香港大學
THE UNIVERSITY OF HONG KONG

从检索增强生成到图增强生成:探索新一代智能问答系统

郭子睿

2024.12.17

001-RAG(检索增强生成)

RAG(检索增强生成)

检索增强生成（Retrieval-Augmented Generation，RAG）是一种结合了检索与生成的大语言模型增强方法，旨在通过引入外部知识库提高生成内容的准确性和上下文相关性。RAG框架包括三个核心组成部分：检索（Retrieval）、增强（Augmentation）和生成（Generation）。

R

检索（Retrieval）

负责从外部知识库或文档集合中找到与用户查询最相关的信息。通过将大规模文本数据进行切块并向量化存储，系统可以根据用户输入生成的查询向量，高效地检索出最匹配的内容。

A

增强（Augmentation）

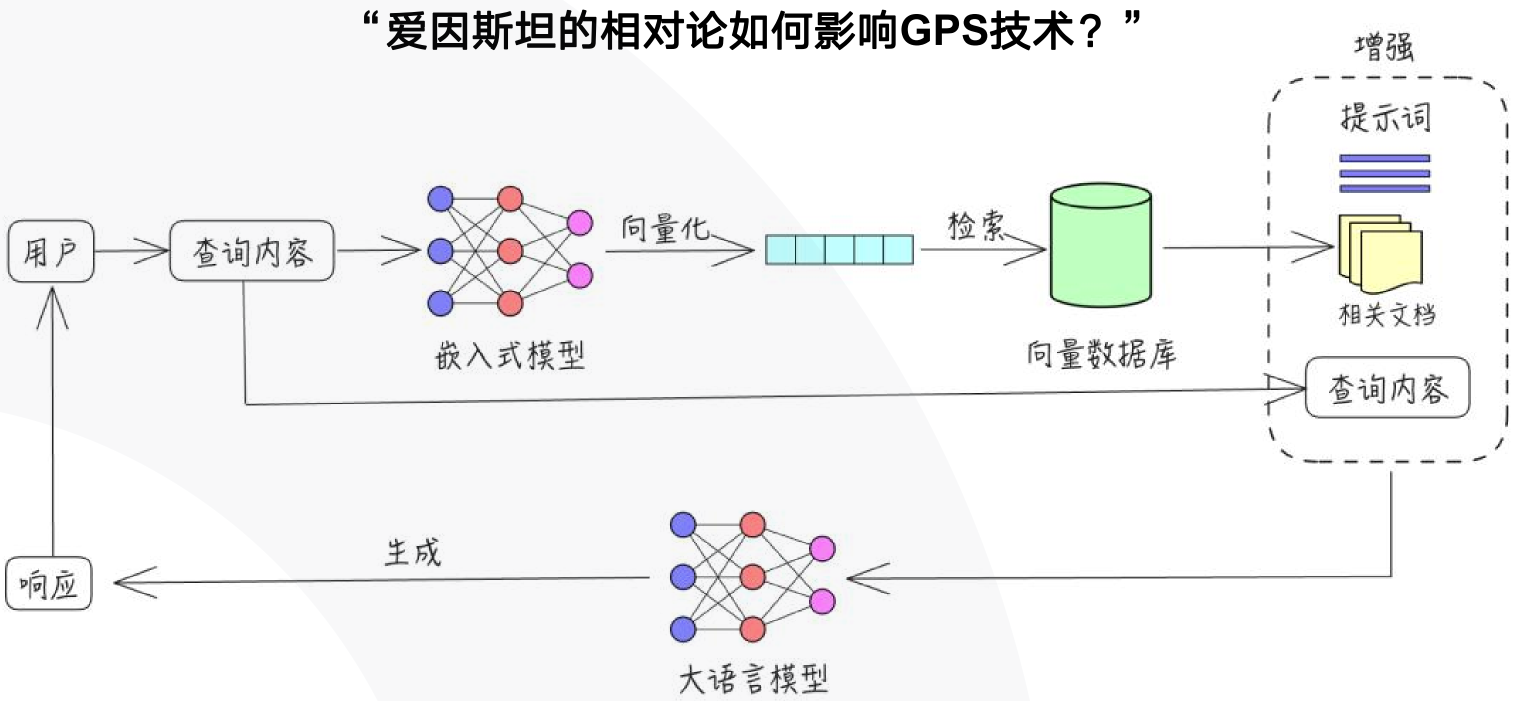
将检索到的相关信息作为额外的上下文，输入到语言模型中，帮助生成模型更好地理解问题并提供更准确的答案。

G

生成（Generation）

由大语言模型（LLM）负责完成，它基于用户的查询以及检索到的增强信息，生成逻辑连贯、上下文相关且高质量的回答。通过结合检索的外部知识和模型自身的语言生成能力，最终输出回答。

002-RAG流程图



1.检索阶段（Retrieval）
系统将用户的查询进行向量化，检索外部知识库，找到与问题相关的内容。例如，系统检索到以下信息片段：

- “爱因斯坦的广义相对论指出，由于重力影响，时间在不同高度的钟走得不一样。”
- “GPS卫星的时钟需要考虑相对论效应，否则每天会出现数公里的定位误差。”

2.增强阶段（Augmentation）
系统将检索到的关键信息与用户的原始查询结合，作为上下文传递给生成模型。输入如下：

查询：爱因斯坦的相对论如何影响GPS技术？
检索上下文：
“广义相对论指出重力影响时间流速，卫星与地面时钟会有差异。”
“GPS系统必须进行相对论时间校正，以保证定位的精确度。”

3.生成阶段（Generation）
最终，大语言模型结合用户的查询和增强的上下文，生成回答：

“爱因斯坦的相对论（特别是广义相对论）指出，由于地球重力场的影响，GPS卫星上的时钟与地面时钟的运行速度不同。为了确保定位准确，GPS系统必须根据相对论对时间进行校正，否则每天会产生约10公里的误差。因此，相对论是GPS技术实现高精度定位的重要理论基础。”

图源：https://www.gptsecurity.info/2024/05/26/RAG/

003-RAG 的优势：缓解大模型落地应用痛点

垂直领域知识的幻觉

通过检索外部权威知识库，RAG 可以提供更准确和可靠的领域特定知识，减少生成幻觉的可能性。

整合长尾语义知识

RAG 能够从广泛的知识库中检索长尾语义知识，从而生成更丰富和全面的响应。

复杂问题的推理能力

支持跨文档、跨领域信息的检索与整合，增强多步推理能力。

灵活适应多领域需求

通过更换或扩展知识库，RAG 能快速适应不同垂直领域的需求。

知识持续更新的困难

无需重新训练模型，RAG 可以通过访问最新的外部知识库，保持输出的时效性和准确性。

减少隐私泄露风险

通过使用外部知识库而不是依赖内部训练数据，RAG 减少了隐私泄露的风险。

提供可解释性

RAG 可以明确区分答案的来源，提高生成系统的透明度。

兼容性强，易于集成

RAG 的模块化架构便于与现有检索系统和生成模型无缝整合。

004-Naive RAG的局限

1

依赖平面数据表示，难以捕捉复杂关系

Naive RAG以向量化的文本切块（chunk）为基础，忽视了实体和关系之间的语义关联。检索结果仅限于表面匹配，无法反映跨文档或跨主题的复杂依赖关系。

2

缺乏上下文感知，导致回答不连贯

检索与生成之间是线性连接，检索的结果以独立片段形式直接输入生成模型，缺少对全局上下文的整合。

3

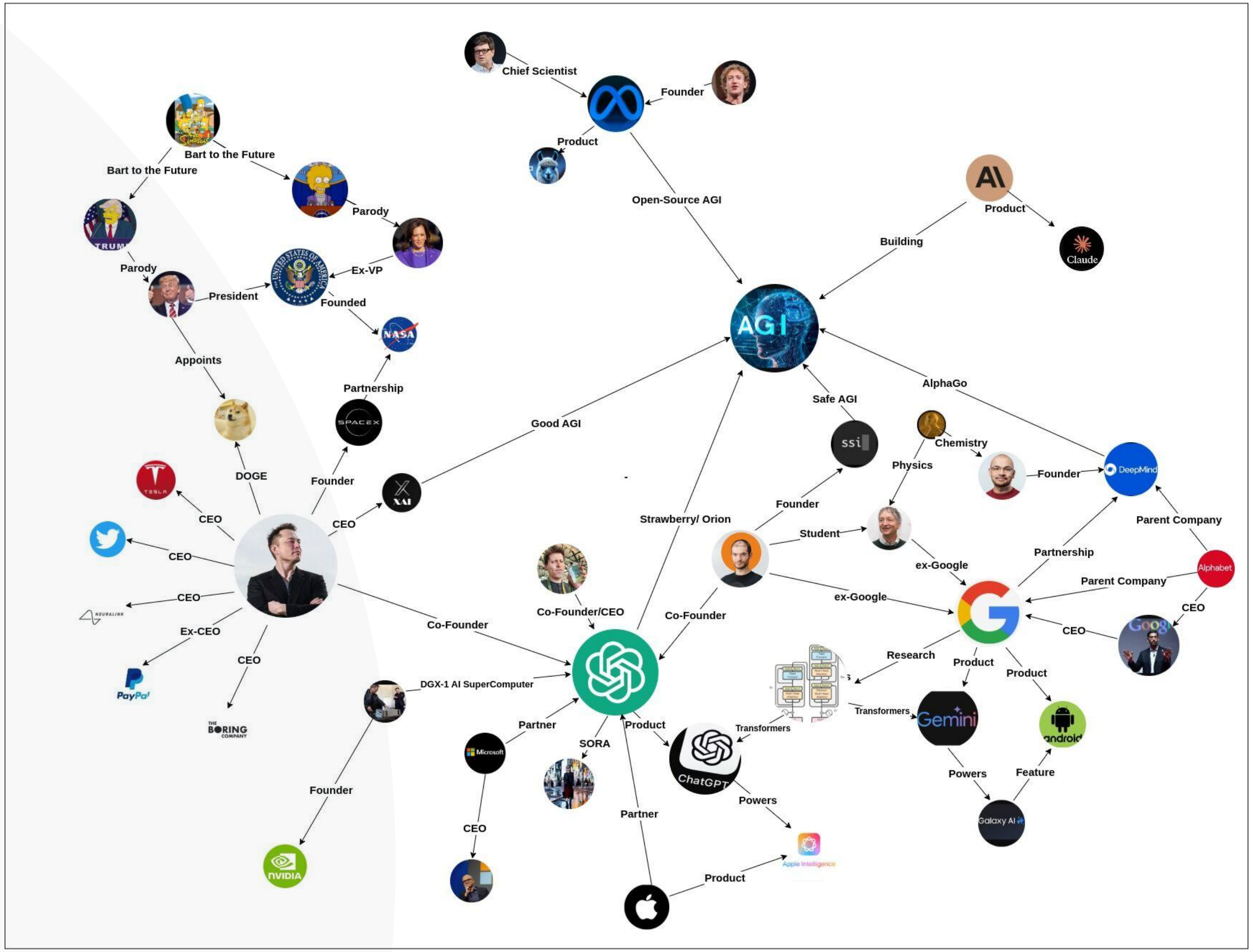
文本块中信息冗余，大量与查询无关内容

Naive RAG检索的文本块中往往包含大量无关信息，这不仅增加了生成模型的计算负担，也可能降低回答的精确性和相关性。

005-Graph的引入——GraphRAG

微软推出了一种名为 GraphRAG 的新方法，**将图形结构融入文本索引和检索过程。**

知识图谱是由一组节点组成的数据结构，这些节点保存了存在于各个数据点上的不同实体之间的关系。结构化知识图谱使 GraphRAG能够通过连接点或对比信息片段在多跳推理中表现出色。

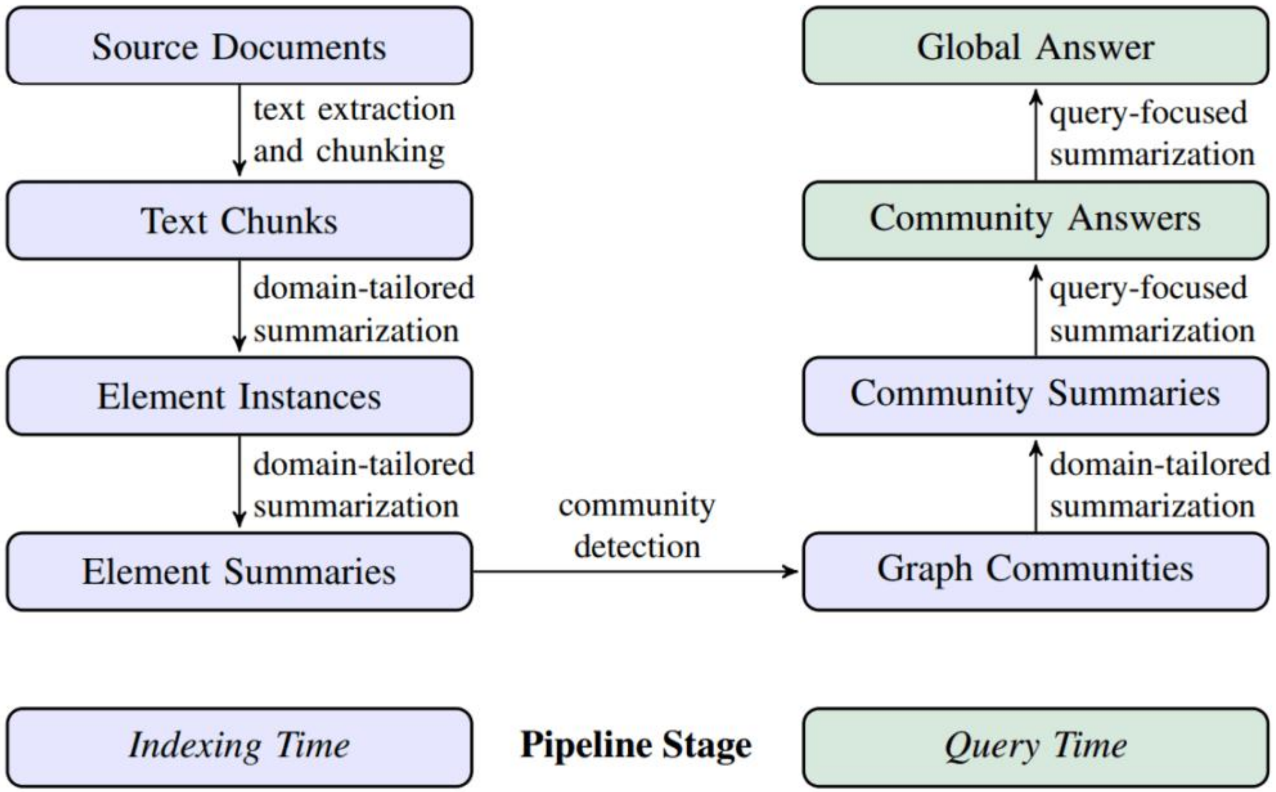


006-GraphRAG——索引阶段

为了构建结构化索引 KG，GraphRAG 使用 LLM 从源文档中提取实体和关系。

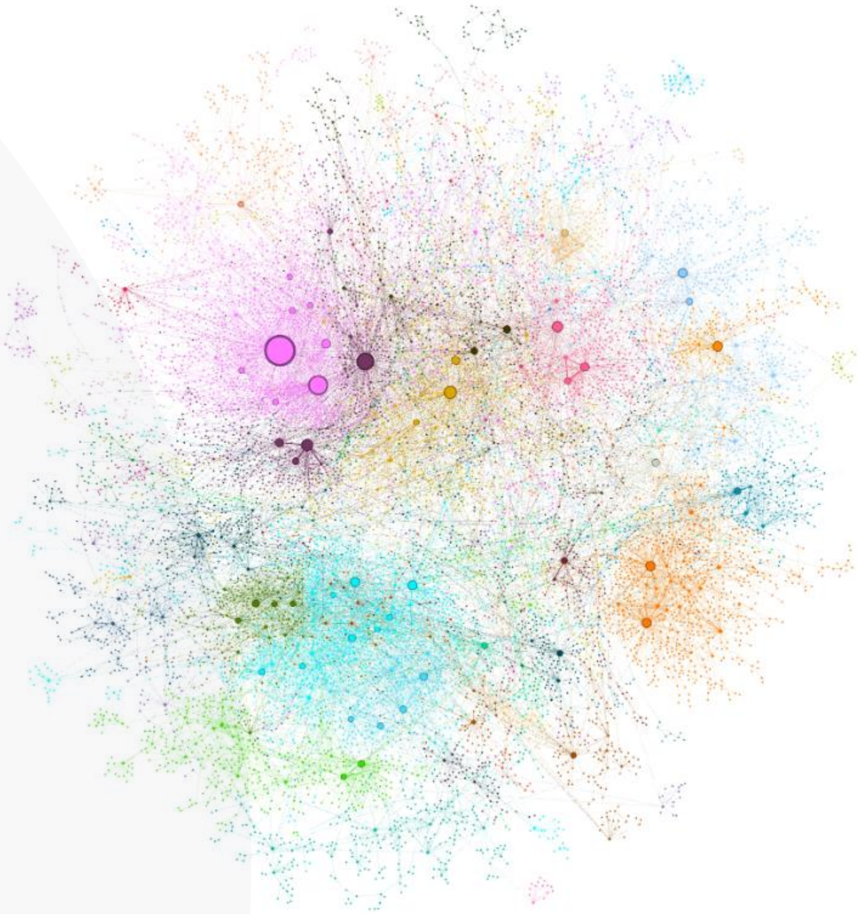
实体表示为 KG 中的单个节点，包含有关名称、组织或类别的信息以及有关实体的简要描述。

关系表示为具有源和目标实体 ID 的不同节点之间的边。这定义了实体之间的关系，并使用连接权重（即强度 分数，范围在 1 到 10 之间）对它们进行评级。



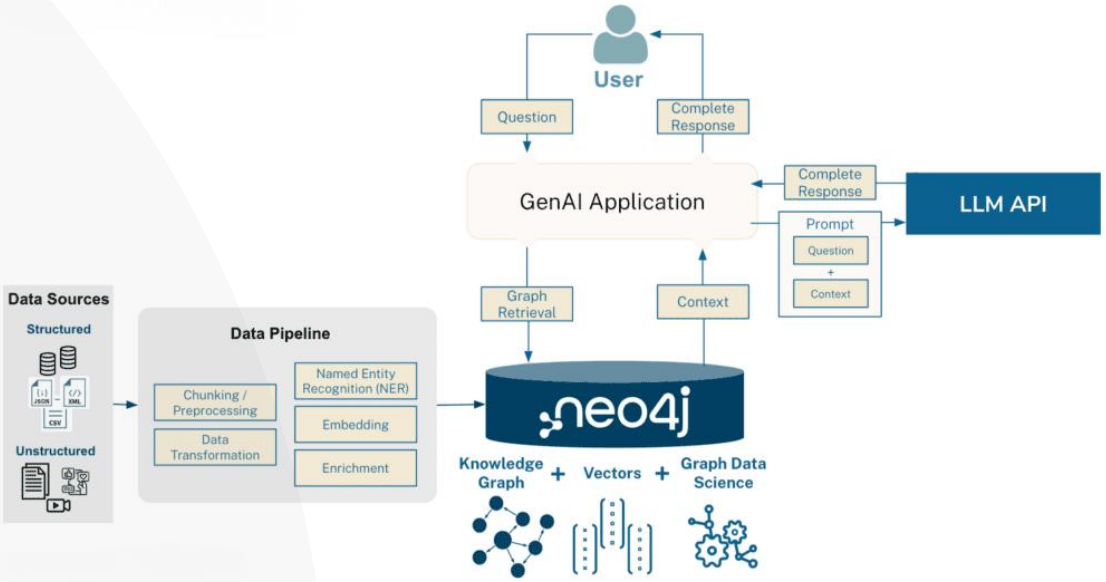
007-GraphRAG——索引阶段

在下一阶段，基于节点连接密度和可扩展性，应用Leiden算法通过将紧密相关的节点分组为层次化集群来发现模块化社区。并为所有的社区生成对应的社区报告，来总结社区内所有实体以及关系，为查询做准备。

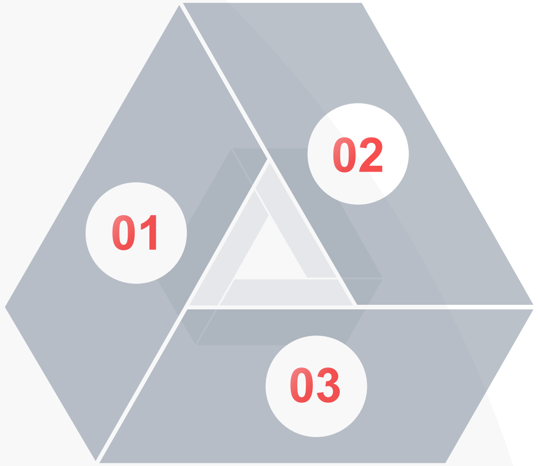


008-GraphRAG——查询阶段

在查询阶段，当用户提出问题时，系统会**遍历所有社区**，并基于社区报告由LLM**生成不同的中间响应**，同时为每个响应分配一个介于0到100之间的有用性分数，该分数表示生成答案与用户查询的相关性。最终，系统根据有用性分数对结果按降序排列，**将相关社区报告作为LLM的上下文输入**，直至达到上下文容量上限。



009-GraphRAG——局限性



成本高昂

在检索阶段，需要为每个社区生成对应的社区报告；在查询阶段，又需遍历所有社区报告以生成回答，这使得整体运行成本极高。

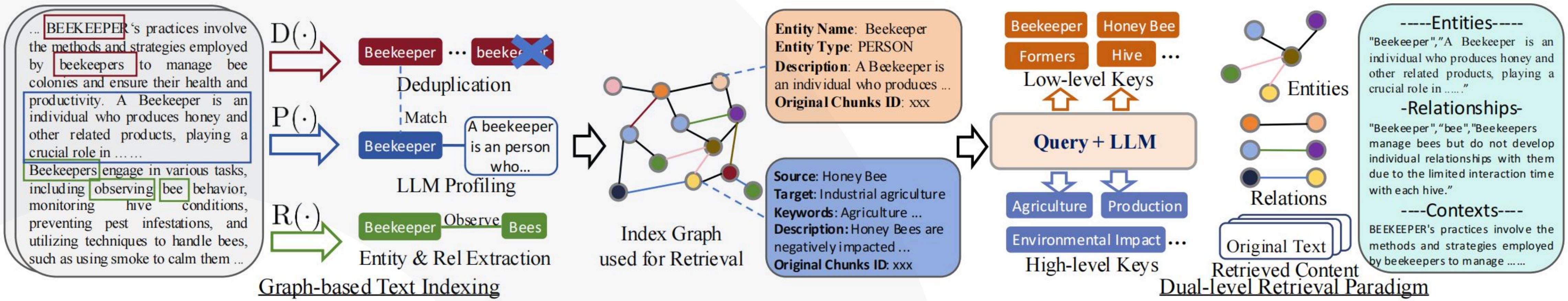
运行速度慢

无论是社区报告的生成还是社区的遍历，都需要耗费大量时间，显著降低了系统的效率。

扩展性差

为将新数据合并到现有图中，必须对先前的数据重新构建社区，这大幅增加了新增数据的处理成本，限制了系统的动态扩展能力。

010-LightRAG——简单、快速、高效



简单

LightRAG减少了社区检测和报告生成的资源开销，优化了检索和生成阶段的计算成本。

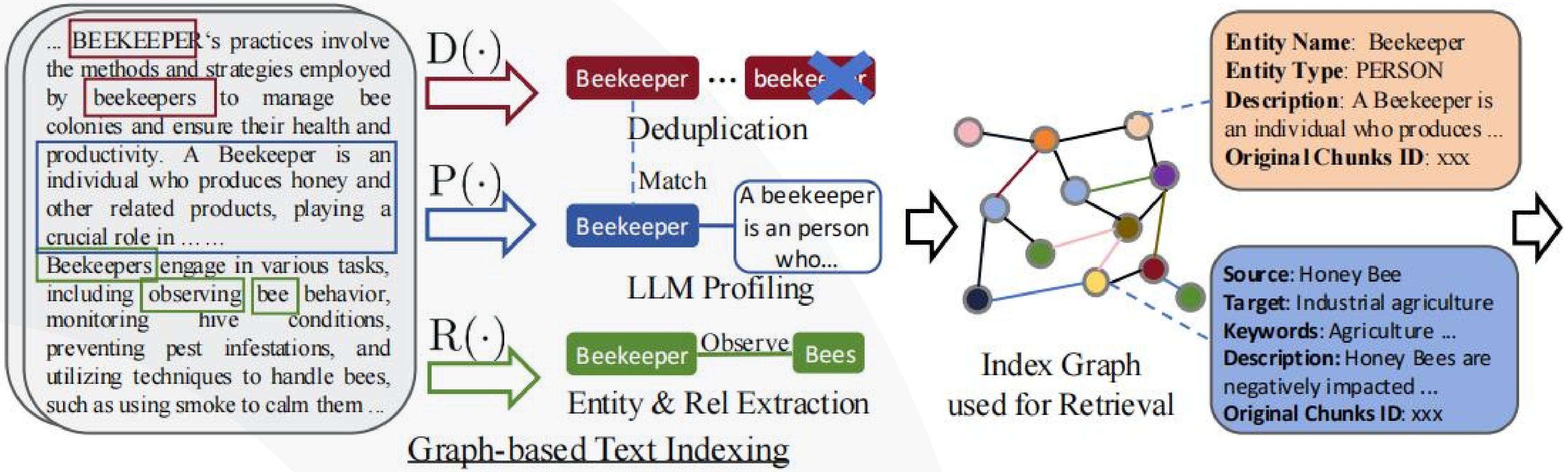
快速

在检索和生成阶段引入双层检索策略，通过低层检索定位具体实体和关系，高层检索抓取全局主题，显著提高信息检索效率。

高效

无需重建整个知识图谱即可快速整合新数据，降低动态更新的时间成本。

011-LightRAG——键值对索引



步骤 1：提取实体和关系

将长文本分割文本块，逐块使用LLM对文本数据进行抽取。提取的**实体**包括名称、地点、事件等，而**关系**则是这些实体之间的语义联系，如“属于”“包含”“依赖”等。

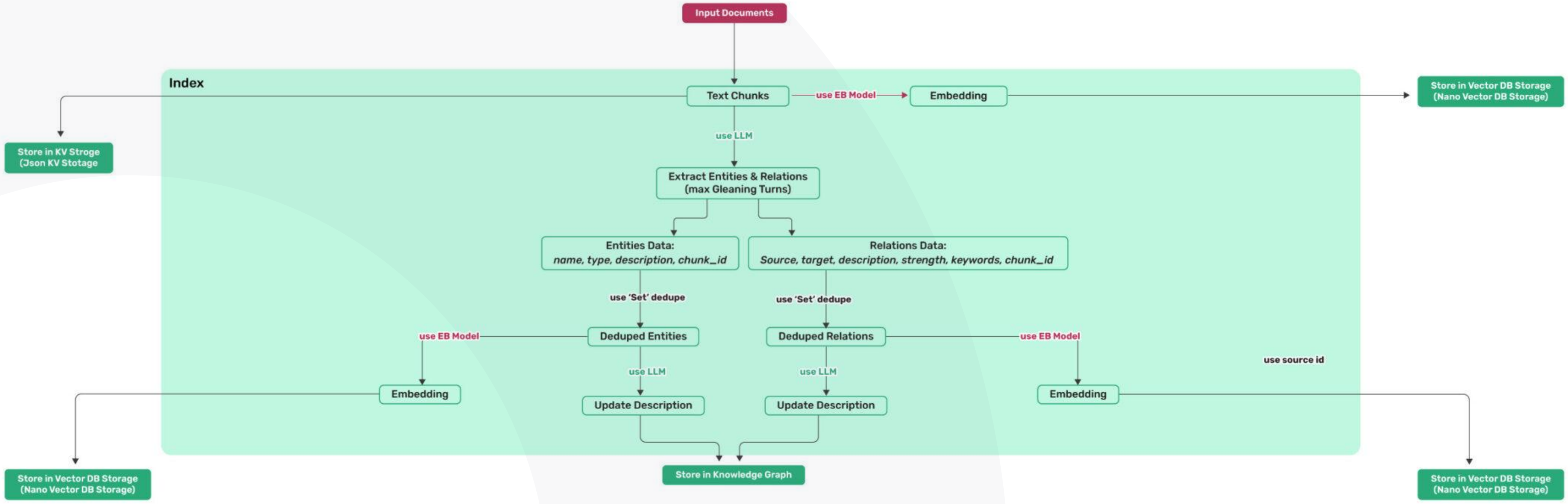
步骤 2：LLM 生成键值对

在完成实体和关系提取后，利用LLM为图谱中的每个实体和关系成结构化的键值对。键是**实体名称**或**关系关键词**，而值则是由LLM生成的**描述性文本**，提供关于该实体或关系的详细背景信息。

步骤 3：重复数据删除

在图谱构建完成后，为了减少冗余并提高查询和生成的效率，系统会对图中的节点进行去重，合并所有重名节点以及对应的值。

012-LightRAG——键值对索引



013-LightRAG——键值对索引

-Goal-
Given a text document that is potentially relevant to this activity and a list of entity types, identify all entities of those types from the text and all relationships among the identified entities.

-Steps-
1. Identify all entities. For each identified entity, extract the following information:
- **entity_name**: Name of the entity, capitalized
- **entity_type**: One of the following types: [organization, person, geo, event]
- **entity_description**: Comprehensive description of the entity's attributes and activities

Format each entity as ("entity"<|><entity_name><|><entity_type><|><entity_description>)

2. From the entities identified in step 1, identify all pairs of (source_entity, target_entity) that are *clearly related* to each other.
For each pair of related entities, extract the following information:
- **source_entity**: name of the source entity, as identified in step 1
- **target_entity**: name of the target entity, as identified in step 1
- **relationship_description**: explanation as to why you think the source entity and the target entity are related to each other
- **relationship_strength**: a numeric score indicating strength of the relationship between the source entity and target entity
- **relationship_keywords**: one or more high-level key words that summarize the overarching nature of the relationship, focusing on concepts or themes rather than specific details

Format each relationship as ("relationship"<|><source_entity><|><target_entity><|><relationship_description><|><relationship_keywords><|><relationship_strength>)

3. Identify high-level key words that summarize the main concepts, themes, or topics of the entire text. These should capture the overarching ideas present in the document.
Format the content-level key words as ("content_keywords"<|><high_level_keywords>)

4. Return output in English as a single list of all the entities and relationships identified in steps 1 and 2. Use **## as the list delimiter.**

5. When finished, output <|COMPLETE|>

-Real Data-
Entity_types: {entity_types}
Text: {input_text}

Output:

Graph Construct Prompt

neo4j aura New Organization / New tenant

Send feedback Learn J Jayakumaran R

Instance: Instance01 Database: neo4j User: neo4j Status: Connected

neo4j$

Database information

Nodes (3,352)

* [ENTITY_TYPE] <"ACTIONS AGAINST EX... <"LEGAL ACTIONS AGAI... <"PENALTY FOR CONTR... <ENTITY_TYPE> AGRICULTURE ARTICLE ARTIFACT ASSET CHAPTER COMMUNICATION COMPANY TYPE CONCEPT CONTRACT CRIMINAL RECORD DOCUMENT EMPLOYMENT STATUS ENTITY_TYPE ENVIRONMENTAL SERVI... EVENT EVENT/DOCUMENT EVENT/LOCATION FINANCE FINANCIAL CONDITION FINANCIAL STATEMENT ... GEO GOODS GROUP HOUSING DEVELOPMENT INDUSTRY INVESTMENT LANGUAGE LAW LEGAL DOCUMENT LEGAL PROVISION LEGISLATION LOAN LOCATION MINING INDUSTRY MISSION OBJECT ORGANIZATION ORGANIZATION, PERSON ORGANIZATION/PERSON PART OF THE CONSTITU... PENALTY

Last update: 18:05:16

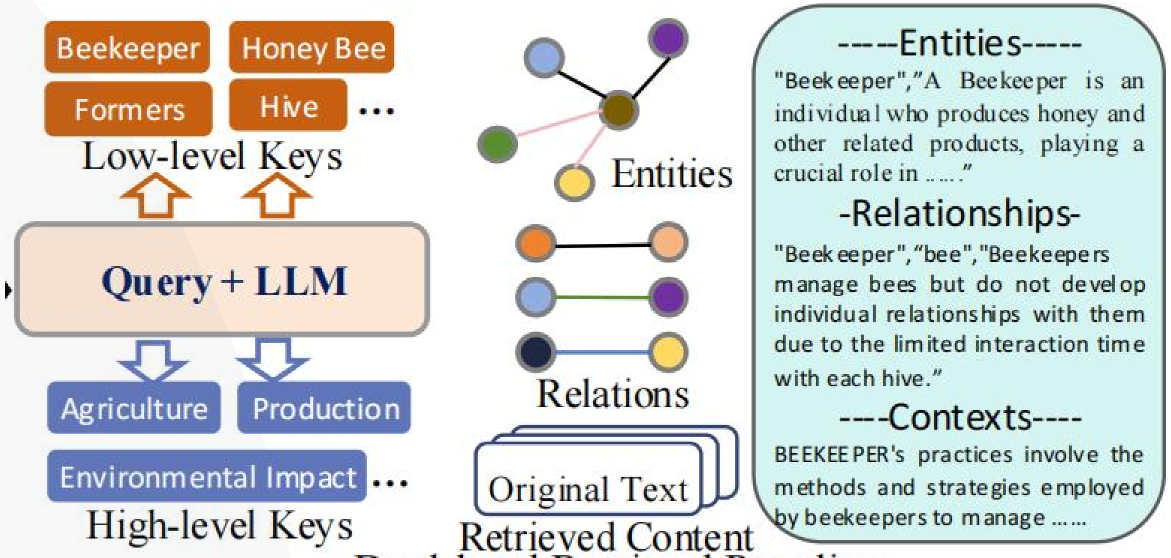
014-LightRAG——双层检索范式

LightRAG首先根据用户的查询抽取出**低层和高层**关键词。

在**低层检索**中，根据低层关键词，从知识图谱中定位与查询直接相关的具体实体，提取出与查询最匹配的实体节点及其邻接关系。

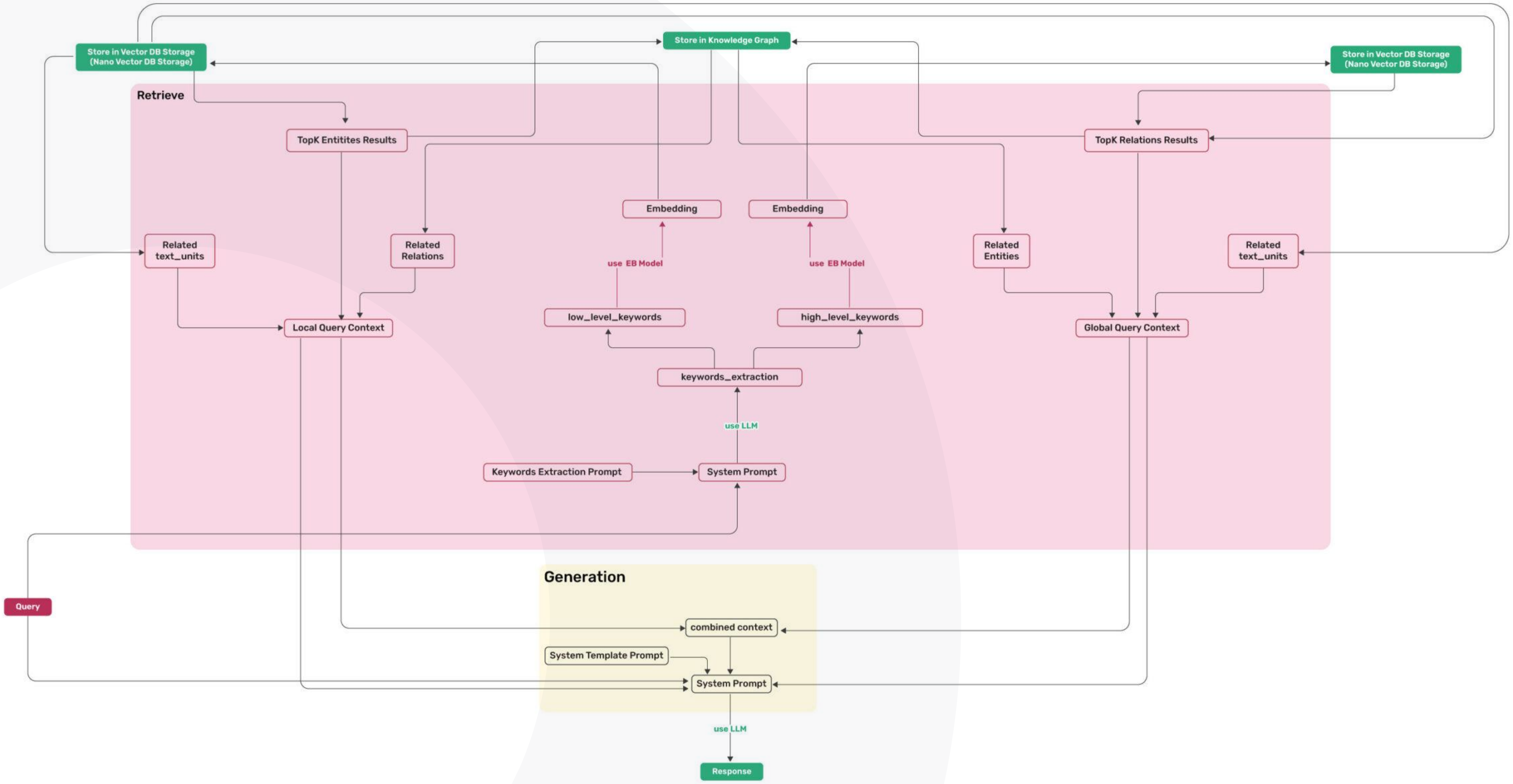
而**高层检索**则更侧重于全局性和抽象性的主题信息，根据高层关键词，检索相关的关系以及对应的实体，提取出更广泛的上下文。

通过这种双层次检索的结合，能够同时兼顾细节的精确性和全局的全面性，最终提供信息完整、逻辑清晰的答案。



Dual-level Retrieval Paradigm

015-LightRAG——双层检索范式



016-LightRAG——双层检索范式

---Role---
You are a helpful assistant tasked with identifying both high-level and low-level keywords in the user's query.

---Goal---
Given the query, list both high-level and low-level keywords. **High-level keywords focus on overarching concepts or themes, while low-level keywords focus on specific entities, details, or concrete terms.**

Keywords Generate Instruction Prompt

- Output the keywords in **JSON format**.
- The JSON should have two keys:
- **"high_level_keywords" for overarching concepts or themes.**
- **"low_level_keywords" for specific entities or details.**

-Examples-

Example 1:
Query: "How does international trade influence global economic stability?"
Output: {{ "**high_level_keywords**": ["International trade", "Global economic stability", "Economic impact"], "**low_level_keywords**": ["Trade agreements", "Tariffs", "Currency exchange", "Imports", "Exports"] }}

Example 2:
Query: "What are the environmental consequences of deforestation on biodiversity?"
Output: {{ "**high_level_keywords**": ["Environmental consequences", "Deforestation", "Biodiversity loss"], "**low_level_keywords**": ["Species extinction", "Habitat destruction", "Carbon emissions", "Rainforest", "Ecosystem"] }}

Example 3:
Query: "What is the role of education in reducing poverty?"
Output: {{ "**high_level_keywords**": ["Education", "Poverty reduction", "Socioeconomic development"], "**low_level_keywords**": ["School access", "Literacy rates", "Job training", "Income inequality"] }}

-Real Data-
Query: {query}
Output:

Keywords Generate Input Prompt

017-LightRAG——评估数据集

Table 4: Statistical information of the datasets.

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

UltraDomain 数据来源于 428 本大学教材，涵盖了农业、社会科学和人文学科等 18 个不同领域。其中，选取了 **Agriculture**（农业）、**CS**（计算机科学）、**Legal**（法律） 和 **Mix**（混合） 这四个数据集作为评估使用，数据集包含 60 万至 500 万个 token。

018-LightRAG——查询生成

Given the following description of a dataset: **{total_description}**

Please identify **5 potential users** who would engage with this dataset. For each user, list **5 tasks** they would perform with this dataset. Then, for each (user, task) combination, generate **5 questions** that require **a high-level understanding of the entire dataset.**

Output the results in the following structure:
- **User 1**: [user description]
  - **Task 1**: [task description] [**Question 1**: {Question 1}, **Question 2**: {Question 2}, **Question 3**: {Question 3}, **Question 4**: {Question 4}, **Question 5**: {Question 5}]
  - **Task 2**: [task description] [**Question 1**: {Question 1}, **Question 2**: {Question 2}, **Question 3**: {Question 3}, **Question 4**: {Question 4}, **Question 5**: {Question 5}]
  - ...
  - **Task 5**: [task description] [**Question 1**: {Question 1}, **Question 2**: {Question 2}, **Question 3**: {Question 3}, **Question 4**: {Question 4}, **Question 5**: {Question 5}]
- **User 2**: [user description]
  ...
- **User 5**: [user description]
  ...

Query Generate Prompt

为了评估RAG系统在高层次语义感知中的有效性，我们参考GraphRAG中的做法，指示LLM生成**五个RAG用户**，并为每个用户生成**五个任务**。随后，LLM针对每个用户任务组合生成**五个问题**，这些问题需要基于整个语料库的理解才能回答。最终，每个数据集生成总计 **125** 个问题。

019-LightRAG——评估指标

为了对涉及整个数据集的查询进行评估，我们采用了以下四个评估维度：

01

Comprehensiveness

How thoroughly does the answer address all aspects and details of the question?

02

Diversity

How varied and rich is the answer in offering different perspectives and insights related to the question?

03

Empowerment

How effectively does the answer enable the reader to understand the topic and make informed judgments?

04

Overall

This dimension assesses the cumulative performance across the three preceding criteria to identify the best overall answer.

020-LightRAG——评估结果

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

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

021-LightRAG——评估结果

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

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

022-LightRAG——开销计算

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

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

$T_{extract}$表示用于实体和关系抽取的Token开销，C_{max} 代表每次API调用允许的最大Token数量，$C_{extract}$表示抽取过程中所需的API调用次数。

在检索阶段，GraphRAG生成了**1,399**个社区，其中有**610**个二级社区在本次实验中被实际用于检索。每个社区报告的平均长度为**1,000**个Token。

在增量数据更新阶段，GraphRAG每个社区报告额外消耗约 **5,000** 个Token，重新构建新旧社区报告的Token成本将高达 **1,399 × 2 × 5,000**。

023-LightRAG——开销计算

插入时间以及增量更新时间

Table 4: Comparison of Document Insertion Times

No.	Token Count	LightRAG (s)	GraphRAG (s)
1	59,870	486	642
2	41,224	418	700
3	73,989	561	953
4	47,502	513	741
5	48,353	453	926

平均查询时间

Table 5: Average Query Times

Metric	LightRAG	GraphRAG
Average Query Time (s)	11.2	23.6

存储空间占用

Table 6: Final Storage Space Usage

Method	Final Storage Space (MB)
LightRAG	39.5
GraphRAG	286.7

024-LightRAG——查询示例

Query: What metrics are most informative for evaluating movie recommendation systems?

Query

High level keywords: ["Metrics", "Movie recommendation systems", "Evaluation methods"],
Low level keywords: ["Accuracy", "Precision", "Recall", "F1 score", "User satisfaction", "Diversity", "Coverage"]

Keywords

---**Entities**---
PERFORMANCE METRICS. Performance Metrics are measures used to evaluate how well the Decision Tree model performs, including RMSLE among others.
PRECISION-RECALL (PR) CURVE. The precision-recall curve is an event used to visualize the trade-off between precision and recall for different thresholds in binary classification.
MOVIE CLUSTERING,.Movie clustering refers to the grouping of movies based on similarities using the K-means algorithm and evaluating their performance metrics.
ACCURACY. Accuracy is a basic metric used to assess the performance of predictive models, expressed as a percentage of correct predictions.
...

---**Relationships**---
ITEMMATRIX, **MOVIES**, ItemMatrix holds the factors for movies, used to derive recommendations for each user based on their preferences.
AVGPRECISIONK, **SPARK**, avgPrecisionK function is utilized within the Spark framework to analyze the precision of recommendations for users.
MEAN SQUARE ERROR, **VARIANCE**, Mean Square Error is calculated based on the total Variance in the dataset, showing the relationship between explained and unexplained variance. + error measurement, variance
ACTUALMOVIES, **PREDICTEDMOVIES**, ActualMovies are compared against predictedMovies to assess the accuracy of the movie recommendation model
...

---**Sources**---
... these metrics are popular for model evaluation and selection purposes ...... These movie IDs are sorted in order of the estimated rating.\nAs we did for MSE and RMSE, we can compute ranking-based evaluation metrics using MLlib's `RankingMetrics` class. ...... Then, for each user, we have the list of actual and predicted movie IDs that we can pass to our APK function. In a manner similar to how we computed MSE, we will sum each of these APK scores using a `reduce` action and divide by the number of users (that is, the count of the `allRecs` RDD):\nFinally, we compute the accuracy for the decision tree...... First, we will compute the MSE and RMSE metrics using `RegressionMetrics`. ...

Retrieval Context

Key Metrics for Evaluating Movie Recommendation Systems
Evaluating movie recommendation systems involves multiple metrics that assess the effectiveness and accuracy of the recommendations provided to users. Below are some of the most informative metrics commonly used in this context, derived from the data and concepts surrounding recommendation systems.
1. **Mean Average Precision at K (MAPK)** [...]
2. **Precision and Recall** [...]
3. **Root Mean Squared Error (RMSE)** and **Mean Squared Error (MSE)** [...]
4. **Area Under the Curve (AUC)** [...]
5. **F-Measure** [...]
6. **User Engagement Metrics** [...]

Conclusion: A comprehensive evaluation of movie recommendation systems employs a range of metrics that capture both the accuracy of the recommendations and user satisfaction. By effectively leveraging these metrics, system designers can optimize recommendation engines to deliver more personalized and relevant experiences to users.

LLM Response

025-LightRAG——自定义KG

由于LightRAG具备低耦合性的设计，用户能够轻松地将自定义的知识图谱插入现有图谱中，灵活补充和扩展系统的知识表示。虽然LLM可以自动抽取文本中的实体和关系，但在医疗、法律、金融等专业垂直领域，LLM可能面临抽取不准确或信息不完整的问题。通过插入自定义KG，用户可以补充领域特定的高质量知识，使系统更好地适配专业场景需求。

```
custom_kg = {
    "entities": [
        {
            "entity_name": "CompanyA",
            "entity_type": "Organization",
            "description": "A major technology company",
            "source_id": "Source1"
        },
        {
            "entity_name": "ProductX",
            "entity_type": "Product",
            "description": "A popular product developed by CompanyA",
            "source_id": "Source1"
        }
    ],
    "relationships": [
        {
            "src_id": "CompanyA",
            "tgt_id": "ProductX",
            "description": "CompanyA develops ProductX",
            "keywords": "develop, produce",
            "weight": 1.0,
            "source_id": "Source1"
        }
    ],
    "chunks": [
        {
            "content": "ProductX, developed by CompanyA, has revolutionized the market with its cutting
            "source_id": "Source1",
        },
        {
            "content": "PersonA is a prominent researcher at UniversityB, focusing on artificial intell
            "source_id": "Source2",
        },
        {
            "content": "None",
            "source_id": "UNKNOWN",
        },
    ],
}
```

Deploy

DATA GRAPHX

API 和数据库设置

文档管理

上传 TXT 文件

Drag and drop file here

Limit 200MB per file • TXT

Browse files

已处理文档

选择文档

小王子前十章

已选择文档: 小王子前十章_20241024_041642

导入图库

LightRAG 知识图谱交互问答系统

智能问答

请输入您的问题：

小王子和玫瑰的关系

hybrid

naive hybrid

发送 清除历史

小王子和玫瑰的关系

小王子与玫瑰之间的关系是《小王子》故事中最核心和最具象征性的情感联结之一，这一关系不仅充满了温情和细腻的情感，也反映了小王子对爱的理解、责任和脆弱的思考。

深厚的情感依赖

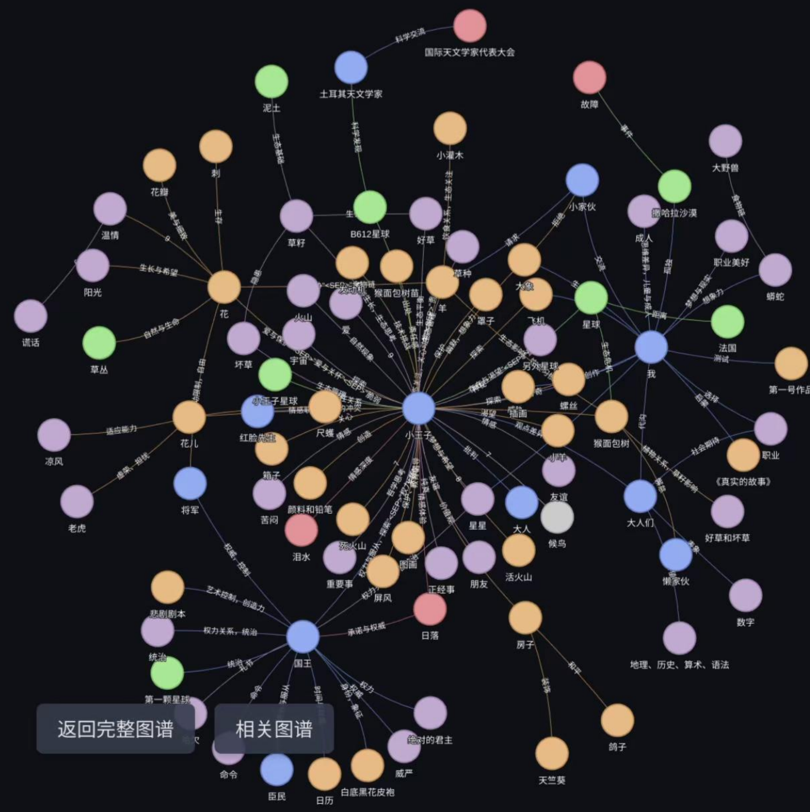
小王子对他的玫瑰抱有深厚的爱慕之情。他从一开始就表现出对玫瑰的关心和关注，努力照顾她的需求，例如为她提供保护性的罩子，以抵御环境的威胁。这种关心体现了他对生命脆弱性的理解和对爱的珍视。在小王子看来，玫瑰是他星球上唯一的花，充满了独特的地位和意义，这种独特性增强了他对她的感情。

矛盾与痛苦

然而，这种依赖的感情中也存在着矛盾和痛苦。小王子在与玫瑰的接触中，常常感到困惑和沮丧。玫瑰有时表现出自恋和娇气，给小王子带来了情感上的痛苦。玫瑰的虚荣心和脆弱让小王子感受到了一种爱的复杂性和脆弱性。在他离开星球后，这种对玫瑰的情感也成为了他追寻自我和理解爱的旅程中的重要一部分。

责任与理解

图谱交互



返回完整图谱 相关图谱

图谱控制

026-改进方向

1 LLM的不稳定性

不同模型（如GPT-4、Claude、PaLM等）在抽取实体和关系时的精度和质量差异较大。这种差异会导致知识图谱中节点和边的质量参差不齐，进而影响检索和生成阶段的效果。

2 检索内容优化

目前的检索机制采用的是“饱和检索”策略，即系统总会检索到预设的内容上限（如上下文容量的最大值），但实际上很多查询并不需要如此多的内容支持，反而可能引入不必要的信息干扰生成结果。这种冗余的内容不仅增加了计算成本，还可能降低生成回答的相关性与精确度。

3 检索文档约束

不同类型的查询往往需要从特定范围的文档中进行检索，而不是在整个知识库中无差别地搜索。例如，针对法律相关的问题，系统应优先检索法律文档；而对于技术问题，则应限制在技术类文档中检索。

4 删除机制

目前系统主要支持增量更新机制，能够有效整合新文档的实体和关系，但在删除旧文档时却缺乏相应的支持。随着知识图谱的规模增长，旧文档中的实体和关系可能因内容过时或与新知识矛盾而需要剥离。然而，合理地从图中删除相关实体和关系是一个复杂的问题。

HKUDS (✨ Data Intelligence L

github.com/HKUDS

完成更新

HKUDS / README.md

Hi there 👋

✨Welcome to the **Data Intelligence Lab @ HKU**!✨

🚀 Our Lab is Passionately Dedicated to Exploring the Forefront of the Data Science & AI 👨‍💻

Home Page | Google Scholar | 公众号 | Stars 17k | Followers 1.1k

✨Data Intelligence Lab@HKU✨

HKUDS

Unfollow

👨‍💻 Welcome to the Data Intelligence Lab! We are a team of dedicated researchers who specialize Data Science at the University of Hong Kong 📚

1.1k followers · **0** following

Followed by **FCode5** and 3 more

University of Hong Kong

Hong Kong

https://sites.google.com/view/chaoh

Achievements

Pinned

LightRAG Public
"LightRAG: Simple and Fast Retrieval-Augmented Generation"
Python ☆ 11.6k 1.5k

GraphGPT Public
[SIGIR'2024] "GraphGPT: Graph Instruction Tuning for Large Language Models"
Python ☆ 644 59

SSLRec Public
[WSDM'2024 Oral] "SSLRec: A Self-Supervised Learning Framework for Recommendation"
Python ☆ 501 64

LLMRec Public
[WSDM'2024 Oral] "LLMRec: Large Language Models with Graph Augmentation for Recommendation"
Python ☆ 399 48

RLMRec Public
[WWW'2024] "RLMRec: Representation Learning with Large Language Models for Recommendation"
Python ☆ 370 46

UrbanGPT Public
[KDD'2024] "UrbanGPT: Spatio-Temporal Large Language Models"
Python ☆ 311 44

135 contributions in the last year

2024

Thank you for listening!