitigate these issues, the L1 system incorporates more sophisticated query analysis techniques and basic knowledge extraction modules. The architecture is expanded to include components that facilitate task decomposition, coordination, and initial stages of knowledge organization (KO), ensuring that the system can manage more complex queries effectively.

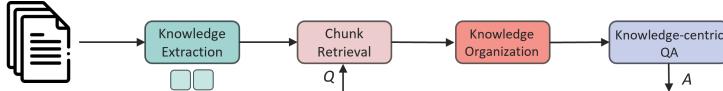


Figure 6: Overview of L1 RAG framework. The squares (■) indicate enhanced chunking and auto-tagging sub-module in knowledge extraction modules.

5.2.1 Enhanced chunking

Chunking involves breaking down a large corpus of text into smaller, more manageable segments. The primary chunking strategies commonly utilized in RAG systems include fixed-size chunking, semantic chunking, and hybrid chunking. Chunking is essential for improving both the efficiency and accuracy of the retrieval process, which consequently affects the overall performance of RAG models in multiple dimensions. In our system, each chunk serves dual purposes: (i) it becomes a unit of information that is vectorized and stored in a database for retrieval, and (ii) it acts as a source for further knowledge extraction and information summarization. Improper chunking not only fails to ensure that text vectors encapsulate the necessary semantic information, but also hinders knowledge extraction based on complete context. For instance, in the context of laws and regulations, a fixed-size chunking approach are prone to destroying text semantics and omitting key conditions, thereby affecting the quality and accuracy of subsequent knowledge extraction.

We propose a text split algorithm to enhance existing chunking methods by breaking down large text documents into smaller, manageable chunks while preserving context and enabling effective summary generation for each chunk. The chunking process is illustrated in Figure 5. Given a source text,

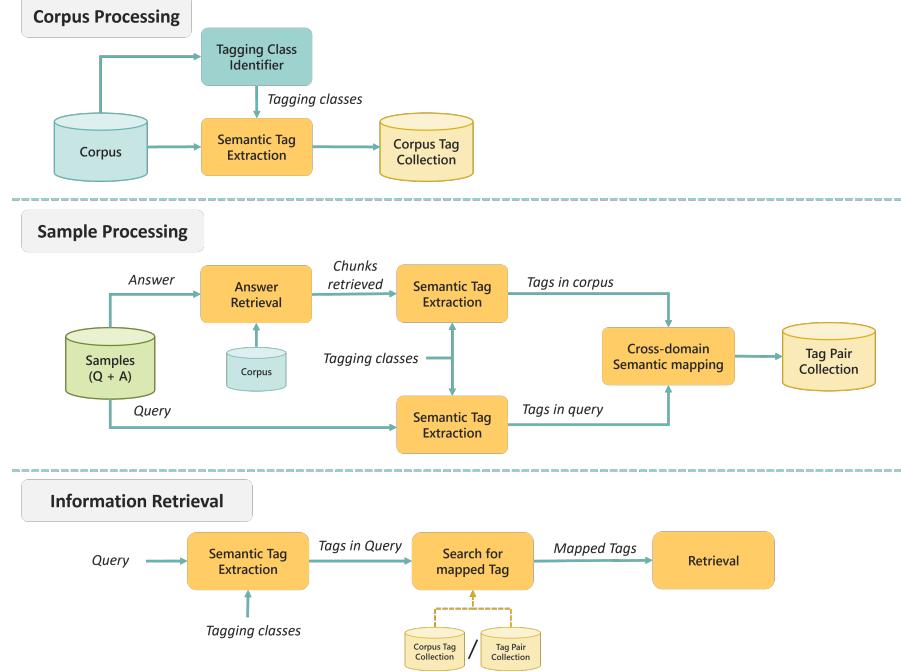


Figure 7: Illustration of the auto-tagging module.

the algorithm iteratively splits the text into chunks. During the first iteration, it generates a forward summary of the initial chunk, providing context for generating summaries of subsequent chunks and maintaining a coherent narrative across splits. Each chunk is summarized using a predefined prompt template that incorporates both the forward summary and the current chunk. This summary is then stored alongside the chunk. The algorithm adjusts the text by removing the processed chunk and updating the forward summary with the summary of the current chunk, preparing for the next iteration. This process continues until the entire text is split and summarized. Additionally, the algorithm can dynamically adjust chunk sizes based on the content and structure of the text.

5.2.2 Auto-tagging

In domain-specific RAG scenarios, the corpus is typically characterized by formal, professional, and rigorously expressed content, whereas the questions posed are often articulated in plain, easily understandable colloquial language. For instance, in medical question-answering (medQA) tasks [33], symptoms of diseases described in the questions are generally phrased in simple, conversational terms. In contrast, the corresponding medical knowledge within the corpus is often expressed using specialized professional terminology. This discrepancy introduces a domain gap that adversely affects the accuracy of chunk retrieval, especially given the limitations of the embedding models employed for this purpose.

To address the domain gap issue, we propose an auto-tagging module designed to minimize the disparity between the source documents and the queries. This module preprocesses the corpus to extract a comprehensive collection of domain-specific tags or to establish tag mapping rules. Prior to the retrieval process, tags are extracted from the query and then mapped to corpus domain using the preprocessed tag collection or tag pair collection. This tag-based domain adaptation can be employed for query rewriting or keyword retrieval within sequential information retrieval frameworks, thereby enhancing both the recall and precision of the retrieval process.

Specifically, we leverage the capabilities of the LLMs to identify key factors within the corpus chunks, summarize these factors, and generalize them into category names, which we refer to as "tag classes." We generate semantic tag extraction prompts based on these tag classes to facilitate accurate tag extraction. In scenarios where only the corpus is available, LLMs are employed with meticulously designed prompts to extract semantic tags from the corpus, thereby forming a comprehensive corpus tag collection. When practical QA samples are available, semantic tag extraction is performed on

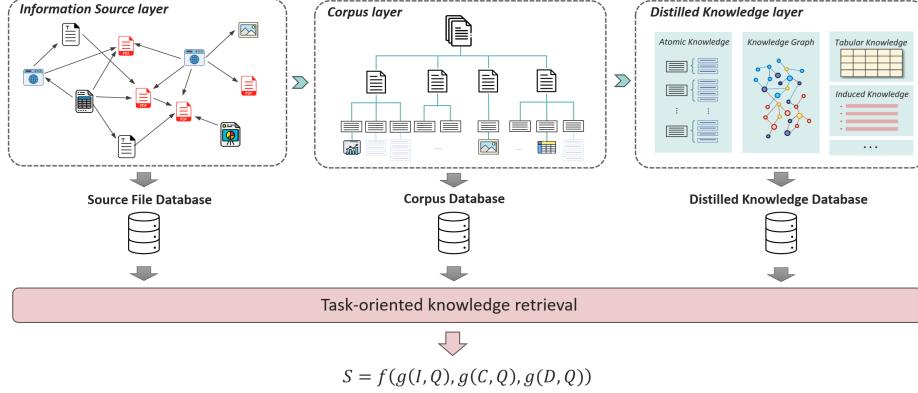


Figure 8: Overview of multi-layer, multi-granularity retrieval over heterogeneous graph

both the queries and the corresponding retrieved answer chunks. Using the tag sets extracted from the chunks and queries, LLMs are utilized to map cross-domain semantic tags and generate a tag pair collection. After establishing both the corpus tag collection and the tag pair collection, tags can be extracted from the query, and the corresponding mapped tags can be identified within the collections. These mapped tags are then used to enhance subsequent information retrieval processes, improving both recall and precision. This workflow leverages the advanced understanding and contextual capabilities of LLMs for domain adaptation.

5.2.3 Multi-Granularity Retrieval

The L1 system is designed to enable multi-layer, multi-granularity retrieval across a heterogeneous knowledge graph, which was constructed in the L0 system. Each layer of the graph (e.g., information source layer, corpus layer, distilled knowledge layer) represents knowledge at different levels of abstraction and granularity, allowing the system to explore and retrieve relevant information at various scales. For example, queries can be mapped to entire documents (information source layer) or specific chunks of text (corpus layer), ensuring that knowledge can be retrieved at the appropriate level for a given task. To support this, similarity scores between queries and graph nodes are computed to measure the alignment between the query and the retrieved knowledge. These scores are then propagated through the layers of the graph, allowing the system to aggregate information from multiple levels. This multi-layer propagation ensures that retrieval can be fine-tuned based on both the broader context (e.g., entire documents) and finer details (e.g., specific chunks or distilled knowledge). The final similarity score is generated through a combination of aggregation and propagation, ensuring that knowledge extraction and utilization are optimized for both precision and efficiency in factual question answering. The retrieval process can be iterative, refining the results based on sub-queries generated through task decomposition, further enhancing the system’s ability to generate accurate and contextually relevant answers.

The overview of multi-layer, multi-granularity retrieval is depicted in Figure 8. For each layer of the graph, both queries Q and graph node are transformed into high-dimensional vector embeddings for similarity evaluation. We denote the similarity evaluation operation as $g(*)$. Here, I , C , and D indicate the node sets in the information source layer, corpus layer, and distilled knowledge layer, respectively. The propagation and aggregation operations are represented by the function $f(*)$. The final chunk similarity score S is obtained by aggregating the scores from other layers and nodes.

5.3 Level-2: Linkable and Reasoning Question focused RAG System

The core functionality of the L2 system lies in its ability to efficiently retrieve multiple sources of relevant information and perform complex reasoning based on it. To facilitate this, the L2 system integrates an advanced knowledge extraction module that comprehensively identifies and extracts pertinent information. Furthermore, a task decomposition and coordination module is implemented to break down intricate tasks into smaller, manageable sub-tasks, thereby enhancing the system’s efficiency in handling them. The proposed framework of L2 RAG system is illustrated in Figure 9.

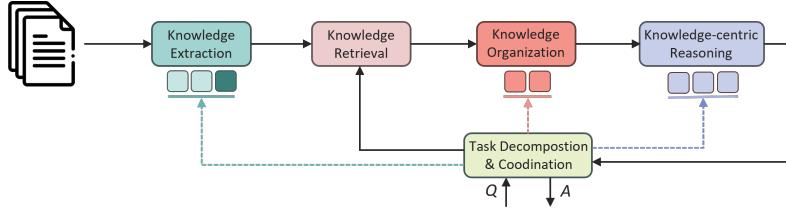


Figure 9: Overview of L2 RAG framework. The square (■) indicates atomic knowledge generation in knowledge extraction module, while the squares (□) represent knowledge reranking and aggregation sub-module in knowledge origination module. Moreover, the squares (□) indicate the multihop reasoning, comparative reasoning, summarizing sub-modules in knowledge-centric reasoning module.

Chunked text contains multifaceted information, increasing the complexity of retrieval. Recent studies have focused on extracting triple knowledge units from chunked text and constructing knowledge graphs to facilitate efficient information retrieval [21, 43]. However, the construction of knowledge graphs is costly, and the inherent knowledge may not always be fully explored. To better present the knowledge embedded in the documents, we propose atomizing the original documents in Knowledge Extraction phase, a process we refer as *Knowledge Atomizing*. Besides, industrial tasks often necessitate multiple pieces of knowledge, implicitly requiring the capability to decompose the original question into several sequential or parallel atomic questions. We refer to this operation as *Task Decomposition*. By combining the extracted atomic knowledge with the original chunks, we construct an atomic hierarchical knowledge base. Each time we decompose a task, the hierarchical knowledge base provides insights into the available knowledge, enabling knowledge-aware task decomposition.

5.3.1 Knowledge Atomizing

We believe that a single document chunk often encompasses multiple pieces of knowledge. Typically, the information necessary to address a specific task represents only a subset of the entire knowledge. Therefore, consolidating these pieces within a single chunk, as traditionally done in information retrieval, may not facilitate the efficient retrieval of the precise information required. To align the granularity of knowledge with the queries generated during task solving, we propose a method called knowledge atomizing. This approach leverages the context understanding and content generation capabilities of LLMs to automatically tag atomic knowledge pieces within each document chunk. Note that, these chunks could be segments of an original reference document, description chunks generated for tables, images, videos, or summary chunks of entire sections, chapters or even documents.

The presentation of atomic knowledge can be various. Instead of utilizing declarative sentences or subject-relationship-object tuples, we propose using questions as knowledge indexes to further bridge the gap between stored knowledge and query. Unlike the semantic tagging process, in knowledge atomizing process, we input the document chunk to LLM as context, ask it to generate relevant questions that can be answered by the given chunk as many as possible. These generated atomic questions are saved as the *atomic question* tags together with the given chunk. An example of knowledge atomizing is demonstrated in Figure 10(c), where the atomic questions encapsulate various aspects of the knowledge contained within the chunk. A hierarchical knowledge base can accommodate queries of varying granularity. Figure 11 illustrates the retrieval process from an atomic knowledge base comprising chunks and atomic questions. Queries can directly retrieve reference chunks as usual. Additionally, since each chunk is tagged with multiple atomic questions, an atomic query can be used to locate relevant atomic questions, which then leads to the associated reference chunks.

5.3.2 Knowledge-Aware Task Decomposition

For a specific task, multiple decomposition strategies might be applicable. Consider Q2 in Figure 1 as an example. The two-step analytical reasoning process depicted may be effective if an interchangeable biosimilar products list is available. However, if only a general list of biosimilar products exists, with attributes dispersed throughout multiple documents, a different decomposition strategy may be necessary: (1) Retrieve the biosimilar product list; (2) Determine whether each product is interchangeable;

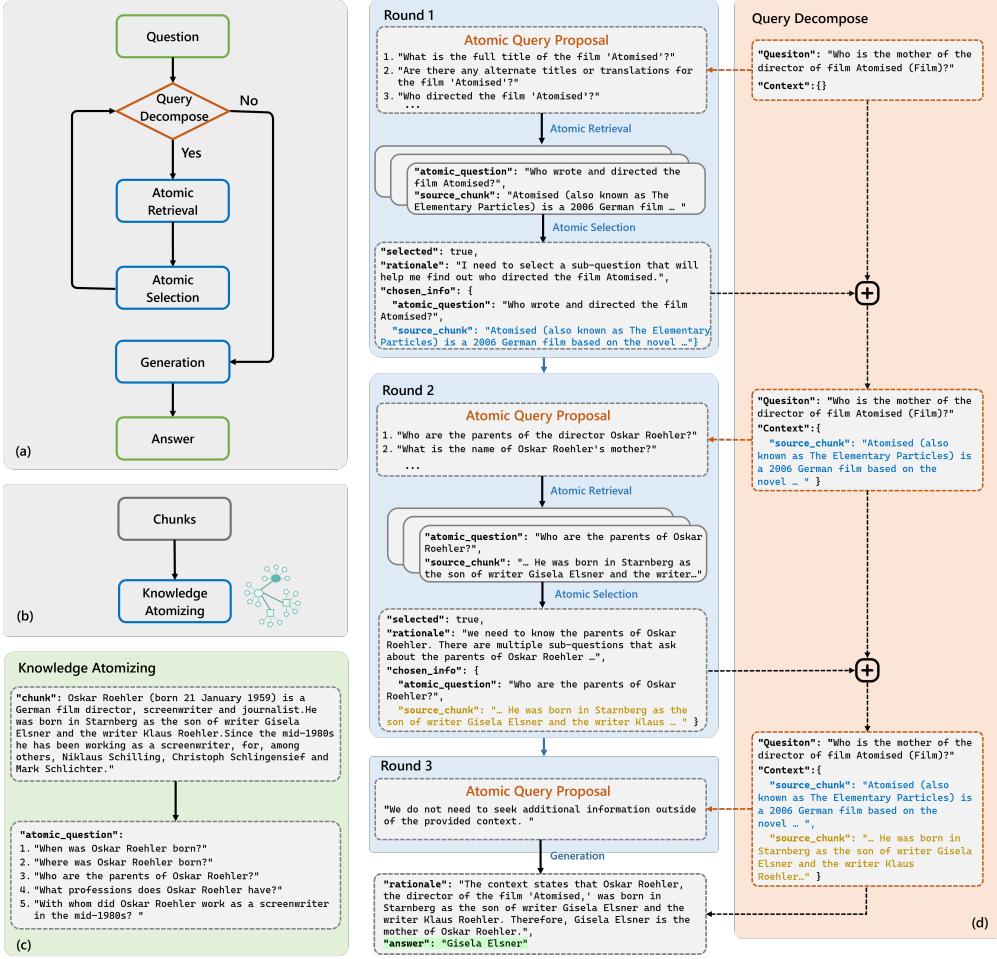


Figure 10: The illustration of knowledge atomizing and knowledge-aware task decomposition: (a) Workflow of task solving with knowledge-aware task decomposition, (b) Workflow of knowledge atomizing, (c) Example of knowledge atomizing, (d) RAG case with knowledge atomizing and knowledge-aware task decomposition.

(3) Count the total number of interchangeable products. The critical factor in selecting the most effective decomposition approach lies in understanding the contents of the specialized knowledge base. Motivated by this, we design the *Knowledge-Aware Task Decomposition* workflow, which is illustrated in Figure 10(a). The complete algorithm for task solving using Knowledge-Aware Task Decomposition is presented in Algorithm 1.

The reference context \mathcal{C}_t is initialized as an empty set, and the original question is denoted by q . As illustrated in the for-loop starting at line 2 of the algorithm, in the t -th iteration, we use an LLM, denoted by \mathcal{LLM} , to generate query proposals potentially useful for task completion, denoted as q_i^t . In this step, the chosen reference chunks \mathcal{C}_t are provided as context to avoid generating proposals linked to already known knowledge. These proposals are then utilized as atomic queries to determine if relevant knowledge exists within the knowledge base. For each atomic question proposal, we retrieve its relevant atomic question candidates along with their source chunks $\{(q_{ij}^t, c_{ij}^t)\}$ from the knowledge base, denoted as \mathcal{KB} . We can use any score metric sim to retrieve atomic questions. In our experiment, we use cosine similarity of their corresponding embeddings to retrieve all top- K atomic questions, provided their similarity to a proposed atomic question is greater than or equal to a given threshold δ . With the original question q , the accumulated context \mathcal{C}_t , and the list of retrieved atomic questions q_{ij}^t , \mathcal{LLM} selects the most useful atomic question q^t from q_{ij}^t and retrieves the relevant chunk c^t . This retrieved chunk is aggregated into the reference context \mathcal{C}_t for the next round of decomposition. Knowledge-aware decomposition can iterate up to N times, where N is

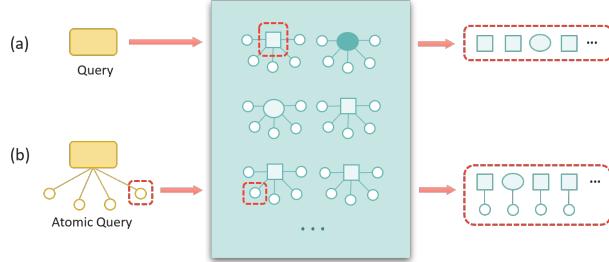


Figure 11: Retrieval process from an atomic knowledge base. It supports two retrieval paths: (a) using queries to directly retrieve chunks as usual; (b) locating atomic nodes first then achieving the associated chunks.

Algorithm 1 Task Solving with Knowledge-Aware Decomposition

```

1: Initialize context  $\mathcal{C}_0 \leftarrow \phi$ 
2: for  $t = 1, 2, \dots, N$  do
3:   Generate atomic question proposals  $\{\hat{q}_i^t\} \leftarrow \mathcal{LLM}(q, \mathcal{C}_{t-1})$ 
4:   For each atomic question proposal  $\hat{q}_i^t$ , retrieve top- $K$  atomic candidates from knowledge base
      
$$\{(q_{ij}^t, c_{ij}^t) \in \mathcal{KB} \mid sim(q_{ij}^t, \hat{q}_i^t) \geq \delta\}$$

5:   Select the most useful atomic question  $q^t \leftarrow \mathcal{LLM}(q, \mathcal{C}_{t-1}, \{q_{ij}^t\})$ 
6:   if  $q^t$  is None then
7:      $\mathcal{C}_t \leftarrow \mathcal{C}_{t-1}$ 
8:     break
9:   else
10:    Fetch the relevant chunk  $c^t$  corresponding to  $q^t$ 
11:    Update context  $\mathcal{C}_t \leftarrow \mathcal{C}_{t-1} \cup \{c^t\}$ 
12:   end if
13: end for
14: Generate answer  $\hat{a} \leftarrow \mathcal{LLM}(q, \mathcal{C}_t)$ 

```

a hyperparameter set to control computational cost. The iteration process can be terminated early if there are no high-quality question proposals, no highly relevant atomic candidates retrieved, no suitable atomic knowledge selections, or if the \mathcal{LLM} determines that the acquired knowledge is sufficient to complete the task. Finally, the accumulated context \mathcal{C}_t is utilized to generate answer \hat{a} for the given question q in line 14.

5.3.3 Knowledge-Aware Task Decomposer Training

It is worth mentioning that knowledge-aware decomposition can be a learnable component. This trained proposer can then directly suggest atomic queries q^t during inference, which means lines 3 to 5 in Algorithm 1 can be replaced by a single call to this learned proposer, thereby reducing both inference time and computational cost. In order to train the knowledge-aware decomposer, we collect data about the rationale behind each step by sampling context and creating diverse interaction trajectories. With this data collected, we train a decomposer that can incorporate domain-specific rationale into the task decomposition and result-seeking process.

The data collection process, as depicted in Figure 12 and Algo. 2, implements a sophisticated dual-dictionary system for managing and tracking information. Our system utilizes two primary data structures: dictionary \mathcal{S} for maintaining comprehensive score records, and dictionary \mathcal{V} for systematically tracking visit frequencies of candidate chunks. During the initialization phase, we establish baseline values by setting all scores to zero and initializing visit counters to one, creating a foundation for dynamic updates throughout the subsequent processing stages.

In each iteration of our decomposition process, the system executes a detailed retrieval operation targeting the top- K' chunks demonstrating maximum relevance to the current atomic question. These chunks must satisfy our similarity threshold criterion (specifically, similarity exceeding δ' , where

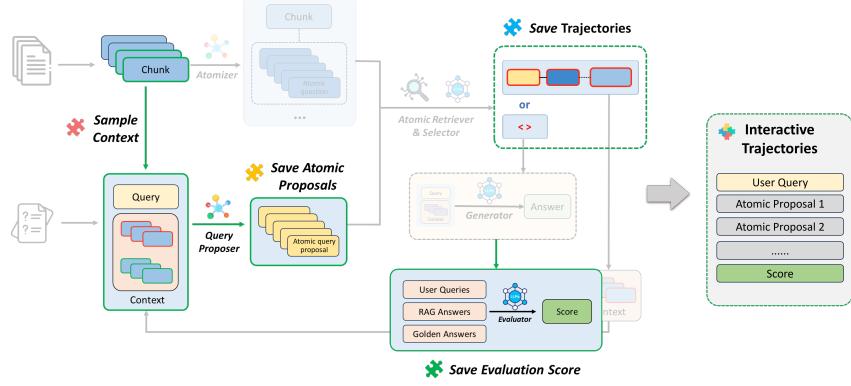


Figure 12: Data collection process for decomposer training, comprising four main components: a) sampling data chunks from the context sampling pool to serve as the reference context for question decomposition, b) saving the generated atomic query proposals, c) after retrieval and selection, saving the chosen atomic query proposals as part of the reasoning trajectories, d) evaluating the answer to generate a score.

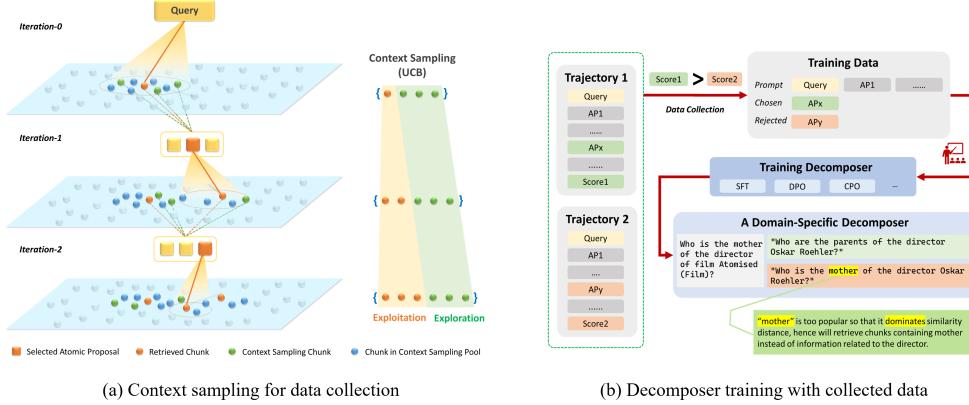


Figure 13: An example of context sampling and an illustration of decomposer training with collected data.

$\delta' < \delta$), with K' intentionally configured to be larger than K to ensure comprehensive coverage. Following this initial retrieval, we carefully select and integrate the data chunks corresponding to the top- K most relevant atomic retrieved pairs into the context. For those retrieved chunks that do not make it into the top- K selection, we systematically incorporate them into \mathcal{S} and methodically update their scores based on precisely calculated relevance metrics.

To ensure comprehensive exploration of the solution space, we have implemented an advanced sampling mechanism that intelligently selects additional chunks from \mathcal{S} when available, incorporating them seamlessly into the reference context. Our implementation leverages the Upper Confidence Bound [9] (UCB) algorithm for context sampling, establishing a balanced approach between exploitation and exploration. The exploitation component manifests through the retriever-selected chunks, focusing on options with currently highest estimated rewards to optimize immediate performance gains. Conversely, the exploration aspect is fulfilled through context sampling from \mathcal{S} , enabling the systematic investigation of less-certain options to accumulate valuable data and potentially uncover superior long-term alternatives.

This meticulously crafted strategy serves a dual purpose: it not only facilitates the generation of diverse and comprehensive atomic query proposals but also enables systematic exploration of multiple potential reasoning pathways. Through this sophisticated approach, we progressively work toward deriving optimal final answers while maintaining a balance between immediate performance optimization and long-term discovery of potentially superior solutions.

Algorithm 2 Collect Data to Train a Knowledge-Aware Decomposer

```

1: Initialize context  $\mathcal{C}_0 \leftarrow \phi$ 
2: Initialize a dictionary to store scores of each chunk  $\mathcal{S} = \{c : 0 \mid \forall c\}$ 
3: Initialize a dictionary to store visits of each chunk  $\mathcal{V} = \{c : 1 \mid \forall c\}$ 
4: for  $t = 1, 2, \dots, N$  do
5:    $c_{sampled} = argmax_c(\mathcal{S}(c) + \alpha \sqrt{\frac{\ln t}{\mathcal{V}(c)}})$ 
6:   Generate atomic question proposals  $\{\hat{q}_i^t\} \leftarrow \mathcal{LLM}(q, \mathcal{C}_{t-1} \cup \{c_{sampled}\})$ 
7:   For each  $\hat{q}_i^t$ , retrieve top- $K'$  atomic candidates from knowledge base
       $\mathbf{AC}_t = \{(q_{ij}^t, c_{ij}^t) \in \mathcal{KB} \mid \exists \hat{q}_i^t. sim(q_{ij}^t, \hat{q}_i^t) \geq \delta'\}$ 
8:   Initialize a list  $\mathbf{RAP}_t = \phi$  to store the most relevant atomic questions
9:   for  $(q, c) \in \mathbf{AC}_t$  do
10:    if  $\exists \hat{q}_i^t. sim(q, \hat{q}_i^t) \geq \delta$  then
11:       $\mathbf{RAP}_t \leftarrow q$ 
12:    else
13:       $\mathcal{S}(c) = \mathcal{S}(c) + \max\{sim(q, \hat{q}_i^t) \mid \forall \hat{q}_i^t\}$ 
14:    end if
15:   end for
16:   Select the most relevant atomic question  $q^t \leftarrow \mathcal{LLM}(q, \mathcal{C}_{t-1}, \mathbf{RAP}_t)$ 
17:   if  $q^t$  is None then
18:      $\mathcal{C}_t \leftarrow \mathcal{C}_{t-1}$ 
19:     break
20:   else
21:     Fetch the relevant chunk  $c^t$  corresponding to  $q^t$ 
22:     Update context  $\mathcal{C}_t \leftarrow \mathcal{C}_{t-1} \cup \{c^t\}$ 
23:     Update score of  $c_t$   $\mathcal{S}(c_t) = 0$ 
24:     Increase visit counts of  $c_t$   $\mathcal{V}(c_t) = \mathcal{V}(c_t) + 1$ 
25:   end if
26: end for
27: Generate answer  $\hat{a} \leftarrow \mathcal{LLM}(q, \mathcal{C}_t)$ 

```

We record atomic proposals (AP), interactive trajectories, and answer scores to support decomposer training. For each specialized domain, interactive trajectories featuring distinct reasoning paths are gathered for decomposer training. This allows us to use the answer score as a supervised signal to train the decomposer. The decomposer training process is depicted in Figure 13. By incorporating preferences in the form of answer scores, the decomposer training can capture domain-specific decomposition rules, thereby adapting the decomposer to meet domain requirements.

Looking ahead, there are several promising avenues for implementing and enhancing our proposed decomposer. We could leverage well-established algorithms such as supervised fine-tuning (SFT) and direct policy optimization (DPO) [46] to train an effective decomposer based on existing LLMs. The practical implementation and performance evaluation of this comprehensive procedure, including detailed empirical analysis and comparative studies, will be addressed in future research work to thoroughly demonstrate its effectiveness and potential applications.

5.4 Level-3: Predictive Question focused RAG System

In the L3 system, there is an increased emphasis on knowledge-based prediction capability, which necessitates effective knowledge collection, organization, and the construction of forecasting rationale. To address this, we leverage the task decomposition and coordination module to build forecasting rationale based on the organized knowledge, which is collected and organized from the retrieved knowledge. The framework of L3 system is illustrated in Figure 14. To ensure the retrieved knowledge is well-prepared for advanced analysis and forecasting, the knowledge organization module is enhanced with specialized submodules dedicated to the structuring and organization of knowledge. These submodules streamline the process of transforming raw retrieved knowledge into a structured, coherent format, optimizing it for subsequent reasoning and predictive tasks. For example, in the

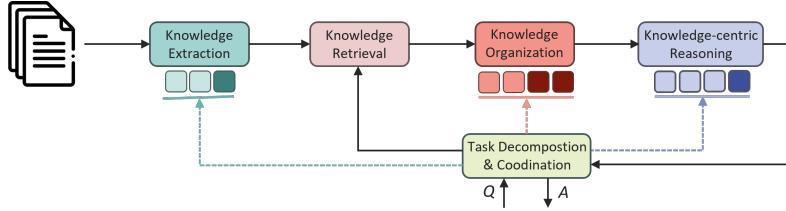


Figure 14: Overview of L3-RAG framework. The squares (■) indicate knowledge structuring and knowledge induction in knowledge organization module, while the square (■) represents forecasting sub-module in knowledge-centric reasoning module.

FDA scenario referred in Figure 1, data from multiple sources—such as medicine labels, clinical trials, and application forms—are integrated into the multi-layer knowledge base. The knowledge structuring submodule follows the instruction from task decomposition module to collect and organize the relevant knowledge (e.g. medicine names with their approval dates) retrieved from knowledge base. The knowledge induction submodule further categorizes this structured knowledge, such as by approval date, to facilitate further statistics analysis and prediction.

Given the limitations of LLMs in applying specialized reasoning logic, their effectiveness in predictive tasks can be restricted. To overcome this, the knowledge-centric reasoning module is enhanced with a forecasting submodule, enabling the system to infer outcomes based on the input queries and the organized knowledge (e.g. total numbers of medicines approved per year). This forecasting submodule allows the system to not only generate answers based on historical knowledge, but also make projections, providing a more robust and dynamic response to complex queries. By integrating advanced knowledge structuring and prediction capabilities, the L3 system can manage and utilize a more complex and dynamic knowledge base effectively.

5.5 Level-4: Creative Question focused RAG System

The L4 system implementation is characterized by the integration of multi-agent systems to facilitate multi-perspective thinking. Addressing creative questions requires creative thinking that draws on factual information and an understanding of underlying principles and rules. At this advanced level, the primary challenges include extracting coherent logical rationales from a retrieved knowledge, navigating complex reasoning processes with numerous influencing factors, and assessing the quality of responses to creative, open-ended questions. To tackle these challenges, the system coordinates multiple agents, each contributing unique insights and reasoning strategies, as illustrated in Figure 15. These agents operate in parallel, synthesizing various thought processes to generate comprehensive and coherent solutions. This multi-agent architecture supports the parallel processing and integration of diverse reasoning paths, ensuring effective management and response to intricate queries. By simulating diverse viewpoints, the L4 system enhances its ability to tackle creative questions, generating innovative ideas rather than predefined solutions. The coordinated outputs from multiple agents not only enrich the reasoning process but also provide users with comprehensive perspectives, fostering creative thinking and inspiring novel solutions to complex problems.

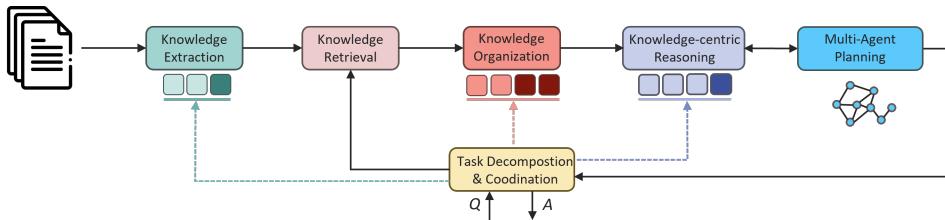


Figure 15: Overview of L4-RAG framework. The multi-agent planning module is introduced to enable multi-perspective thinking.

6 Evaluation and Metrics

To validate the effectiveness of our proposed method, we conduct experiments on both open-domain benchmarks and domain-specific benchmarks. We delineate the evaluation metrics and methods employed to assess the performance of proposed knowledge-aware task decomposition method in Section 6.1. The evaluation results on three open-domain benchmarks are presented in Section 6.2, while the results on two legal domain-specific benchmarks in Section 6.3. Furthermore, we present in-depth analysis through three real case studies in Section 6.5, which highlight the superiority of our method compared to existing decomposition approaches.

6.1 Experimental Setup

Methods To thoroughly evaluate the performance of our proposed knowledge-aware decomposition approach (described in Section 5.3), we have selected a variety of baseline methods that represent different strategies for task-solving with LLMs. We include **Zero-Shot CoT**[35] to assess the inherent reasoning capabilities and embedded knowledge of the underlying LLM without additional context. **Naive RAG**[36], which introduces external knowledge through retrieval, serves as a benchmark for evaluating the incremental benefits of augmented knowledge. The **Self-Ask** framework[44] is employed to investigate the impact of an iterative question decomposition and answering strategy on task performance. Additionally, **GraphRAG** [21] is evaluated in both local and global modes to assess the impact of knowledge graph-based methods on multi-hop reasoning tasks.

To ensure a fair comparison and to highlight the influence of hierarchical knowledge structures, we have extended Naive RAG and Self-Ask to utilize both a general flat knowledge base, denoted as R , and a hierarchical retriever, denoted as $H\text{-}R$, as introduced in Figure 11. The hierarchical retriever ($H\text{-}R$) utilizes the questions or follow-up questions to retrieve chunks through both path (a) and path (b) at the same time. The retrieved chunks from both paths are then aggregated to form a comprehensive reference context for LLM to answer each question, potentially enhancing the relevance of the provided context.

The experimental methods are summarized as follows:

- **Zero-Shot CoT:** Questions are addressed using solely the Chain-Of-Thought (CoT) technique, which prompts the LLMs to articulate its reasoning process step-by-step without the aid of example demonstrations or supplemental context. This method assesses the LLMs' intrinsic knowledge and reasoning capabilities in a zero-shot setting.
- **Naive RAG w/ R:** This approach employs dense retrieval from a flat knowledge base to procure relevant information for each question. The knowledge base consists of pre-embedded chunks are matched to the original question based on semantic similarity. The retrieval process is direct, without any intermediate task decomposition.
- **Naive RAG w/ H-R:** This method extends the Naive RAG framework by incorporating a hierarchical retrieval process ($H\text{-}R$) that operates through two concurrent paths. Path (a) performs a direct retrieval of knowledge chunks in response to the original question, similar to the flat retrieval approach. Path (b), on the other hand, uses the original question again to find the relevant atomic questions and obtain the corresponding chunks. The combined output from both paths is then aggregated, creating a rich reference context.
- **Self-Ask:** This method employs a task decomposition strategy wherein the LLMs is prompted to iteratively generate and answer follow-up questions, thereby breaking down complex problems into more manageable sub-tasks. General demonstrations illustrating the logic and methodology of task decomposition are provided for all benchmarks to guide the LLMs' reasoning process. As detailed in the original paper [44], the framework encourages the LLMs to engage in a recursive dialogue with itself, generating intermediate answers that progressively build towards the final answer. In this setting, the LLMs relies solely on its inherent knowledge base, as no external contexts are introduced to aid in answering the follow-up questions.
- **Self-Ask w/ R:** Building upon the **Self-Ask** method, this setting introduces an additional retrieval component, for each follow-up question generated by the LLMs, relevant chunks are retrieved from a flat knowledge base to provide a reference context. The retrieval process uses the follow-up question as the query. This approach seeks to combine the benefits

of iterative task decomposition with rich external knowledge from retrieval, potentially improving the LLMs’ performance on complex reasoning tasks.

- **Self-Ask w/ H-R:** This variant of the Self-Ask method enhances the retrieval process by utilizing a hierarchical knowledge base, as opposed to the flat one used in **Self-Ask w/ R**. When the LLMs generates follow-up questions, these are employed as queries in a dual-path retrieval system, specifically paths (a) and (b) in Figure 11. The outputs from both retrieval paths are then aggregated to form a richer reference context.
- **GraphRAG Local:** In this approach, the flat knowledge base is pre-processed to construct a knowledge graph in accordance with the public guidance. The inference is run in local mode.
- **GraphRAG Global:** The inference is run in global mode in this setting.
- **Ours:** The proposed knowledge-aware decomposition method iteratively decomposes complex questions into sub-questions and retrieves relevant knowledge up to a maximum of five iterations. This process limits the context for the final answer to the five most useful knowledge chunks.

Metrics For maintaining consistency with established benchmarks, two conventional metrics are adopted in our experimental evaluation: **Exact Match (EM)**, which assesses whether the response is identical to a predefined correct answer, and the **F1** score, which is the harmonic mean of precision and recall at the token level. During evaluation, we noticed that the LLM sometimes produced responses more verbose than expected, even when the QA prompt aimed to limit output style. To more accurately gauge the responses’ alignment with the intended answers—beyond mere lexical matching—we introduced a novel evaluation metric employing *GPT-4*. In this process, *GPT-4* acts as an evaluator, assessing the correctness of a response in relation to the question and the correct answer labels. We refer to this metric as **Accuracy (Acc)**. Upon manual inspection of a sample set, the judgments rendered by *GPT-4* demonstrate complete agreement with human evaluators, affirming the reliability of this metric.

Furthermore, we encountered situations where a method achieves high accuracy (Acc) scores yet registers low F1 scores. To elucidate the underlying factors of such discrepancies, we also report on the **Recall** and **Precision** of the generated responses. Recall measures the proportion of relevant tokens from the answer labels that are captured in the response, while precision evaluates the relevance of the tokens in the generated answer with respect to the correct labels. Specifically, in cases where multiple correct answer labels are available, we employ a conservative scoring approach for **EM**, **F1**, **Precision**, and **Recall** by retaining the highest score achieved. This approach is designed to equitably consider the range of correct answers that the LLM may generate. It should be noted that, in the context of computing **Accuracy (Acc)**, all admissible answer labels are furnished concurrently to the evaluation process, resulting in a singular Accuracy score.

The metrics employed in this evaluation — **Exact Match (EM)**, **F1**, **Precision**, **Recall**, and **Accuracy (Acc)** — are primarily suited for questions categorized as L1 and L2, which are characterized by their association with ground truth answers that are factual and definitive. However, the utility of these metrics diminishes for predictive and creative questions, namely the L3 and L4 questions, where answers are inherently uncertain or subjective, and no single correct response exists. For L3 questions, alternative assessment methods such as trend judgment and qualitative analysis become more appropriate to capture the predictive validity of the responses. Furthermore, for L4 questions, which demand a higher degree of insight or innovation, it is essential to evaluate answers through a multi-faceted lens, considering criteria such as relevance, diversity, comprehensiveness, uniqueness, and inspiration to fully appreciate the depth and originality of the approaches’ responses.

LLM and Hyper-parameters In our experiments, we employ GPT-4 (1106-Preview version) across all the methods outlined previously. For the knowledge extraction phase, we utilize a *temperature* setting of 0.7 specifically for the *Knowledge Atomizing* process, promoting a balance between diversity and determinism in the generated atomic knowledge. Conversely, during all question-answering (QA) steps in each method, we implement a *temperature* of 0, ensuring consistent responses from the model. Regarding the retrieval component, we engage the *text-embedding-ada-002* (version 2) as our embedding model for both the general flat knowledge bases and the hierarchical knowledge bases. For the general flat knowledge bases, the retriever is configured to fetch up to 16 knowledge chunks, applying a retrieval score threshold of 0.2. In the case of hierarchical knowledge bases, the

retriever is initially set to retrieve a maximum of 8 chunks with a more stringent threshold of 0.5. Subsequently, an additional 4 chunks can be retrieved via each atomic query posed.

6.2 Evaluation on Open-Domain Benchmarks

In this subsection, we demonstrate the performance of our method across three open-domain benchmarks. To ensure a fair and objective evaluation, particularly in the context of real-world industrial applications, we have selected three widely-recognized multi-hop datasets: HotpotQA [64], 2WikiMultiHopQA [29], and MuSiQue [54]. Below, we provide a brief overview of these datasets, noting that our method does not leverage the question type information nor the number of hops information during the solving process, as our approach is designed to be agnostic to such classifications.

HotpotQA The HotpotQA dataset is a well-known multi-hop QA benchmark primarily consisting of 2-hop questions, each associated with 10 Wikipedia paragraphs. Among these, some paragraphs contain supporting facts essential to answering the question, while the rest serve as distractors. The dataset also includes a *question type* field, which delineates the logical reasoning required—*comparison* questions involve contrasting two entities, and *bridge* questions require inferring the bridge entity, or inferring the property of an entity through an intermediary entity, or locating the answer entity [64]. The *comparison* questions in HotpotQA align with the *comparative* questions defined in Section 3.2. Similarly, *bridge* questions correspond to either *bridging* questions or *summarizing* questions, depending on the complexity of the rationale required. Although our method operates independently of these types, their description here exemplifies the nature of questions within the dataset and contextualizes the expected performance variance across different benchmarks.

2WikiMultiHopQA Inspired by HotpotQA, 2WikiMultiHopQA expands the diversity of question types. It retains the *comparison* type from HotpotQA and introduces *inference* and *compositional* questions that evolve from the *bridge* type by focusing on entity attribute deduction and entity location, respectively. Additionally, the *bridge comparison* type is a novel category that requires a synthesis of *bridge* and *comparison* reasoning. In this dataset, the *comparison* questions correspond to the *comparative* questions defined in Section 3.2, akin to those in HotpotQA. The *inference* questions are analogous to *bridging* questions, and the *compositional* questions are similar to *summarizing* questions as described in the same section. The *bridge_comparison* questions, due to their hybrid nature and increased complexity, also fall under the *summarizing* questions category. This dataset typically presents 2-hop to 4-hop questions, each accompanied by 10 Wikipedia paragraphs containing supporting facts and distractors. While these types inform the dataset’s structure, they are not utilized by our method, which treats all questions uniformly regardless of their categorization.

MuSiQue Addressing the issue that many multi-hop questions can be solved via shortcuts—arriving at correct answers without proper reasoning—MuSiQue implements stringent filters and additional mechanisms specifically designed to encourage connected reasoning, as reported by Trivedi et al. [54]. Unlike the other datasets, MuSiQue does not categorize questions by type, but it does provide explicit information on the number of hops required for each question, ranging from 2 to 4 hops. Each question is associated with 20 context paragraphs, which introduce a mix of relevant and irrelevant information, further complicating the task of discerning the correct reasoning path. This explicit hop information, while not used by our method, underscores the complexity of the dataset and the robustness required by models to handle such challenges effectively.

In our experiments, we randomly sample 500 QA data from the *dev* set of each dataset, without consideration for question type nor number of hops, to ensure randomness. We compile the context paragraphs from all sampled QA data into a single knowledge base for each benchmark, creating a more complex retrieval scenario. This design choice is aimed at rigorously assessing our model’s question decomposition and relevant context retrieval abilities. Table 3 outlines the distribution of question types within our sampled sets, offering insight into the variety of reasoning challenges presented in our evaluation, though this does not directly impact our method.

Overall Performance The evaluation results across HotpotQA, 2WikiMultiHopQA, and MuSiQue are presented in Table 4, Table 5, and Table 6, respectively. If we hypothesize that the highest achievable performance on each benchmark may reflect its relative difficulty, a tentative ranking from easiest to most challenging would be: HotpotQA, 2WikiMultiHopQA, and MuSiQue. Our

Table 3: Distribution of question types across three multi-hop QA datasets.

Type	Count	Ratio	#Hops	Count	Ratio
comparison	107	21.4%	2	263	52.6%
bridge	393	78.6%	3	169	33.8%
			4	68	13.6%

Table 4: Performance comparison on HotpotQA. Best in bold, second-best underlined.

Method	EM	F1	Acc	Precision	Recall
Zero-Shot CoT	32.60	43.94	53.60	46.56	43.97
Naive RAG w/ R	<u>56.80</u>	<u>72.67</u>	82.60	<u>74.52</u>	74.86
Naive RAG w/ H-R	54.80	70.25	81.60	72.56	72.24
Self-Ask	28.80	43.61	59.60	43.49	56.21
Self-Ask w/ R	44.80	63.08	81.00	63.23	74.57
Self-Ask w/ H-R	47.20	64.24	82.20	64.27	75.95
GraphRAG Local	0.00	10.66	89.00	5.90	83.07
GraphRAG Global	0.00	7.42	64.80	4.08	63.16
Ours	61.20	76.26	<u>87.60</u>	78.10	<u>78.95</u>

observations suggest that for HotpotQA, considered the least challenging, the GraphRAG in local mode and our method are closely competitive, with minor performance disparities. However, as the difficulty increases for 2WikiMultiHopQA and MuSiQue, our method outperforms others.

The inclusion of retrieved context significantly enhances accuracy, with gains ranging from approximately 10% (comparing Zero-Shot CoT and Naive RAG on MuSiQue) to around 29% (on HotpotQA). This indicates that for simpler benchmarks, RAG equipped with naive knowledge retrieval could address simple multihop questions, leading to a significant accuracy boost. However, for more challenging benchmarks involving complex multihop questions, the accuracy improvement from naive knowledge retrieval is limited, underscoring the constrained reasoning capabilities of the LLMs. By incorporating decomposition mechanisms, Self-Ask significantly enhances accuracy, especially on more challenging benchmarks. The combination of knowledge retrieval and Self-Ask decomposition yields superior results on 2WikiMultiHopQA and MuSiQue, compared to using a single mechanism. However, in the case of HotpotQA, all methods employing retrieval (except for GraphRAG in Global mode, which will be discussed later) attain accuracies above 80%, with negligible differences between them.

Interestingly, the application of a hierarchical atomic knowledge base does not significantly impact Naive RAG’s performance compared to Naive RAG with general flat knowledge base, potentially due to the embedding distance between the original multi-hop questions and the atomic questions of relevant contexts. Nonetheless, when combined with task decomposition, a hierarchical knowledge base shows more promise, as evidenced by the performance boost observed in Self-Ask with Hierarchical Retrieval (Self-Ask w/ H-R) compared to Self-Ask with Retrieval (Self-Ask w/ R), particularly on MuSiQue, which requires more complex reasoning. This improvement underscores the potential of hierarchical knowledge bases in enhancing the effectiveness of decomposition mechanisms in complex reasoning tasks.

Our proposed method focuses on knowledge-aware task decomposition, which performs decomposition with an awareness of available knowledge, effectively leveraging the atomic information provided by the hierarchical knowledge base. Experimental results demonstrate that our approach consistently outperforms other methods, validating its effectiveness in complex reasoning scenarios.

Regarding GraphRAG, originally designed for the query-focused summarization (QFS) task as outlined by [21], we observe its suboptimal performance in both local and global modes compared to our method. Notably, GraphRAG exhibits a curious trend: it achieves higher accuracy and recall

Table 5: Performance comparison on 2WikiMultiHopQA. Best in bold, second-best underlined.

Method	EM	F1	Acc	Precision	Recall
Zero-Shot CoT	35.67	41.40	43.87	41.43	43.11
Naive RAG w/ R	51.20	59.74	62.80	59.06	62.30
Naive RAG w/ H-R	<u>51.40</u>	59.73	63.00	59.36	62.43
Self-Ask	23.80	37.49	51.60	34.56	60.72
Self-Ask w/ R	46.80	<u>64.17</u>	79.80	61.17	80.21
Self-Ask w/ H-R	48.00	63.99	<u>80.00</u>	<u>61.30</u>	<u>79.56</u>
GraphRAG Local	0.00	11.83	71.20	6.74	75.17
GraphRAG Global	0.00	7.35	45.00	4.09	55.43
Ours	66.80	75.19	82.00	74.04	78.87

Table 6: Performance comparison on MuSiQue. Best in bold, second-best underlined.

Method	EM	F1	Acc	Precision	Recall
Zero-Shot CoT	12.93	22.90	23.47	24.40	24.10
Naive RAG w/ R	<u>32.00</u>	43.31	44.40	<u>44.42</u>	47.29
Naive RAG w/ H-R	30.40	41.30	43.40	42.06	44.53
Self-Ask	16.40	27.27	35.40	26.33	37.65
Self-Ask w/ R	28.40	42.54	49.80	41.13	53.37
Self-Ask w/ H-R	29.80	<u>44.05</u>	<u>54.00</u>	42.47	<u>55.89</u>
GraphRAG Local	0.60	9.62	49.80	5.73	55.82
GraphRAG Global	0.00	5.16	44.60	2.82	52.19
Ours	46.40	56.62	59.60	57.45	59.53

scores while performing lower on EM, F1, and Precision metrics. A closer analysis of GraphRAG’s outputs reveals a tendency to echo the query and include meta-information about the answer within its graph structure. Despite attempts to refine its QA prompt, this behavior persists. An illustrative example is presented in Table 7, which shows GraphRAG Local’s response to a question from HotpotQA.

6.3 Evaluation on Legal Benchmarks

In this subsection, we present the performance of our approach on two legal benchmarks: LawBench [23] and Open Australian Legal QA [15]. Before doing so, we provide a brief description of each benchmark.

LawBench LawBench is a comprehensive legal benchmark for Chinese laws. It comprises 20 meticulously designed tasks aimed at accurately assessing the legal capabilities of LLMs. Unlike some existing benchmarks that rely solely on multiple-choice questions, LawBench includes a variety of task types that are closely related to real-world applications. These tasks encompass legal entity recognition, reading comprehension, crime amount calculation, and legal consulting, among others. Since not all tasks are RAG-oriented (e.g., reading comprehension), we have selected 6 specific tasks, which are detailed in Table 8. The number of questions of each task is 500.

We also provide example questions of these tasks for the readers reference.

- 1-1: Answer the following question by directly providing the content of the article:
 ↪ What is the content of Article 76 of the Securities Law?
- 1-2: According to the ‘Securities Law’, which of the following statements about stock exchanges is incorrect? A: Without the permission of the stock exchange, no entity or individual may publish real-time securities trading information; B: The stock exchange may restrict trading on securities accounts that exhibit major abnormal trading conditions as needed, and report to the securities regulatory authority under the State Council for record; C: The accumulated property of a member-based stock exchange belongs

Table 7: An Example of GraphRAG Local output on a HotpotQA question. The table showcases the tendency to repeat the question and include meta-information in its response.

Question	Which country is home to Alsa Mall and Spencer Plaza?
Answer Labels	India
Answer of GraphRAG	Alsa Mall and Spencer Plaza are both located in Chennai, India [Data: India and Chennai Community (2391); Entities (4901, 4904); Relationships (9479, 1687, 5215, 5217)].

Table 8: Overview of LawBench tasks

Task No.	Task	Type	Metric
1-1	Statute Recitation	Generation	F1
1-2	Legal Knowledge Q&A	Single Choice	EM
3-1	Statute Prediction (Fact-based)	Multiple Choices	EM
3-2	Statute Prediction (Scenario-based)	Generation	F1
3-6	Case Analysis	Single Choice	EM
3-8	Consultation	Generation	F1

- to the members, and their rights are jointly enjoyed by the members; during → its existence, the accumulated property may not be distributed to the → members; D: The stock exchange formulates listing rules, trading rules, → member management rules, and other relevant rules in accordance with → securities laws and administrative regulations, and reports to the → securities regulatory authority under the State Council for record.
- 3-1: Based on the following facts and charges, provide the relevant articles of the → Criminal Law. Facts: The Yushu City, Jilin Province, accused that on → November 15, 2015, the defendant He signed a car rental agreement with Guo, → the owner of a taxi with license plate number xxx. The agreement stipulated → a monthly rent of RMB 3,900.00, payable monthly. On January 19, 2016, → without the knowledge of Guo, the defendant He concealed the truth and → falsely claimed to be the owner of the taxi. He signed a car rental → agreement with the victim Ma, with a monthly rent of RMB 3,800.00 and a → rental period of one year, collecting a total of RMB 50,600.00 from Ma for → one year's rent and vehicle deposit. On February 26, 2016, the taxi was → retrieved by its owner Guo from the victim Ma. The victim Ma repeatedly → asked the defendant He to return the rent and deposit, but the defendant He → refused to return them. The prosecution provided evidence including the → defendant's confession, the victim's statement, witness testimonies, and → documentary evidence, and believed that the defendant He, with the purpose → of illegal possession, defrauded others of their property by fabricating → facts and concealing the truth during the signing and performance of the → contract. The amount was relatively large, and his actions violated the → provisions of Article xx of the Criminal Law of the People's Republic of → China, and he should be held criminally responsible for xx. Charge: Contract → Fraud.
- 3-2: Please provide the legal basis according to the specific scenario and question, → only the content of the specific legal provision is needed, each scenario → involves only one legal provision. Scenario: A cargo ship arrives at the → port of discharge, but the consignee fails to arrive in time to collect the → goods. Under which legal provision can the captain unload the goods at → another appropriate place?
- 3-6: One year after the bar opened, the business environment changed drastically, → and all partners held a meeting to discuss countermeasures. According to the → 'Partnership Enterprise Law,' the following voting matters are considered → valid votes: A: Zhang believes that the name 'Tongcheng' is not attractive → and proposes to change it to 'Tongsheng Bar.' Wang and Zhao agree, but Li → opposes; B: In view of the sluggish business, Wang proposes to suspend → operations for one month for renovation and reorganization. Zhang and Zhao → agree, but Li opposes; C: Due to the urgent needs of the bar, Zhao proposes

Table 9: Evaluation Results on Legal Benchmarks (Metric is **F1** / **EM** as indicated in Table 8)

Task		Zero-Shot CoT	GraphRAG Local	Ours (N=5)
LawBench	1-1	21.31	<u>23.27</u>	78.58
	1-2	54.24	<u>62.60</u>	70.60
	3-1	53.32	<u>74.60</u>	83.16
	3-2	<u>27.51</u>	25.98	46.05
	3-6	<u>51.16</u>	47.64	61.91
	3-8	17.44	<u>18.43</u>	23.58
	Open Australian Legal QA	25.10	<u>34.35</u>	63.34

Table 10: Evaluation Results on Legal Benchmarks (Metric is **Acc**)

Task		Zero-Shot CoT	GraphRAG Local	Ours (N=5)
LawBench	1-1	1.23	<u>16.60</u>	90.12
	1-2	54.00	<u>63.40</u>	70.60
	3-1	49.90	<u>75.40</u>	88.82
	3-2	15.83	<u>27.60</u>	67.54
	3-6	51.12	<u>57.00</u>	62.73
	3-8	49.70	<u>58.80</u>	61.72
	Open Australian Legal QA	16.48	<u>88.27</u>	98.59

- to sell a batch of coffee machines to the bar. Zhang and Wang agree, but Li opposes; D: Given the four partners' lack of experience in bar management, → Li proposes to appoint his friend Wang as the managing partner. Zhang and → Wang agree, but Zhao opposes.
- 3-8: Resident A rented out the house to B. With A's consent, B renovated the rented → house and sublet it to C. C unilaterally altered the load-bearing structure → of the house. Why can A request B to bear liability for breach of contract?

Open Australian Legal QA The benchmark consists of 2,124 questions and answers synthesized by GPT-4 from the Australian legal corpus. All questions are of the generation type. One example is: "What is the landlord's general obligation under section 63 of the Act in the case of Anderson v Armitage [2014] NSWCATCD 157 in New South Wales?"

Evaluation results are listed in Table 9, where we only compare to "GraphRAG Local", as it generally performs better than "GraphRAG Global" on these tasks.

For the aforementioned reasons, we also use GPT-4 to evaluate all experimental results, reporting the accuracy (**Acc**) in Table 10. When comparing the results in Table 9 and Table 10, we observe that the order of the results is preserved, even though some metrics change significantly. In the following section, we aim to identify the reasons behind these changes, which may provide valuable insights for designing better metrics to evaluate RAG frameworks in the future.

1. The accuracy of our approach increases significantly for generation tasks (1-1, 3-2, Open Australian Legal QA). For these tasks, our answers are often semantically equivalent but syntactically different from the golden answers. This explains the improved metric performance, as GPT-4 can compare the semantic content of the answers. This also applies to the "GraphRAG Local" results for the "Open Australian Legal QA" task.
2. The accuracy of "GraphRAG Local" decreases for generation tasks 1-1 and 3-2. These tasks involve statute recitation and prediction, requiring the retrieval of specific articles. Upon detailed examination, We find that "GraphRAG Local" often fails to retrieve the correct articles or references the wrong ones, but it tends to repeat the legal information. Therefore, token-level recall can be improved by simply rephrasing legal names and common prefixes, such as "According to XX law, XX articles...".

3. Both our approach and "GraphRAG Local" show significant accuracy improvements on task 3-8. Besides the reason mentioned in the first point, the quality of the golden answers may also contribute to this difference. The questions and golden answers in task 3-8 are sourced from a consulting website, resulting in varying quality. For example, one question asks "Do the children from the original marriage have an obligation to support the father?" However, the provided golden answer includes an irrelevant article, "Article 1067," which pertains to parents' obligations to support minor children.

Question: In the case where both parents are divorced and have formed their
 ↵ own families with new children, and according to the court's judgment
 ↵ , the father is required to pay monthly child support to the mother
 ↵ until the child is 18 years old. Do the children from the original
 ↵ marriage have an obligation to support the father?

Reference Answer: In our country, biological children have an obligation to
 ↵ support their parents who are divorced. The relationship between
 ↵ children and parents does not dissolve because of the divorce of the
 ↵ children or parents. Supporting parents is a legal obligation of the
 ↵ children. If the children do not support the elderly, the parents can
 ↵ directly file a lawsuit in the People's Court requesting the court
 ↵ to order the daughter to pay support. Legal basis: Article 1067 of
 ↵ the Civil Code of the People's Republic of China states that if
 ↵ parents do not fulfill their obligation to support, minor children or
 ↵ adult children who cannot live independently have the right to
 ↵ request support from their parents. If adult children do not fulfill
 ↵ their obligation to support, parents who are incapacitated or have
 ↵ difficulties in living have the right to request support from adult
 ↵ children. Article 1084 states that the relationship between parents
 ↵ and children does not dissolve because of the parents' divorce. After
 ↵ the divorce, whether the children are directly raised by the father
 ↵ or the mother, they are still the children of both parents. Lawyer's
 ↵ explanation: After the parents are divorced, adult children still
 ↵ have the obligation to support both parents. The obligation of adult
 ↵ children to support their parents does not change because of changes
 ↵ in the parents' relationship. If adult children do not fulfill their
 ↵ obligation to support, parents who are incapacitated or have
 ↵ difficulties in living have the right to request support from adult
 ↵ children.

Our Answer: Yes, there is an obligation to support. Legal basis: According
 ↵ to Article 1069 of the Civil Code of the People's Republic of China,
 ↵ it is stipulated that: 'The obligation of children to support their
 ↵ parents does not terminate because of changes in the parents' marital
 ↵ relationship.' Therefore, even if the parents are divorced and
 ↵ remarried with new children, the original children still have an
 ↵ obligation to support their parents.

4. The accuracy of all methods on choice tasks 1-2, 3-1, and 3-6 almost coincides with the F1 score, as expected. An exception is task 3-1, where the difference is mainly due to GPT-4's capacity to understand Chinese, particularly in distinguishing numbers in Arabic and Chinese. In Chinese law, all numbers are written in Chinese, while in the golden answers, all numbers are given in Arabic.

6.4 Experiments on Domain Aligned Atomic Proposers

Given that all methods demonstrate notably low accuracy scores on the MuSiQue dataset, we specifically selected this dataset to investigate whether implementing a fine-tuned atomic proposer could potentially enhance overall performance metrics and improve the quality of results.

Data Collection To establish a comprehensive and efficient experimental framework, we initiate our process by randomly sampling 1000 questions from the dataset. These questions are subsequently divided into two distinct sets following a standard split ratio: 80% of the questions are designated as training data, specifically utilized for collecting and analyzing decomposition trajectories, while the remaining 20% are reserved as an evaluation dataset to assess the model's performance. For each question within the training dataset, we meticulously follow the procedural framework outlined

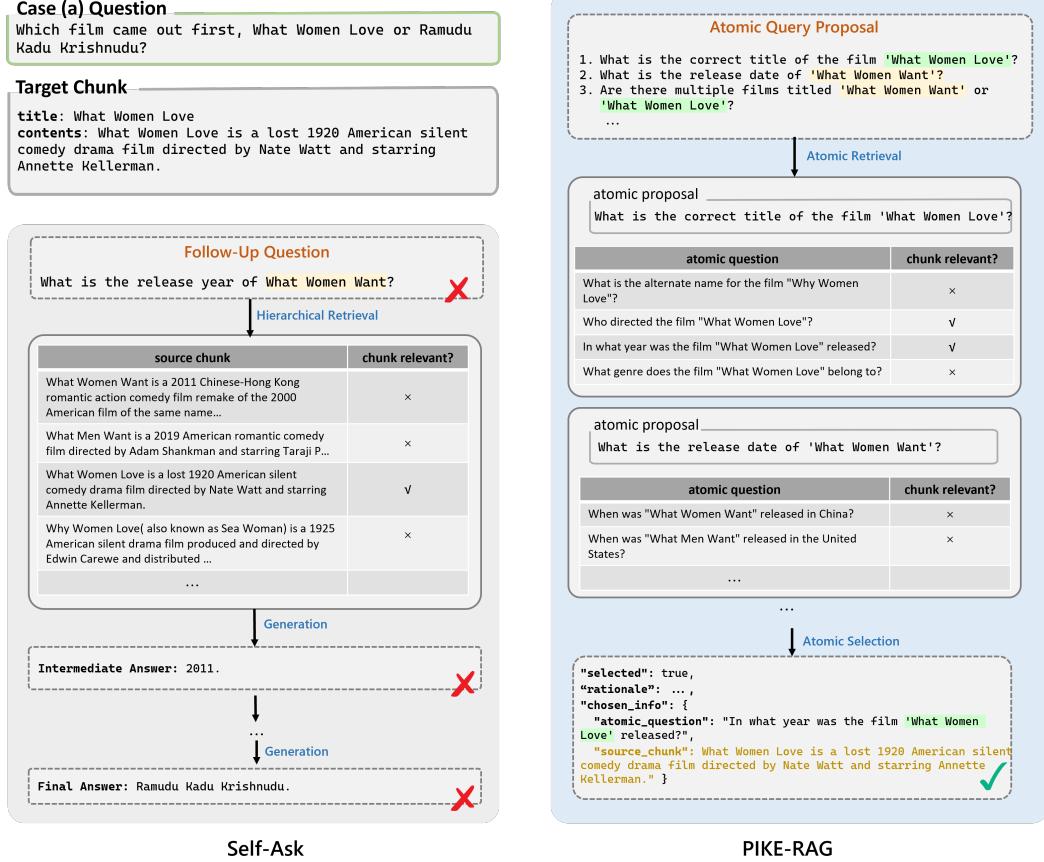


Figure 16: Case (a): Given the lesser-known film "What Women Love" as opposed to the more popular "What Women Want," single-path methods like Self-Ask on the left are predisposed to generating follow-up questions about the latter, leading to an incorrect final answer. Conversely, PIKE-RAG can effectively discern the intended meaning of the original question by positing several atomic queries and postpone the task understanding to atomic selection phase with relevant atomic questions provided, and subsequently arriving at an accurate conclusion.

in Figure 12, systematically generating follow-up questions in a stepwise progression until we successfully obtain a decomposition trajectory that leads to the correct answer. We establish a maximum limit of 10 rounds per question. The results of our data collection process prove to be efficient - we successfully obtain correct answers along with their corresponding decomposition trajectories for 84% of the training data, which represents a substantial improvement compared to the baseline accuracy of 58% achieved without implementing context sampling techniques.

Training Method Our training methodology is designed to optimize model performance by focusing exclusively on decomposition trajectories that successfully lead to correct answers. The data structure for each record follows a format that captures the complete problem-solving path: $(q, [(q^1, a^1), \dots, (q^t, a^t)], a)$, where q and a are the original question and its answer, respectively. Each pair (q^i, a^i) in the trajectory is composed of an atomic question and its answer. This structured approach ensures that we maintain a clear record of the decomposition process from start to finish.

For our LLM training, we employ supervised fine-tuning (SFT) to enhance model performance. Our process converts each trajectory into structured training data pairs of prompts and responses. These pairs are formatted as (x_i, y_i) according to Algorithm 3. The algorithm incorporates two key functions: $prompt_x$ and $prompt_y$, which generate contextual descriptions based on input parameters. These functions are integral to the transformation process, and we detail their implementation in the following sections.

Algorithm 3 Transform each decomposition trajectory into data pairs for SFT

```
1: Let  $(q, [(q^1, a^1), \dots, (q^t, a^t)], a)$  be the given decomposition trajectory.  
2: Initialize an empty list to store pairs of prompts and responses  $\mathcal{D}_{sft} = \emptyset$ .  
3: # Prepare training data when a further decomposition is necessary.  
4: for  $i = 1, 2, \dots, t$  do  
5:   #  $prompt_x$  takes the original question plus any previously collected sub-questions and  
   answers as input.  
6:    $x_i = prompt_x(q, [(q^1, a^1), \dots, (q^{i-1}, a^{i-1})])$   
7:   #  $prompt_y$  takes two inputs: a decomposition indicator and an optional sub-question.  
8:    $y_i = prompt_y(True, a^i)$   
9:    $\mathcal{D}_{sft} \leftarrow (x_i, y_i)$   
10: end for  
11: # Prepare training data when no further decomposition is necessary.  
12:  $x_{t+1} = prompt_x(q, [(q^1, a^1), \dots, (q^t, a^t)])$   
13:  $y_{t+1} = prompt_y(False, None)$   
14:  $\mathcal{D}_{sft} \leftarrow (x_{t+1}, y_{t+1})$ 
```

Prompt Template for $prompt_x$

Based on the given information, determine whether a follow-up question is necessary or not.

Original Question
{the original problem}

Existing Context
{A list of sub-questions and their answers}
Make sure your output align with the following format:
<decompose>False</decompose>
OR
<decompose>True</decompose>
<sub-question>a follow-up question</sub-question>

Prompt Template for $prompt_y$ if a sub-question is provided

<decompose>True</decompose>
<sub-question>{The given sub-question}</sub-question>

Prompt Template for $prompt_y$ if no sub-question is provided

<decompose>False</decompose>

This transformation is particularly efficient because it generates multiple training instances from each trajectory - specifically, for a trajectory containing t sub-questions, we generate $t + 1$ distinct training data points, maximizing the learning potential from each successful decomposition.

For all experiments, we use a learning rate of $1.5e^{-5}$ and parameter-efficient fine-tuning (PEFT) with LoRA configuration parameters (lora=16, alpha=64). To ensure reproducibility, we maintain all other hyperparameters at their default values as specified in reference [56]. We run all experiments three times with different random seeds and report the mean results. We perform the experiments on a computing node equipped with a single NVIDIA A100-80G GPU for all training processes.

Experiment Results Table 11 presents our experimental findings. We select three open-source LLMs with varying parameter sizes and computational capabilities as our base models for atomic proposers: meta-llama/Llama-3.1-8B [6], Qwen/Qwen2.5-14B [52], and microsoft/phi-4 [51]. Through SFT, we develop these into fine-tuned atomic proposers. For inference, we integrated two advanced LLMs-GPT-4o (version 2024-11-20) and Llama-3.1-70B-Instruct [6]-to process the retrieved data and generate the final answers. These models were chosen for their enhanced capabilities in delivering accurate, high-quality outputs.

For illustration, the second row and second column in Table 11 shows our use of the original meta-llama/Llama-3.1-8B as the atomic proposer for question decomposition. After gathering sufficient

Table 11: Evaluation Effects of Domain Aligned Atomic Proposers (Columns represent different generation LLMs while rows denote smaller LMs that we adopt as atomic decomposer.)

	GPT-4o	GPT-4o+FT	Llama-3.1-70B-Instruct	Llama-3.1-70B-Instruct+FT
Llama-3.1-8B	47.83%	62.14%	48.37%	58.70%
Qwen2.5-14B	56.52%	63.95%	57.61%	63.04%
phi-4-14B	60.33%	65.76%	58.70%	62.50%

context, we use GPT-4o to generate the final answer. In contrast, the second row and third column shows results when using a fine-tuned version of meta-llama/Llama-3.1-8B as the atomic proposer, while still using GPT-4o for final answer generation. The difference in values demonstrates how fine-tuning atomic proposers with domain-specific decomposition trajectories improves performance.

6.5 Real Case Studies

This section presents three case studies from our evaluation benchmark to illustrate the underlying principles of our proposed decomposition pipeline, as detailed in Algorithm 1. Through these real-world examples, we aim to highlight the benefits of our systematic approach. These cases will shed light on how each step of the pipeline contributes to improved performance and the insights gained from their implementation.

Our task decomposition strategy involves generating multiple atomic queries rather than producing a single deterministic follow-up question, as demonstrated in the Self-Ask approach. Contemporary decomposition methods typically employ a generative model to formulate a singular follow-up question. However, this approach carries an intrinsic risk of generating erroneous questions, potentially leading to an incorrect decomposition pathway and, ultimately, an erroneous answer. Consider the Case (a) depicted in Figure 16, where the original question pertains to a film titled "What Women Love." Due to the existence of a more prominent film, "What Women Want," the employed language model tends to 'correct' the original question. Consequently, methods like Self-Ask (as shown on the left side of Figure 16) generate only one follow-up question related to this erroneously assumed object. In the illustrated instance, although the target chunk has been retrieved due to the similarity in embeddings, a 'false' intermediate answer is produced for the 'false' follow-up question, culminating in an incorrect final response. In contrast, our methodology posits atomic queries concerning both "What Women Love" and "What Women Want," thereby seeking to clarify the true intent of the initial question. With both films existing and relevant atomic questions being retrieved, our approach subsequently gains the advantage of verifying the question's intent and selecting the correct and most pertinent chunk during the atomic selection phase.

Furthermore, the discrepancy between the formulation of the corpus and the query, is another critical factor advocating for a multi-query approach over a singular deterministic one. The presentation gap can impede the retrieval process even when the generated follow-up question is semantically accurate. For instance, as illustrated in Case (b) in Figure 17, a single-path method such as Self-Ask on the left side might directly inquire 'Who is the mother of Oskar Roehler?' However, the knowledge base articulates familial relationships using a different schema, 'A is the son of B and C' in this case, thus the retrieval process falters despite the correctness of the question. Even when we applied the hierarchical retrieval to Self-Ask, the Self-Ask with Hierarchical Retrieval did not succeed in bridging this gap. In contrast, our approach, which generates multiple atomic queries, encompasses a broader range of phrasings that correspond to the diverse representations in the knowledge base. In the depicted case, while the atomic query specifically asking for Oskar Roehler's mother encounters the same retrieval issue, an alternative query seeking information about his parents successfully retrieves the target chunk. This exemplifies how our method's flexibility in query generation enhances the likelihood of aligning with the knowledge base's structure and obtaining accurate information.

Our methodology emphasizes the retrieval of atomic questions rather than directly retrieving chunks. This design choice is exemplified in Case (b) depicted in Figure 17. The knowledge chunk in the corpus is structured using the pattern 'A ... as the son of B and C', which poses challenges for direct retrieval by queries such as 'Who is the mother of ...'. In our specialized knowledge base, such direct queries tend to retrieve chunks conforming to the patterns 'A is the mother of B' or 'A is the father of B'. By utilizing atomic questions as intermediaries for retrieval, our approach effectively narrows

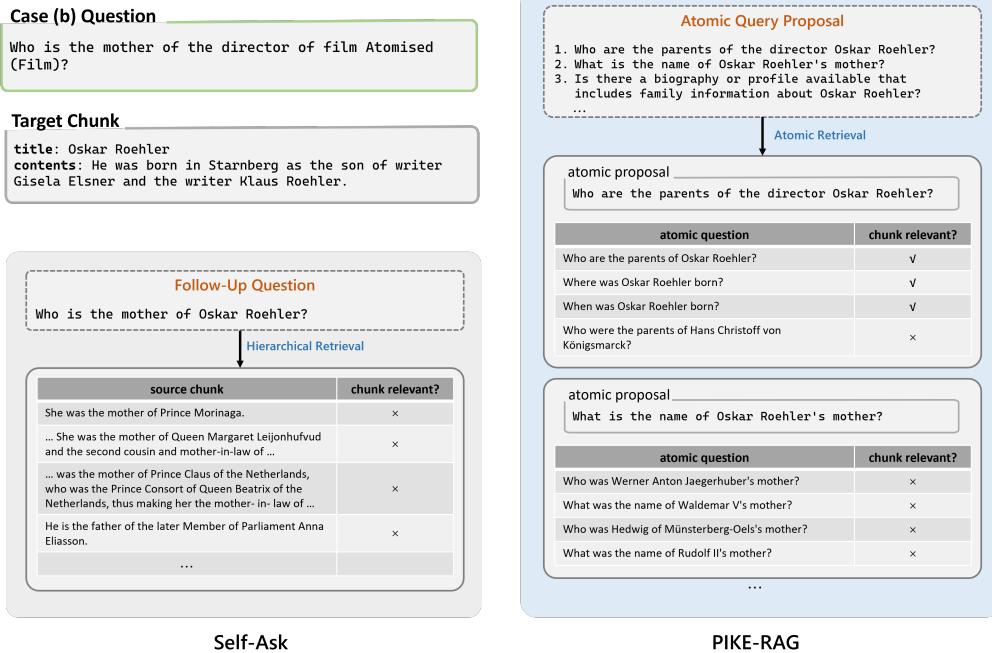


Figure 17: Case (b): By proposing multiple atomic queries, PIKE-RAG effectively retrieves the relevant knowledge chunk, whereas the single deterministic follow-up question approach employed by Self-Ask fails to align with the knowledge base’s schema, resulting in a retrieval failure.

the gap between a single query and the multiple sentence structures found in the knowledge base. It facilitates bridging the expression pattern differences exemplified by ‘the mother of’ versus ‘the son of’ in this scenario.

In contrast to methods like Self-Ask, which only retains intermediate answers for subsequent processing, our method preserves the entire chunk as contextual information. During the atomic selection phase, we present a list of atomic questions as candidate summaries of the relevant content from the original chunk. This strategy significantly reduces token usage and simplifies the process of selecting the pertinent information. Case (c) in Figure 18 demonstrates the dual benefits of our approach: first, by selecting from a curated list of atomic questions, we streamline the identification of relevant information; second, by retaining the entire selected chunk rather than just the intermediate answer, we ensure a rich context is maintained for accurate and comprehensive subsequent processing. While the Self-Ask method on the left retrieves the target chunk, it fails to correctly identify the pertinent ‘Ernie Watts’ due to the excessive contextual information. Since retrieved chunks in Self-Ask are discarded after generating an intermediate answer, the method potentially follows an incorrect pathway, leading to an inaccurate conclusion. In contrast, our approach can efficiently filter and select the appropriate atomic question from a concise list. Although the atomic question in this round pertains to the role of Ernie Watts, there is no need to inquire further about his birthplace, as this information is encapsulated within the selected chunk, which remains available for context in subsequent rounds.

7 Conclusion

To address the diverse challenges faced by RAG systems in industrial applications, we propose that the core foundation of RAG systems should extend beyond traditional retrieval mechanisms to the effective construction and utilization of specialized knowledge and rationale. Therefore, we introduce a new paradigm that classifies tasks based on their difficulty in knowledge extraction, comprehension, and utilization, providing a novel framework for system design and evaluation. Applying this paradigm allows for phased exploration of RAG capabilities, which facilitates the progressive refinement of RAG algorithms and the staged implementation of RAG applications. Moreover, we introduce the specialized Knowledge and Rationale Augmented Generation (PIKE-

Case (c) Question

In which district was Ernie Watts born?

Target Chunk

title: Ernie Watts (footballer, born 1872)
contents: Born in Woolhampton in Berkshire, his earliest known club was Reading, before he moved to Notts County, where he made 17 appearances in The Football League. He had a second spell with Reading and also played for ...

Follow-Up Question: Who is Ernie Watts?

Hierarchical Retrieval

source chunk	chunk relevant?
Born in Woolhampton in Berkshire, his earliest known club was Reading, before he moved to Notts County, where he made 17 appearances in The Football ...	v
Gentle Thoughts is ... It features Ernie Watts, Dave Grusin, Patrice Rushen, Anthony Jackson, ...	x
...	

Generation

Intermediate Answer: Ernie Watts is a musician who played on the album Gentle Thoughts.

Follow-Up Question: What is the birthplace of Ernie Watts?

Self-Ask

Atomic Query Proposal

1. Who is Ernie Watts?
2. What is the full name of Ernie Watts?
3. What is the date of birth of Ernie Watts?

...

Atomic Retrieval

atomic proposal

Who is Ernie Watts?

atomic question	chunk relevant?
Where was Ernie Watts born?	v
...	

Atomic Selection

```
"selected": true,
"rationale": ...,
"chosen_info": {
  "atomic_question": "Where was Ernie Watts born?",
  "source_chunk": "Born in Woolhampton in Berkshire, his earliest known club was Reading, before he moved to Notts County, where he made 17 appearances in The Football League. He had a second spell with Reading and also played for ... "
}
```

Generation

```
"Quesiton": ...
"Context": {
  "source_chunk": ...
}
```

PIKE-RAG

Figure 18: Case (c): PIKE-RAG’s benefits from leveraging a concise list of atomic questions for targeted selection and retaining full chunks for rich contextual support. Conversely, Self-Ask’s approach, although successful in retrieving relevant chunks, is compromised by its dependency on intermediate answers for context, which ultimately results in the generation of incorrect final answers.

RAG) framework, focusing on specialized knowledge extraction and rationale construction. PIKE-RAG effectively extracts, comprehends, and organizes specialized knowledge and construct coherent rationale for accurate answers, offering customizable system capabilities to meet varying requirements. Additionally, we propose knowledge atomizing and knowledge-aware task decomposition to tackle complex questions, such as multihop queries, achieving significant performance improvements on various open-domain and legal benchmarks.

References

- [1] Document intelligence. <https://azure.microsoft.com/en-us/products/ai-services/ai-document-intelligence>. Accessed: 2024-08-25.
- [2] paddlepaddle. <https://github.com/PaddlePaddle/PaddleOCR>. Accessed: 2024-08-25.
- [3] Tesseract ocr. <https://github.com/tesseract-ocr/tesseract>. Accessed: 2024-08-25.
- [4] Textract. <https://aws.amazon.com/cn/textract/>. Accessed: 2024-08-25.
- [5] J Achiam, S Adler, S Agarwal, L Ahmad, I Akkaya, FL Aleman, D Almeida, J Altenschmidt, S Altman, S Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [6] AI@Meta. Llama 3 model card. 2024.
- [7] R Anil, S Borgeaud, Y Wu, J-B Alayrac, J Yu, R Soricut, J Schalkwyk, AM Dai, A Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [8] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023.
- [9] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.
- [10] Aram Bahrini, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J Riggs, Maryam Esmaeili, Rastin Mastali Majdabadkohne, and Morteza Pasehvar. Chatgpt: Applications, opportunities, and threats. In *2023 Systems and Information Engineering Design Symposium (SIEDS)*, pages 274–279. IEEE, 2023.
- [11] Iz Beltagy, Arman Cohan, and Kyle Lo. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, 2020.
- [12] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. ACM, 2021.
- [13] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens, 2022.
- [14] T Brown, B Mann, N Ryder, M Subbiah, JD Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [15] Umar Butler. Open australian legal qa, 2023.
- [16] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms, 2024.
- [17] Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. Dense x retrieval: What retrieval granularity should we use? *arXiv preprint arXiv:2312.06648*, 2023.
- [18] Wenhua Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text, 2022.
- [19] Cohere. Say goodbye to irrelevant search results: Cohere rerank is here. <https://txt.cohere.com/rerank/>, 2023. Accessed: 2023-08-28.
- [20] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.
- [21] 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, 2024.

- [22] Yu Fang and other authors. Hierarchical graph network for multi-hop question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2020.
- [23] Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*, 2023.
- [24] Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. Retrieval-generation synergy augmented large language models, 2023.
- [25] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels, 2022.
- [26] Y Gao, Y Xiong, X Gao, K Jia, J Pan, Y Bi, Y Dai, J Sun, and H Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.
- [27] Yunfan Gao, Yun Xiong, Meng Wang, and Haofen Wang. Modular rag: Transforming rag systems into lego-like reconfigurable frameworks, 2024.
- [28] Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. WikiAsp: A Dataset for Multi-domain Aspect-based Summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225, 2021.
- [29] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*, 2020.
- [30] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models, 2022.
- [31] H. Jiang, Q. Wu, X. Luo, D. Li, C.-Y. Lin, Y. Yang, and L. Qiu. Longllmilingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839*, 2023.
- [32] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation, 2023.
- [33] Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.
- [34] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp, 2023.
- [35] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- [36] P Lewis, E Perez, A Piktus, F Petroni, V Karpukhin, N Goyal, H Kuttler, M Lewis, WT Yih, T Rocktaschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [37] Ruosen Li and Xinya Du. Leveraging structured information for explainable multi-hop question answering and reasoning, 2023.
- [38] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhui Chen. Long-context llms struggle with long in-context learning, 2024.
- [39] Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhagash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, Carl Yang, and Liang Zhao. Domain specialization as the key to make large language models disruptive: A comprehensive survey, 2024.
- [40] X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*, 2023.

- [41] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.
- [42] Feifei Pan, Mustafa Canim, Michael Glass, Alfio Gliozzo, and James Hendler. End-to-end table question answering via retrieval-augmented generation, 2022.
- [43] Pranoy Panda, Ankush Agarwal, Chaitanya Devaguptapu, Manohar Kaul, and Prathosh A P. Holmes: Hyper-relational knowledge graphs for multi-hop question answering using llms, 2024.
- [44] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models, 2023.
- [45] Minghui Qiu and other authors. Dynamically fusing recurrent neural networks for multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2019.
- [46] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [47] O Ram, Y Levine, I Dalmedigos, D Muhlgay, A Shashua, K Leyton-Brown, and Y Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.
- [48] Sohini Roychowdhury, Marko Krema, Anvar Mohammad, Brian Moore, Arijit Mukherjee, and Punit Prakashchandra. Eratta: Extreme rag for table to answers with large language models, 2024.
- [49] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy, 2023.
- [50] Jiahuo Sun, Chengjin Xu, Lumingyan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph, 2024.
- [51] Phi-4 Team. Phi-4 technical report, 2024.
- [52] Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- [53] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [54] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.
- [55] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions, 2023.
- [56] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>, 2020.
- [57] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. Freshllms: Refreshing large language models with search engine augmentation, 2023.
- [58] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity, 2023.
- [59] Xiantao Wang, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao, and Wei Wang. Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases, 2023.

- [60] Yu Wang, Nedim Lipka, Ryan A. Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. Knowledge graph prompting for multi-document question answering, 2023.
- [61] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models, 2024.
- [62] S. Yang. Advanced rag 01: Small-to-big retrieval. <https://towardsdatascience.com/advanced-rag-01-small-to-big-retrieval-172181b396d4>, 2023. Accessed: 2023-08-28.
- [63] Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen tau Yih, and Xin Luna Dong. Crag – comprehensive rag benchmark, 2024.
- [64] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.
- [65] Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Yong Liu, and Shen Huang. End-to-end beam retrieval for multi-hop question answering, 2024.
- [66] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, and B. Cui. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, 2024.
- [67] H. S. Zheng, S. Mishra, X. Chen, H.-T. Cheng, E. H. Chi, Q. V. Le, and D. Zhou. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*, 2023.
- [68] Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models, 2024.
- [69] Zijie Zhong, Hanwen Liu, Xiaoya Cui, Xiaofan Zhang, and Zengchang Qin. Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation, 2024.