

# From Local to Global: A GraphRAG Approach to Query-Focused Summarization

从局部到全局: 一种面向查询的 GraphRAG 摘要方法

Darren Edge<sup>1†</sup> Ha Trinh<sup>1†</sup> Newman Cheng<sup>2</sup> Joshua Bradley<sup>2</sup> Alex Chao<sup>3</sup>

Darren Edge<sup>1†</sup> Ha Trinh<sup>1†</sup> Newman Cheng<sup>2</sup> Joshua Bradley<sup>2</sup> Alex Chao<sup>3</sup>

Apurva Mody<sup>3</sup> Steven Truitt<sup>2</sup> Dasha Metropolitansky<sup>1</sup> Robert Osazuwa Ness<sup>1</sup>

Apurva Mody<sup>3</sup> Steven Truitt<sup>2</sup> Dasha Metropolitansky<sup>1</sup> Robert Osazuwa Ness<sup>1</sup>

Jonathan Larson<sup>1</sup>

Jonathan Larson<sup>1</sup>

<sup>1</sup> Microsoft Research

<sup>1</sup> 微软研究院

<sup>2</sup> Microsoft Strategic Missions and Technologies

<sup>2</sup> 微软战略任务与技术

<sup>3</sup> Microsoft Office of the CTO

<sup>3</sup> 微软首席技术官办公室

{daedge, trinhha, newmancheng, joshbradley, achao, moapurva, steventruitt, dasham, robertness, jolarso}@microsoft.com

{daedge, trinhha, newmancheng, joshbradley, achao, moapurva, steventruitt, dasham, robertness, jolarso}@microsoft.com

<sup>†</sup> These authors contributed equally to this work

<sup>†</sup> 这些作者对本工作贡献相等

## Abstract

### 摘要

The use of retrieval-augmented generation (RAG) to retrieve relevant information from an external knowledge source enables large language models (LLMs) to answer questions over private and/or previously unseen document collections. However, RAG fails on global questions directed at an entire text corpus, such as "What

are the main themes in the dataset?”, since this is inherently a query-focused summarization (QFS) task, rather than an explicit retrieval task. Prior QFS methods, meanwhile, do not scale to the quantities of text indexed by typical RAG systems. To combine the strengths of these contrasting methods, we propose GraphRAG, a graph-based approach to question answering over private text corpora that scales with both the generality of user questions and the quantity of source text. Our approach uses an LLM to build a graph index in two stages: first, to derive an entity knowledge graph from the source documents, then to pregenerate community summaries for all groups of closely related entities. Given a question, each community summary is used to generate a partial response, before all partial responses are again summarized in a final response to the user. For a class of global sensemaking questions over datasets in the 1 million token range, we show that GraphRAG leads to substantial improvements over a conventional RAG baseline for both the comprehensiveness and diversity of generated answers.

使用检索增强生成 (RAG) 从外部知识源检索相关信息，使大语言模型 (LLM) 能够针对私有和/或先前未见的文档集合回答问题。然而，RAG 在面向整个文集的全局问题上会失败，例如“数据集中主要主题是什么？”，因为这是固有的面向查询的摘要 (QFS) 任务，而不是显式检索任务。与此同时，现有的 QFS 方法无法扩展到典型 RAG 系统所索引的大量文本。为结合这两类方法的优点，我们提出了 GraphRAG，一种面向私有文本语料库的基于图的方法，能同时随用户问题的普适性和源文本量而扩展。我们的方法使用 LLM 分两阶段构建图索引：首先，从源文档中推导实体知识图；然后为所有紧密相关实体群预生成社区摘要。给定问题时，每个社区摘要用于生成部分回应，然后将所有部分回应再汇总成对用户的最终回应。对于约 100 万标记量级数据集上的一类全局感知问题，我们展示了 GraphRAG 在生成答案的全面性和多样性方面，相较于传统 RAG 基线有显著提升。

## 1 Introduction

### 1 引言

Retrieval augmented generation (RAG) (Lewis et al., 2020) is an established approach to using LLMs to answer queries based on data that is too large to contain in a language model’s context window, meaning the maximum number of tokens (units of text) that can be processed by the LLM at once (Kuratov et al., 2024; Liu et al., 2023). In the canonical RAG setup, the system has access to a large external corpus of text records and retrieves a subset of records that are individually relevant to the query and collectively small enough to fit into the context window of the LLM. The LLM then generates a response based on both the query and the retrieved records (Baumel et al., 2018; Dang, 2006; Laskar et al., 2020; Yao et al., 2017). This conventional approach, which we collectively call vector RAG, works well for queries that can be answered with information localized within a small set of records. However, vector RAG approaches do not support sensemaking queries, meaning queries that require global understanding of the entire dataset, such as “What are the key trends in how scientific discoveries are influenced by interdisciplinary research over the past decade?”

检索增强生成 (RAG)(Lewis 等, 2020) 是将 LLM 用于基于超出模型上下文窗口容量的数据的查询回答的一种既有方法, 上下文窗口指可被 LLM 一次处理的最大标记数 (Kuratov 等, 2024; Liu 等, 2023)。在典型的 RAG 设置中, 系统可访问大型外部文本语料库, 并检索对查询各自相关且总体足够小以适配 LLM 上下文窗口的记录子集。随后 LLM 基于查询和检索到的记录生成回答 (Baumel 等, 2018; Dang, 2006; Laskar 等, 2020; Yao 等, 2017)。我们统称为向量 RAG 的这种常规方法对可在少量记录内本地化回答的问题效果良好。然而, 向量 RAG 无法支持需对整个数据集进行全局理解的感知查询, 例如“过去十年跨学科研究如何影响科学发现的主要趋势是什么?”

Sensemaking tasks require reasoning over “connections (which can be among people, places, and events) in order to anticipate their trajectories and act effectively” (Klein et al., 2006). LLMs such as GPT (Achiam et al., 2023; Brown et al., 2020), Llama (Touvron et al., 2023), and Gemini (Anil et al., 2023) excel at sensemaking in complex domains like scientific discovery (Microsoft, 2023) and intelligence analysis (Ranade and Joshi, 2023). Given a sensemaking query and a text with an implicit and interconnected set of concepts, an LLM can generate a summary that answers the query. The challenge, however, arises when the volume of data requires a RAG approach, since vector RAG approaches are unable to support sensemaking over an entire corpus.

感知任务需要推理“用于预判其轨迹并有效行动的连接(可发生在人物、地点和事件之间)” (Klein 等, 2006)。诸如 GPT(Achiam 等, 2023; Brown 等, 2020)、Llama(Touvron 等, 2023) 和 Gemini(Anil 等, 2023) 等 LLM 在复杂领域的感知方面表现出色, 例如科学发现(微软, 2023) 和情报分析(Ranade 与 Joshi, 2023)。在给定感知查询和隐含且相互关联概念的文本时, LLM 能生成回答该查询的摘要。然而, 当数据量需要采用 RAG 方法时, 问题出现了, 因为向量 RAG 方法无法支持针对整个语料库的感知。

In this paper, we present GraphRAG - a graph-based RAG approach that enables sensemaking over the entirety of a large text corpus. GraphRAG first uses an LLM to construct a knowledge graph, where nodes correspond to key entities in the corpus and edges represent relationships between those entities. Next, it partitions the graph into a hierarchy of communities of closely related entities, before using an LLM to generate community-level summaries. These summaries are generated in a bottom-up manner following the hierarchical structure of extracted communities, with summaries at higher levels of the hierarchy recursively incorporating lower-level summaries. Together, these community summaries provide global descriptions and insights over the corpus. Finally, GraphRAG answers queries through map-reduce processing of community summaries; in the map step, the summaries are used to provide partial answers to the query independently and in parallel, then in the reduce step, the partial answers are combined and used to generate a final global answer.

在本文中, 我们提出 GraphRAG——一种基于图的 RAG 方法, 能够对大型文本语料库的整体进行意义构建。GraphRAG 首先使用 LLM 构建知识图谱, 节点对应语料库中的关键实体, 边表示这些实体之间的关系。接着, 它将图划分为一层层紧密相关实体的社区层级, 然后使用 LLM 生成社区级摘要。这些摘要按提取社区的层级结构自底向上生成, 层级较高的摘要递归地整合较低层的摘要。综合这些社区摘要可以为语料库提供全局性的描述与洞见。最后, GraphRAG 通过对社区摘要进行 map-reduce 处理来回答查询; 在 map 步骤中, 摘要被用来独立且并行地为查询提供部分答案, 而在 reduce 步骤中, 将这些部分答案合并并用于生成最终的全局答案。

The GraphRAG method and its ability to perform global sensemaking over an entire corpus form the main contribution of this work. To demonstrate this ability, we developed a novel application of the LLM-as-

a-judge technique (Zheng et al., 2024) suitable for questions targeting broad issues and themes where there is no ground-truth answer. This approach first uses one LLM to generate a diverse set of global sensemaking questions based on corpus-specific use cases, before using a second LLM to judge the answers of two different RAG systems using predefined criteria (defined in Section 3.3). We use this approach to compare GraphRAG to vector RAG on two representative real-world text datasets. Results show GraphRAG strongly outperforms vector RAG when using GPT-4 as the LLM.

GraphRAG 方法及其对整个语料库执行全局意义构建的能力构成了本工作的主要贡献。为证明该能力，我们开发了一种适用于针对广泛议题和主题且无标准答案的问题的新颖 LLM-as-a-judge 技术应用 (Zheng et al., 2024)。该方法首先使用一个 LLM 基于语料库特定用例生成多样的全局意义构建问题，然后使用第二个 LLM 按预定义标准 (在第 3.3 节定义) 对两种不同 RAG 系统的答案进行评判。我们使用该方法在两个具代表性的真实文本数据集上将 GraphRAG 与向量 RAG 进行比较。结果显示，在使用 GPT-4 作为 LLM 时，GraphRAG 显著优于向量 RAG。

GraphRAG is available as open-source software at <https://github.com/microsoft/graphrag>. In addition, versions of the GraphRAG approach are also available as extensions to multiple open-source libraries, including LangChain (LangChain, 2024), LlamaIndex (LlamaIndex, 2024), NebulaGraph (NebulaGraph, 2024), and Neo4J (Neo4J, 2024).

GraphRAG 已作为开源软件发布于 <https://github.com/microsoft/graphrag>。此外，GraphRAG 方法的若干版本也作为多种开源库的扩展提供，包括 LangChain (LangChain, 2024)、LlamaIndex (LlamaIndex, 2024)、NebulaGraph (NebulaGraph, 2024) 和 Neo4J (Neo4J, 2024)。

## 2 Background

### 2 背景

### 2.1 RAG Approaches and Systems

#### 2.1 RAG 方法与系统

RAG generally refers to any system where a user query is used to retrieve relevant information from external data sources, whereupon this information is incorporated into the generation of a response to the query by an LLM (or other generative AI model, such as a multi-media model). The query and retrieved records populate a prompt template, which is then passed to the LLM (Ram et al., 2023). RAG is ideal when the total number of records in a data source is too large to include in a single prompt to the LLM, i.e. the amount of text in the data source exceeds the LLM’s context window.

RAG 通常指任何使用用户查询从外部数据源检索相关信息、然后由 LLM(或其他生成式 AI 模型，如多模态模型) 将这些信息纳入生成查询响应的系统。查询与检索到的记录填入提示模板，随后传入 LLM (Ram et al., 2023)。当数据源中的记录总量过大，无法全部纳入 LLM 的单次提示中，即数据源文本量超过 LLM 的上下文窗口时，RAG 是理想的选择。

In canonical RAG approaches, the retrieval process returns a set number of records that are semantically similar to the query and the generated answer uses only the information in those retrieved records. A common approach to conventional RAG is to use text embeddings, retrieving records closest to the query in vector space where closeness corresponds to semantic similarity (Gao et al., 2023). While some RAG approaches may use alternative retrieval mechanisms, we collectively refer to the family of conventional approaches as vector RAG. GraphRAG contrasts with vector RAG in its ability to answer queries that require global sensemaking over the entire data corpus.

在规范的 RAG 方法中，检索过程返回与查询语义相似的一定数量记录，生成的答案仅使用这些检索到记录中的信息。传统 RAG 的常见做法是使用文本嵌入，检索在向量空间中与查询最接近的记录，距离近表示语义相似 (Gao et al., 2023)。虽然一些 RAG 方法可能使用替代的检索机制，我们将传统方法统称为向量 RAG。GraphRAG 与向量 RAG 的对比在于其能回答需要对整个数据语料库进行全局意义构建的查询。

GraphRAG builds upon prior work on advanced RAG strategies. GraphRAG leverages summaries over large sections of the data source as a form of “self-memory” (described in Cheng et al. 2024), which are later used to answer queries as in Mao et al. 2020). These summaries are generated in parallel and iteratively aggregated into global summaries, similar to prior techniques (Feng et al., 2023; Gao et al., 2023; Khattab et al., 2022; Shao et al., 2023; Su et al., 2020; Trivedi et al., 2022; Wang et al., 2024). In particular, GraphRAG is similar to other approaches that use hierarchical indexing to create summaries (similar to Kim et al. 2023; Sarthi et al. 2024). GraphRAG contrasts with these approaches by generating a graph index from the source data, then applying graph-based community detection to create a thematic partitioning of the data.

GraphRAG 建立在先前关于高级 RAG 策略的工作之上。GraphRAG 利用覆盖大段数据源的摘要作为一种“自我记忆” (见 Cheng et al. 2024)，随后用于回答查询 (见 Mao et al. 2020)。这些摘要并行生成并迭代聚合为全局摘要，类似于先前的技术 (Feng et al., 2023; Gao et al., 2023; Khattab et al., 2022; Shao et al., 2023; Su et al., 2020; Trivedi et al., 2022; Wang et al., 2024)。特别地，GraphRAG 类似于其他使用层级索引来创建摘要的方法 (类似 Kim et al. 2023; Sarthi et al. 2024)。GraphRAG 与这些方法的区别在于它从源数据生成图索引，然后应用基于图的社区检测来创建主题化的数据划分。

## 2.2 Using Knowledge Graphs with LLMs and RAG

### 2.2 将知识图与 LLM 和 RAG 结合使用

Approaches to knowledge graph extraction from natural language text corpora include rule-matching, statistical pattern recognition, clustering, and embeddings (Etzioni et al., 2004; Kim et al., 2016; Mooney and Bunescu, 2005; Yates et al., 2007). GraphRAG falls into a more recent body of research that use of LLMs for knowledge graph extraction (Ban et al., 2023; Melnyk et al., 2022; OpenAI, 2023; Tan et al., 2017; Trajanoska et al., 2023; Yao et al., 2023; Yates et al., 2007; Zhang et al., 2024a). It also adds to a growing body of RAG approaches that use a knowledge graph as an index (Gao et al., 2023). Some techniques use subgraphs, elements of the graph, or properties of the graph structure directly in the prompt (Baek et al., 2023; He et al., 2024; Zhang, 2023) or as factual grounding for generated outputs (Kang et al., 2023; Ranade and Joshi, 2023). Other techniques (Wang et al., 2023b) use the knowledge graph to enhance retrieval, where at query time an LLM-based agent dynamically traverses a graph with nodes representing document elements (e.g., passages,

tables) and edges encoding lexical and semantical similarity or structural relationships. GraphRAG contrasts with these approaches by focusing on a previously unexplored quality of graphs in this context: their inherent modularity (Newman, 2006) and the ability to partition graphs into nested modular communities of closely related nodes (e.g., Louvain, Blondel et al. 2008; Leiden, Traag et al. 2019). Specifically, GraphRAG recursively creates increasingly global summaries by using the LLM to create summaries spanning this community hierarchy.

从自然语言文本语料中抽取知识图的方法包括规则匹配、统计模式识别、聚类和嵌入 (Etzioni et al., 2004; Kim et al., 2016; Mooney and Bunescu, 2005; Yates et al., 2007)。GraphRAG 属于最近一批使用大型语言模型从文本中抽取知识图的研究 (Ban et al., 2023; Melnyk et al., 2022; OpenAI, 2023; Tan et al., 2017; Trajanoska et al., 2023; Yao et al., 2023; Yates et al., 2007; Zhang et al., 2024a)。它也补充了越来越多将知识图用作索引的 RAG 方法的研究 (Gao et al., 2023)。一些技术在提示中直接使用子图、图的元素或图结构的属性 (Baek et al., 2023; He et al., 2024; Zhang, 2023), 或将其作为生成输出的事实依据 (Kang et al., 2023; Ranade and Joshi, 2023)。其他技术 (Wang et al., 2023b) 利用知识图增强检索, 在查询时基于 LLM 的代理动态遍历图, 节点表示文档元素 (如段落、表格), 边编码词汇和语义相似性或结构关系。GraphRAG 与这些方法的区别在于关注在此背景下未被充分探索的图的一个特性: 其内在的模块性 (Newman, 2006) 以及将图划分为紧密相关节点的嵌套模块化社区的能力 (例如 Louvain, Blondel et al. 2008; Leiden, Traag et al. 2019)。具体而言, GraphRAG 通过使用 LLM 在该社区层次上生成摘要, 递归地创建越来越全局化的摘要。

## 2.3 Adaptive benchmarking for RAG Evaluation

### 2.3 面向 RAG 评估的自适应基准测试

Many benchmark datasets for open-domain question answering exist, including HotPotQA (Yang et al., 2018), MultiHop-RAG (Tang and Yang, 2024), and MT-Bench (Zheng et al., 2024). However, these benchmarks are oriented towards vector RAG performance, i.e., they evaluate performance on explicit fact retrieval. In this work, we propose an approach for generating a set of questions for evaluating global sensemaking over the entirety of the corpus. Our approach is related to LLM methods that use a corpus to generate questions whose answers would be summaries of the corpus, such as in Xu and Lapata (2021). However, in order to produce a fair evaluation, our method avoids generating the questions directly from the corpus itself (as an alternative implementation, one can use a subset of the corpus held out from subsequent graph extraction and answer evaluation steps).

存在许多用于开放域问答的基准数据集, 包括 HotPotQA (Yang et al., 2018)、MultiHop-RAG (Tang and Yang, 2024) 和 MT-Bench (Zheng et al., 2024)。然而, 这些基准侧重于向量 RAG 的性能评估, 即评估对显式事实检索的表现。在本文中, 我们提出了一种生成问题集合的方法, 用以评估对整个语料库的全局理解能力。我们的方法与使用语料生成问题、其答案为语料摘要的 LLM 方法相关, 如 Xu 和 Lapata (2021)。但为保证公平评估, 我们的方法避免直接从语料本身生成问题 (作为替代实现, 可以使用从随后的图抽取和答案评估步骤中保留出来的语料子集)。

Adaptive benchmarking refers to the process of dynamically generating evaluation benchmarks tailored to specific domains or use cases. Recent work has used LLMs for adaptive benchmarking to ensure relevance, diversity, and alignment with the target application or task (Yuan et al., 2024; Zhang et al., 2024b). In this

work, we propose an adaptive benchmarking approach to generating global sensemaking queries for the LLM. Our approach builds on prior work in LLM-based persona generation, where the LLM is used to generate diverse and authentic sets of personas (Kosinski, 2024; Salminen et al., 2024; Shin et al., 2024). Our adaptive benchmarking procedure uses persona generation to create queries that are representative of real-world RAG system usage. Specifically, our approach uses the LLM to infer the potential users would use the RAG system and their use cases, which guide the generation of corpus-specific sensemaking queries.

自适应基准测试指动态生成针对特定领域或用例的评估基准的过程。近期工作已使用 LLM 进行自适应基准测试，以确保与目标应用或任务的相关性、多样性和一致性 (Yuan et al., 2024; Zhang et al., 2024b)。在本文中，我们提出了一种用于为 LLM 生成全局理解查询的自适应基准方法。我们的方法建立在之前基于 LLM 的角色 (persona) 生成工作之上，其中 LLM 被用于生成多样且真实的角色集合 (Kosinski, 2024; Salminen et al., 2024; Shin et al., 2024)。我们的自适应基准流程使用角色生成来创建代表真实世界 RAG 系统使用情形的查询。具体而言，我们的方法使用 LLM 推断潜在用户将如何使用 RAG 系统及其用例，从而指导生成与语料库相关的理解查询。

## 2.4 RAG evaluation criteria

### 2.4 RAG 评估标准

Our evaluation relies on the LLM to evaluate how well the RAG system answers the generated questions. Prior work has shown LLMs to be good evaluators of natural language generation, including work where LLMs evaluations were competitive with human evaluations (Wang et al., 2023a; Zheng et al., 2024). Some prior work proposes criteria for having LLMs quantify the quality of generated texts such as "fluency" (Wang et al., 2023a). Some of these criteria are generic to vector RAG systems and not relevant to global sensemaking, such as "context relevance", "faithfulness", and "answer relevance" (RAGAS, Es et al. 2023). Lacking a gold standard for evaluation, one can quantify relative performance for a given criterion by prompting the LLM to compare generations from two different competing models (LLM-as-a-judge, (Zheng et al., 2024)). In this work, we design criteria for evaluating RAG-generated answers to global sensemaking questions and evaluate our results using the comparative approach. We also validate results using statistics derived from LLM-extracted statements of verifiable facts, or "claims."

我们的评估依赖于 LLM 来评判 RAG 系统回答生成问题的效果。先前工作表明 LLM 是自然语言生成的良好评估者，其中一些研究显示 LLM 的评估可与人工评估相媲美 (Wang et al., 2023a; Zheng et al., 2024)。已有工作提出让 LLM 量化生成文本质量的标准，例如“流畅性” (Wang et al., 2023a)。其中一些标准对向量 RAG 系统通用但与全局意义建构无关，例如“上下文相关性”、“忠实性”和“答案相关性” (RAGAS, Es et al. 2023)。在缺乏金标准的情况下，可以通过提示 LLM 比较来自两个不同竞品模型的生成结果来量化某一标准的相对性能 (LLM 作为裁判, (Zheng et al., 2024))。在本工作中，我们为评估 RAG 生成的全局意义建构问题答案设计了评判标准，并使用比较方法评估结果。我们还利用从 LLM 提取的可验证事实陈述或“主张”所衍生的统计数据来验证结果。

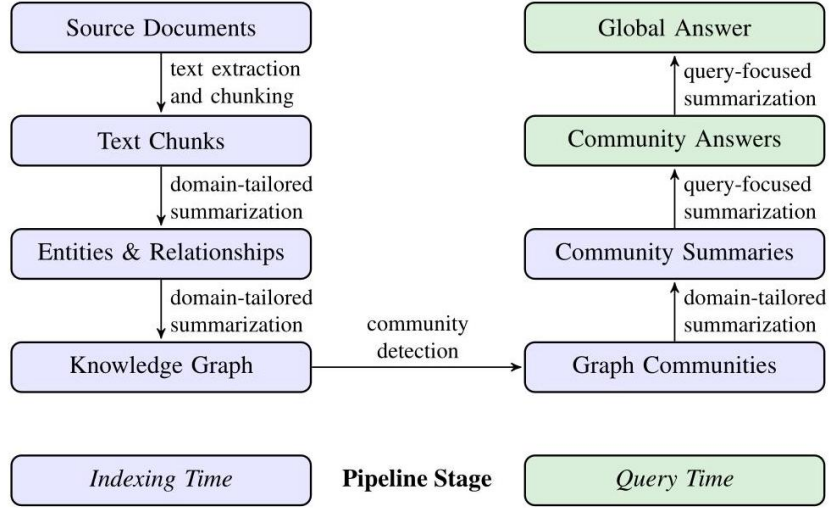


Figure 1: Graph RAG pipeline using an LLM-derived graph index of source document text. This graph index spans nodes (e.g., entities), edges (e.g., relationships), and covariates (e.g., claims) that have been detected, extracted, and summarized by LLM prompts tailored to the domain of the dataset. Community detection (e.g., Leiden, Traag et al., 2019) is used to partition the graph index into groups of elements (nodes, edges, covariates) that the LLM can summarize in parallel at both indexing time and query time. The “global answer” to a given query is produced using a final round of query-focused summarization over all community summaries reporting relevance to that query.

图 1: 使用由 LLM 派生的源文档文本图索引的 Graph RAG 管道。该图索引包含节点 (如实体)、边 (如关系) 和协变量 (如主张), 这些元素均由针对数据集领域定制的 LLM 提示检测、提取并摘要。社区检测 (例如 Leiden, Traag et al., 2019) 用于将图索引划分为元素组 (节点、边、协变量), 以便 LLM 在索引时和查询时并行地对这些组进行摘要。针对给定查询的“全局答案”由对所有社区摘要进行最终一轮以查询为中心的摘要产生, 这些摘要报告与该查询的相关性。

## 3 Methods

### 3 方法

### 3.1 GraphRAG Workflow

#### 3.1 GraphRAG workflow

Figure 1 illustrates the high-level data flow of the GraphRAG approach and pipeline. In this section, we describe the key design parameters, techniques, and implementation details for each step.

图 1 描述了 GraphRAG 方法与管道的数据流高层次结构。本节我们将介绍每一步的关键设计参数、技术和实现细节。

### 3.1.1 Source Documents → Text Chunks

#### 3.1.1 源文档 → 文本块

To start, the documents in the corpus are split into text chunks. The LLM extracts information from each chunk for downstream processing. Selecting the size of the chunk is a fundamental design decision; longer text chunks require fewer LLM calls for such extraction (which reduces cost) but suffer from degraded recall of information that appears early in the chunk (Kuratov et al., 2024; Liu et al., 2023). See Section A. 1 for prompts and examples of the recall-precision trade-offs.

首先，将语料库中的文档拆分为文本块。LLM 从每个块中提取信息以供后续处理。选择块的大小是一个基本的设计决策；较长的文本块需要更少的 LLM 调用来完成此类提取（从而降低成本），但会降低对出现在块前部信息的召回率 (Kuratov et al., 2024; Liu et al., 2023)。关于召回-精确度权衡的提示和示例见附录 A.1。

### 3.1.2 Text Chunks → Entities & Relationships

#### 3.1.2 文本块 → 实体与关系

In this step, the LLM is prompted to extract instances of important entities and the relationships between the entities from a given chunk. Additionally, the LLM generates short descriptions for the entities and relationships. To illustrate, suppose a chunk contained the following text:

在此步骤中，LLM 被提示从给定文本块中提取重要实体的实例及实体间的关系。此外，LLM 还为实体和关系生成简短描述。举例而言，假设某文本块包含以下内容：

NeoChip’s (NC) shares surged in their first week of trading on the NewTech Exchange. However, market analysts caution that the chipmaker’s public debut may not reflect trends for other technology IPOs. NeoChip, previously a private entity, was acquired by Quantum Systems in 2016. The innovative semiconductor firm specializes in low-power processors for wearables and IoT devices.

NeoChip(NC) 在 NewTech 交易所首周交易中股价飙升。然而，市场分析师警告说，该芯片制造商的公开亮相可能并不代表其他科技 IPO 的趋势。NeoChip 曾为私有实体，2016 年被 Quantum Systems 收购。这家创新的半导体公司专注于可穿戴设备和物联网设备的低功耗处理器。

The LLM is prompted such that it extracts the following:

LLM 被提示以便其提取如下内容：

- The entity NeoChip, with description “NeoChip is a publicly traded company specializing in low-power processors for wearables and IoT devices.”
- 实体 NeoChip，描述为 “NeoChip 是一家专注于可穿戴设备和物联网设备低功耗处理器的上市公司。”

- The entity Quantum Systems, with description "Quantum Systems is a firm that previously owned NeoChip."

• 实体 Quantum Systems, 描述为 "Quantum Systems 是曾经拥有 NeoChip 的公司。"

- A relationship between NeoChip and Quantum Systems, with description "Quantum Systems owned NeoChip from 2016 until NeoChip became publicly traded."

• NeoChip 与 Quantum Systems 之间的关系, 描述为 "Quantum Systems 自 2016 年起拥有 NeoChip, 直到 NeoChip 变为上市公司。"

These prompts can be tailored to the domain of the document corpus by choosing domain appropriate few-shot exemplars for in-context learning (Brown et al., 2020). For example, while our default prompt extracts the broad class of "named entities" like people, places, and organizations and is generally applicable, domains with specialized knowledge (e.g., science, medicine, law) will benefit from few-shot exemplars specialized to those domains.

这些提示可以通过选择适合语料库领域的少量示例进行上下文学习来定制。例如, 尽管我们的默认提示提取通用类别的“命名实体”, 如人物、地点和组织, 通常适用, 但具有专门知识的领域(如科学、医学、法律)将从针对这些领域的少量示例中受益。

The LLM can also be prompted to extract claims about detected entities. Claims are important factual statements about entities, such as dates, events, and interactions with other entities. As with entities and relationships, in-context learning exemplars can provide domain-specific guidance. Claim descriptions extracted from the example text chunk are as follows:

LLM 也可以被提示提取关于被检测实体的主张。主张是关于实体的重要事实性陈述, 例如日期、事件以及与其他实体的交互。如同实体和关系, 少量上下文学习示例可以提供领域特定的指导。示例文本块中提取的主张描述如下:

- NeoChip's shares surged during their first week of trading on the NewTech Exchange.

• NeoChip 的股票在其于 NewTech 交易所上市的首周交易中飙升。

- NeoChip debuted as a publicly listed company on the NewTech Exchange.

• NeoChip 在 NewTech 交易所首次作为上市公司亮相。

- Quantum Systems acquired NeoChip in 2016 and held ownership until NeoChip went public.

• Quantum Systems 于 2016 年收购了 NeoChip, 并一直持有所有权直到 NeoChip 上市。

See Appendix A for prompts and details on our implementation of entity and claim extraction.

• 有关实体和主张抽取的提示和实施细则, 请参见附录 A。

### 3.1.3 Entities & Relationships → Knowledge Graph

#### 3.1.3 实体与关系 → 知识图谱

The use of an LLM to extract entities, relationships, and claims is a form of abstractive summarization - these are meaningful summaries of concepts that, in the case of relationships and claims, may not be explicitly stated in the text. The entity/relationship/claim extraction processes creates multiple instances of a single element because an element is typically detected and extracted multiple times across documents.

使用大语言模型抽取实体、关系与主张是一种抽象式归纳——这些是对概念的有意义总结，在关系和主张的情况下，可能在文本中并未明示。实体/关系/主张抽取过程中会为单一元素创建多个实例，因为同一元素通常会在不同文档中被多次检测和抽取。

In the final step of the knowledge graph extraction process, these instances of entities and relationships become individual nodes and edges in the graph. Entity descriptions are aggregated and summarized for each node and edge. Relationships are aggregated into graph edges, where the number of duplicates for a given relationship becomes edge weights. Claims are aggregated similarly.

在知识图谱抽取流程的最后一步，这些实体和关系的实例会成为图中的各个节点和边。节点与边的实体描述会被聚合并摘要化。关系被聚合为图的边，某一关系的重复次数即成为边权重。主张也以类似方式聚合。

In this manuscript, our analysis uses exact string matching for entity matching - the task of reconciling different extracted names for the same entity (Barlaug and Gulla, 2021; Christen and Christen, 2012; Elmagarmid et al., 2006). However, softer matching approaches can be used with minor adjustments to prompts or code. Furthermore, GraphRAG is generally resilient to duplicate entities since duplicates are typically clustered together for summarization in subsequent steps.

在本文中，我们的分析使用精确字符串匹配进行实体匹配——也就是将不同提取名称的同一实体进行调和 (Barlaug 和 Gulla, 2021; Christen 和 Christen, 2012; Elmagarmid 等, 2006)。不过，也可以通过对提示或代码做小幅调整来采用更宽松的匹配方法。此外，GraphRAG 通常对重复实体具有鲁棒性，因为重复项通常会在后续步骤中聚类以便摘要化。

### 3.1.4 Knowledge Graph → Graph Communities

#### 3.1.4 知识图谱 → 图社区

Given the graph index created in the previous step, a variety of community detection algorithms may be used to partition the graph into communities of strongly connected nodes (e.g., see the surveys by Fortunato (2010) and Jin et al. (2021)). In our pipeline, we use Leiden community detection (Traag et al., 2019) in a hierarchical manner, recursively detecting sub-communities within each detected community until reaching leaf communities that can no longer be partitioned.

鉴于前一步创建的图索引，可使用多种社区检测算法将图划分为强连接节点的社区（例如，参见 Fortunato (2010) 和 Jin 等 (2021) 的综述）。在我们的流程中，我们采用 Leiden 社区检测 (Traag 等, 2019) 以分层方式递归检测每个已检测社区内的子社区，直到达到无法再细分的叶子社区。

Each level of this hierarchy provides a community partition that covers the nodes of the graph in a mutually exclusive, collectively exhaustive way, enabling divide-and-conquer global summarization. An illustration of such hierarchical partitioning on an example dataset can be found in Appendix B.

该层级结构的每一层提供一个社区划分，覆盖图的节点，互斥且穷尽，从而实现分而治之的全局摘要。在附录 B 可见该层级划分在示例数据集上的示意图。

### 3.1.5 Graph Communities → Community Summaries

#### 3.1.5 图社区 → 社区摘要

The next step creates report-like summaries of each community in the community hierarchy, using a method designed to scale to very large datasets. These summaries are independently useful as a way to understand the global structure and semantics of the dataset, and may themselves be used to make sense of a corpus in the absence of a specific query. For example, a user may scan through community summaries at one level looking for general themes of interest, then read linked reports at a lower level that provide additional details for each subtopic. Here, however, we focus on their utility as part of a graph-based index used for answering global queries.

下一步为社区层级中的每个社区创建类似报告的摘要，采用一种可扩展到非常大数据集的方法。这些摘要本身即可用于理解数据集的全局结构与语义，也可在没有特定查询的情况下用来理解语料。例如，用户可在某一层浏览社区摘要以寻找感兴趣的一般主题，然后阅读下层的链接报告以获取每个子主题的详细信息。此处我们关注的是其作为基于图的索引用于回答全局查询的实用性。

GraphRAG generates community summaries by adding various element summaries (for nodes, edges, and related claims) to a community summary template. Community summaries from lower-level communities are used to generate summaries for higher-level communities as follows:

GraphRAG 通过将各种元素摘要（节点、边及相关主张）加入社区摘要模板来生成社区摘要。低层社区的社区摘要被用于生成高层社区的摘要，方法如下：

- Leaf-level communities. The element summaries of a leaf-level community are prioritized and then iteratively added to the LLM context window until the token limit is reached. The prioritization is as follows: for each community edge in decreasing order of combined source and target node degree (i.e., overall prominence), add descriptions of the source node, target node, the edge itself, and related claims.
- 叶子级社区。优先考虑叶子级社区的元素摘要，然后迭代地将其加入 LLM 的上下文窗口直到达到令牌限制。优先顺序为：按社区中每条边的源节点和目标节点度数之和（即总体显著性）降序，对每条边依次加入源节点描述、目标节点描述、该边描述及相关主张。

- Higher-level communities. If all element summaries fit within the token limit of the context window, proceed as for leaf-level communities and summarize all element summaries within the community. Otherwise, rank sub-communities in decreasing order of element summary tokens and iteratively substitute sub-community summaries (shorter) for their associated element summaries (longer) until they fit within the context window.

- 高层社区。如果所有元素摘要都能在上下文窗口的令牌限制内，则按叶子级社区同法对社区内所有元素摘要进行总结。否则，按元素摘要令牌数降序对子社区进行排序，迭代地用子社区摘要(较短)替换其对应的元素摘要(较长)，直至它们适配上下文窗口。

### 3.1.6 Community Summaries → Community Answers → Global Answer

#### 3.1.6 社区摘要 → 社区答案 → 全局答案

Given a user query, the community summaries generated in the previous step can be used to generate a final answer in a multi-stage process. The hierarchical nature of the community structure also means that questions can be answered using the community summaries from different levels, raising the question of whether a particular level in the hierarchical community structure offers the best balance of summary detail and scope for general sensemaking questions (evaluated in section 4).

给定用户查询，上一步生成的社区摘要可在多阶段过程中用于生成最终答案。社区结构的层级性也意味着可以使用不同层级的社区摘要来回答问题，从而引出一个问题：层级社区结构中的某一特定层级是否在摘要细节与广义理解问题的覆盖范围之间提供了最佳平衡（在第 4 节评估）。

For a given community level, the global answer to any user query is generated as follows:

对于给定的社区层级，任何用户查询的全局答案按如下方式生成：

- Prepare community summaries. Community summaries are randomly shuffled and divided into chunks of pre-specified token size. This ensures relevant information is distributed across chunks, rather than concentrated (and potentially lost) in a single context window.

- 准备社区摘要。将社区摘要随机打乱并分成预先指定令牌大小的块。这可确保相关信息分布在各块中，而不是集中（并可能在单一上下文窗口中丢失）。

- Map community answers. Intermediate answers are generated in parallel. The LLM is also asked to generate a score between 0-100 indicating how helpful the generated answer is in answering the target question. Answers with score 0 are filtered out.

- 映射社区答案。并行生成中间答案。还要求 LLM 生成一个 0-100 的评分，表示所生成答案对回答目标问题的帮助程度。得分为 0 的答案会被过滤掉。

- Reduce to global answer. Intermediate community answers are sorted in descending order of helpfulness score and iteratively added into a new context window until the token limit is reached. This final context is used to generate the global answer returned to the user.

- 汇总为全局答案。按帮助度得分降序排列中间社区答案，迭代地将它们加入新的上下文窗口，直至达到令牌限制。该最终上下文用于生成返回给用户的全局答案。

## 3.2 Global Sensemaking Question Generation

### 3.2 全局理解问题生成

To evaluate the effectiveness of RAG systems for global sensemaking tasks, we use an LLM to generate a set of corpus-specific questions designed to assess high-level understanding of a given corpus, without requiring retrieval of specific low-level facts. Instead, given a high-level description of a corpus and its purposes, the LLM is prompted to generate personas of hypothetical users of the RAG system. For each hypothetical user, the LLM is then prompted to specify tasks that this user would use the RAG system to complete. Finally, for each combination of user and task, the LLM is prompted to generate questions that require understanding of the entire corpus. Algorithm 1 describes the approach.

为了评估 RAG 系统在全局理解任务上的有效性，我们使用 LLM 生成一组特定语料库的问题，旨在评估对给定语料库的高层次理解，而无需检索具体的低级事实。相反，在给出语料库及其目的的高层描述后，LLM 会被提示生成 RAG 系统假想用户的人物设定。对于每个假想用户，LLM 然后被提示指出该用户会使用 RAG 系统完成的任务。最后，对于每个用户与任务的组合，LLM 被提示生成需要理解整个语料库的问题。算法 1 描述了该方法。

#### Algorithm 1: Prompting Procedure for Question Generation

##### 算法 1: 问题生成的提示程序

1: Input: Description of a corpus, number of users  $K$ , number of tasks per user  $N$ , number of questions per (user, task) combination  $M$ .

1: 输入: 语料库描述, 用户数  $K$ , 每用户任务数  $N$ , 每 (用户, 任务) 组合的问题数  $M$ 。

2: Output: A set of  $K * N * M$  high-level questions requiring global understanding of the corpus.

2: 输出: 一组  $K * N * M$  需要对语料库进行全局理解的高层问题。

3: procedure GENERATEQUESTIONS

3: 过程 GENERATEQUESTIONS

4: Based on the corpus description, prompt the LLM to:

4: 基于语料库描述, 提示 LLM 执行:

1. Describe personas of  $K$  potential users of the dataset.

1. 描述  $K$  名潜在数据集用户的人物设定。

2. For each user, identify  $N$  tasks relevant to the user.

2. 对每个用户，识别与该用户相关的  $N$  项任务。

3. Specific to each user & task pair, generate  $M$  high-level questions that:

3. 针对每个用户与任务对，生成  $M$  个具体的高层问题，要求：

- Require understanding of the entire corpus.

- 要求理解整个语料库。

- Do not require retrieval of specific low-level facts.

- 不要求检索具体的低层事实。

5: Collect the generated questions to produce  $K * N * M$  test questions for the dataset.

5: 收集生成的问题以为数据集制作  $K * N * M$  道测试题。

6: end procedure

6: 结束过程

For our evaluation, we set  $K = M = N = 5$  for a total of 125 test questions per dataset. Table 1 shows example questions for each of the two evaluation datasets.

在我们的评估中，我们为每个数据集设置了共计  $K = M = N = 5$  道测试题。表 1 显示了两个评估数据集的示例问题。

### 3.3 Criteria for Evaluating Global Sensemaking

#### 3.3 评估全局理解的标准

Given the lack of gold standard answers to our activity-based sensemaking questions, we adopt the head-to-head comparison approach using an LLM evaluator that judges relative performance according to specific criteria. We designed three target criteria capturing qualities that are desirable for global sensemaking activities.

鉴于我们的基于活动的理解问题缺乏金标准答案，我们采用基于 LLM 评估器的对头比较方法，根据特定标准判断相对表现。我们设计了三个目标标准，体现了对全局理解活动有益的特性。

Appendix F shows the prompts for our head-to-head measures computed using an LLM evaluator, summarized as:

附录 F 显示了用于基于 LLM 评估器计算的对头衡量提示，概括如下：

- **Comprehensiveness.** How much detail does the answer provide to cover all aspects and details of the question?

- 全面性。答案在多大程度上提供细节以涵盖问题的所有方面和细节？

- **Diversity.** How varied and rich is the answer in providing different perspectives and insights on the question?

- 多样性。答案在提供不同视角和见解方面有多丰富多样？

- **Empowerment.** How well does the answer help the reader understand and make informed judgments about the topic?

- 赋能。答案在多大程度上帮助读者理解并就该主题做出有根据的判断？

Table 1: Examples of potential users, tasks, and questions generated by the LLM based on short descriptions of the target datasets. Questions target global understanding rather than specific details.

表 1: 基于目标数据集的简短描述由 LLM 生成的潜在用户、任务和问题示例。问题着眼于全局理解而非具体细节。

Dataset	Example activity framing and generation of global sensemaking questions
Podcast transcripts	<p>User: A tech journalist looking for insights and trends in the tech industry</p> <p>Task: Understanding how tech leaders view the role of policy and regulation</p> <p>Questions:</p> <ol style="list-style-type: none"> <li>1. Which episodes deal primarily with tech policy and government regulation?</li> <li>2. How do guests perceive the impact of privacy laws on technology development?</li> <li>3. Do any guests discuss the balance between innovation and ethical considerations?</li> <li>4. What are the suggested changes to current policies mentioned by the guests?</li> <li>5. Are collaborations between tech companies and governments discussed and how?</li> </ol>
News articles	<p>User: Educator incorporating current affairs into curricula</p> <p>Task: Teaching about health and wellness</p> <p>Questions:</p> <ol style="list-style-type: none"> <li>1. What current topics in health can be integrated into health education curricula?</li> <li>2. How do news articles address the concepts of preventive medicine and wellness?</li> <li>3. Are there examples of health articles that contradict each other, and if so, why?</li> <li>4. What insights can be gleaned about public health priorities based on news coverage?</li> <li>5. How can educators use the dataset to highlight the importance of health literacy?</li> </ol>

数据集	示例活动构思与生成宏观理解性问题
播客文字记录	用户: 一位寻求科技行业见解与趋势的科技记者 任务: 理解科技领袖如何看待政策与监管的角色问题: 1. 哪些集主要讨论科技政策与政府监管? 2. 嘉宾如何看待隐私法对技术发展的影响? 3. 是否有嘉宾讨论创新与伦理考量之间的平衡? 4. 嘉宾提出了哪些对现行政策的建议改变? 5. 是否讨论了科技公司与政府的合作, 以及如何讨论的?
新闻文章	用户: 将时事融入课程的教育工作者 任务: 教授健康与保健 问题: 1. 哪些当前健康话题可以纳入健康教育课程? 2. 新闻文章如何论述预防医学与健康的概念? 3. 是否有相互矛盾的健康报道示例, 如有, 原因何在? 4. 基于新闻报道可得出哪些有关公共卫生优先事项的见解? 5. 教育工作者如何利用该数据集强调健康素养的重要性?

Furthermore, we use a “control criterion” called Directness that answers “How specifically and clearly does the answer address the question?”. In plain terms, directness evaluates the concision of an answer in a generic sense that applies to any generated LLM summarization. We include it to behave as a reference against which we can judge the soundness of results for the other criteria. Since directness is effectively in opposition to comprehensiveness and diversity, we would not expect any method to win across all four criteria.

此外, 我们使用一种称为直接性 (Directness) 的“控制标准”, 用于回答“答案在多大程度上具体且清晰地回应了问题?”。通俗地说, 直接性评估答案的简洁性——一种适用于任何生成型大型模型摘要的通用衡量。我们将其作为参考标准, 以便评判其他准则结果的合理性。由于直接性与全面性和多样性在效果上存在权衡, 我们不期望任何方法能在所有四项准则上同时胜出。

In our evaluations, the LLM is provided with the question, the generated answers from two competing systems, and prompted to compare the two answers according to the criterion before giving a final judgment of which answer is preferred. The LLM either indicates a winner; or, it returns a tie if they are fundamentally similar. To account for the inherent stochasticity of LLM generation, we run each comparison with multiple replicates and average the results across replicates and questions. An illustration of LLM assessment for answers to a sample question can be found in Appendix D.

在我们的评估中, 大型模型会被提供问题、来自两套竞品系统的生成答案, 并被提示根据相应准则比较两份答案, 随后给出最终判定以决定更偏好的答案。模型要么指出赢家; 要么若两者本质相似则判为并列。为考虑生成的固有随机性, 我们对每次比较进行多次重复, 并将重复与问题间的结果取平均。附录 D 中展示了对示例问题答案进行模型评估的示例。

## 4 Analysis

### 4 分析

## 4.1 Experiment 1

### 4.1 实验一

#### 4.1.1 Datasets

##### 4.1.1 数据集

We selected two datasets in the one million token range, each representative of corpora that users may encounter in their real-world activities:

我们挑选了两个约为一百万标记量级的数据集，分别代表用户在现实活动中可能遇到的语料类型：

Podcast transcripts. Public transcripts of Behind the Tech with Kevin Scott, a podcast featuring conversations between Microsoft CTO Kevin Scott and various thought leaders in science and technology (Scott, 2024). This corpus was divided into  $1669 \times 600$ -token text chunks, with 100-token overlaps between chunks ( $\sim 1$  million tokens).

播客转录。公开的 Behind the Tech with Kevin Scott 节目转录，该播客由微软 CTO Kevin Scott 与各类科技与科学领域思想领袖对话 (Scott, 2024)。该语料被划分为  $1669 \times 600$  标记的文本片段，片段间有 100 标记重叠 ( $\sim 1$  百万标记)。

News articles. A benchmark dataset comprised of news articles published from September 2013 to December 2023 in a range of categories, including entertainment, business, sports, technology, health, and science (Tang and Yang, 2024). The corpus is divided into  $3197 \times 600$ -token text chunks, with 100-token overlaps between chunks ( $\sim 1.7$  million tokens).

新闻文章。一个基准数据集，包含 2013 年 9 月至 2023 年 12 月间发布的各类新闻文章，涵盖娱乐、商业、体育、技术、健康和科学等类别 (Tang and Yang, 2024)。语料被划分为  $3197 \times 600$  标记的文本片段，片段间有 100 标记重叠 ( $\sim 1.7$  百万标记)。

#### 4.1.2 Conditions

##### 4.1.2 条件

We compared six conditions including GraphRAG at four different graph community levels (C0, C1, C2, C3), a text summarization method that applies our map-reduce approach directly to source texts (TS), and a vector RAG "semantic search" approach (SS):

我们比较了六种条件，包括在四个不同图社群层级 (C0、C1、C2、C3) 下的 GraphRAG，直接将我们的 map-reduce 方法应用于源文本的文本摘要方法 (TS)，以及基于向量检索的 RAG “语义搜索”方法 (SS)：

- CO. Uses root-level community summaries (fewest in number) to answer user queries.

- C0。使用根级社群摘要 (数量最少) 来回答用户查询。

- C1. Uses high-level community summaries to answer queries. These are sub-communities of C0, if present, otherwise C0 communities projected downwards.

- C1。使用高层级社群摘要来回答查询。这些是 C0 的子社群 (如存在), 否则为向下投影的 C0 社群。

- C2. Uses intermediate-level community summaries to answer queries. These are subcommunities of C1, if present, otherwise C1 communities projected downwards.

- C2。使用中间层级社群摘要来回答查询。这些是 C1 的子社群 (如存在), 否则为向下投影的 C1 社群。

- C3. Uses low-level community summaries (greatest in number) to answer queries. These are sub-communities of C2, if present, otherwise C2 communities projected downwards.

- C3。使用低层级社群摘要 (数量最多) 来回答查询。这些是 C2 的子社群 (如存在), 否则为向下投影的 C2 社群。

- TS. The same method as in Section 3.1.6, except source texts (rather than community summaries) are shuffled and chunked for the map-reduce summarization stages.

- TS。与第 3.1.6 节中的方法相同, 但在 map-reduce 摘要阶段对源文本 (而非社群摘要) 进行洗牌与分块处理。

- SS. An implementation of vector RAG in which text chunks are retrieved and added to the available context window until the specified token limit is reached.

- SS。向量 RAG 的一种实现, 其中检索文本片段并将其添加到可用上下文窗口, 直至达到指定的标记限制。

The size of the context window and the prompts used for answer generation are the same across all six conditions (except for minor modifications to reference styles to match the types of context information used). Conditions only differ in how the contents of the context window are created.

上下文窗口的大小和用于生成答案的提示在六种条件中相同 (除为匹配所使用的上下文信息类型对参考样式做了少量修改)。条件之间仅在如何构建上下文窗口的内容上有所不同。

The graph index supporting conditions C0-C3 was created using our generic prompts for entity and relationship extraction, with entity types and few-shot examples tailored to the domain of the data.

支持条件 C0-C3 的图索引是使用我们用于实体和关系抽取的通用提示创建的, 实体类型和少样本示例针对数据领域进行了定制。

### 4.1.3 Configuration

#### 4.1.3 配置

We used a fixed context window size of 8k tokens for generating community summaries, community answers, and global answers (explained in Appendix C). Graph indexing with a 600 token window (explained in Section A.2) took 281 minutes for the Podcast dataset, running on a virtual machine (16GB RAM, Intel(R) Xeon(R) Platinum 8171M CPU @ 2.60GHz) and using a public OpenAI endpoint for gpt-4-turbo (2M TPM, 10k RPM).

我们对生成社区摘要、社区答案和全局答案使用固定的上下文窗口大小 8k 令牌 (详见附录 C)。对 Podcast 数据集进行 600 令牌的图索引 (详见附录 A.2) 耗时 281 分钟, 运行在一台虚拟机上 (16GB 内存, Intel(R) Xeon(R) Platinum 8171M CPU @ 2.60GHz), 并使用公共 OpenAI 端点的 gpt-4-turbo (2M TPM, 10k RPM)。

We implemented Leiden community detection using the graspologic library (Chung et al., 2019). The prompts used to generate the graph index and global answers can be found in Appendix E, while the prompts used to evaluate LLM responses against our criteria can be found in Appendix F. A full statistical analysis of the results presented in the next section can be found in Appendix G.

我们使用 graspologic 库 (Chung 等人, 2019) 实现了 Leiden 社区检测。用于生成图索引和全局答案的提示见附录 E, 用于根据我们的标准评估 LLM 回答的提示见附录 F。下一节中结果的完整统计分析见附录 G。

## 4.2 Experiment 2

### 4.2 实验 2

To validate the comprehensiveness and diversity results from Experiment 1, we implemented claim-based measures of these qualities. We use the definition of a factual claim from Ni et al. (2024), which is "a statement that explicitly presents some verifiable facts." For example, the sentence "California and New York implemented incentives for renewable energy adoption, highlighting the broader importance of sustainability in policy decisions" contains two factual claims: (1) California implemented incentives for renewable energy adoption, and (2) New York implemented incentives for renewable energy adoption.

为验证实验 1 的全面性和多样性结果, 我们实现了基于主张的度量。我们采用 Ni 等人 (2024) 对事实主张的定义, 即“明确表述可验证事实的陈述”。例如句子“加利福尼亚和纽约实施了可再生能源采纳激励措施, 凸显了可持续性在政策决策中的更广泛重要性”包含两个事实主张:(1) 加利福尼亚实施了可再生能源采纳激励措施, 和 (2) 纽约实施了可再生能源采纳激励措施。

To extract factual claims, we used Claimify (Metropolitansky and Larson, 2025), an LLM-based method that identifies sentences in an answer containing at least one factual claim, then decomposes these sentences into simple, self-contained factual claims. We applied Claimify to the answers generated under the conditions from Experiment 1. After removing duplicate claims from each answer, we extracted 47,075 unique claims,

with an average of 31 claims per answer.

为提取事实主张，我们使用了 Claimify(Metropolitansky 和 Larson, 2025)，这是一种基于 LLM 的方法，识别答案中包含至少一个事实主张的句子，然后将这些句子分解为简单的、独立的事实主张。我们将 Claimify 应用于实验 1 各条件下生成的答案。去除每个答案中的重复主张后，我们提取了 47,075 条唯一主张，平均每个答案 31 条主张。

We defined two metrics, with higher values indicating better performance:

我们定义了两个指标，数值越高表示性能越好:

1. Comprehensiveness: Measured as the average number of claims extracted from the answers generated under each condition.

1. 全面性: 衡量为每个条件生成的答案中提取主张的平均数量。

2. Diversity: Measured by clustering the claims for each answer and calculating the average number of clusters.

2. 多样性: 通过对每个答案的主张进行聚类并计算平均簇数来衡量。

For clustering, we followed the approach described by Padmakumar and He (2024), which involved using Scikit-learn’s implementation of agglomerative clustering (Pedregosa et al., 2011). Clusters were merged through “complete” linkage, meaning they were combined only if the maximum distance between their farthest points was less than or equal to a predefined distance threshold. The distance metric used was 1 - ROUGE-L. Since the distance threshold influences the number of clusters, we report results across a range of thresholds.

在聚类上，我们遵循 Padmakumar 和 He(2024) 描述的方法，使用 Scikit-learn 中的凝聚层次聚类实现 (Pedregosa 等人, 2011)。簇通过“complete”链接合并，即仅当两个簇最远点之间的最大距离小于或等于预设距离阈值时才合并。使用的距离度量为 1 - ROUGE-L。由于距离阈值会影响簇的数量，我们在一系列阈值下报告结果。

## 5 Results

### 5 结果

### 5.1 Experiment 1

#### 5.1 实验 1

The indexing process resulted in a graph consisting of 8,564 nodes and 20,691 edges for the Podcast dataset, and a larger graph of 15,754 nodes and 19,520 edges for the News dataset. Table 2 shows the number of community summaries at different levels of each graph community hierarchy.

索引过程为 Podcast 数据集生成了包含 8,564 个节点和 20,691 条边的图，对于 News 数据集则生成了更大的图，包含 15,754 个节点和 19,520 条边。表 2 显示了每个图社区层级在不同层次的社区摘要数量。

Global approaches vs. vector RAG. As shown in Figure 2 and Table 6, global approaches significantly outperformed conventional vector RAG (SS) in both comprehensiveness and diversity criteria across datasets. Specifically, global approaches achieved comprehensiveness win rates between 72- 83% ( $p<.001$ ) for Podcast transcripts and 72-80% ( $p<.001$ ) for News articles, while diversity win rates ranged from 75-82% ( $p<.001$ ) and 62-71% ( $p<.01$ ) respectively. Our use of directness as a validity test confirmed that vector RAG produces the most direct responses across all comparisons.

全局方法 vs. 向量 RAG。如图 2 和表 6 所示，全局方法在各数据集的全面性和多样性指标上均显著优于传统的向量 RAG(SS)。具体地，全局方法在 Podcast 转录上的全面性胜率为 72-83%( $p<.001$ )，在 News 文章上为 72-80%( $p<.001$ )，而多样性胜率在前者为 75-82%( $p<.001$ )，在后者为 62-71%( $p<.01$ )。我们将直接性用作有效性检验，确认在所有比较中向量 RAG 产生了最直接的回应。

Empowerment. Empowerment comparisons showed mixed results for both global approaches versus vector RAG (SS) and GraphRAG approaches versus source text summarization (TS). Using an LLM to analyze LLM reasoning for this measure indicated that the ability to provide specific examples, quotes, and citations was judged to be key to helping users reach an informed understanding. Tuning element extraction prompts may help to retain more of these details in the GraphRAG index.

授权。关于授权的比较显示，整体方法与向量 RAG(SS) 以及 GraphRAG 方法与源文本摘要 (TS) 之间结果喜忧参半。使用大型模型分析大型模型推理表明，提供具体示例、引用和引证被认为是帮助用户获得知情理解的关键。调整要素抽取提示可能有助于在 GraphRAG 索引中保留更多此类细节。

#### Podcast transcripts

	SS	TS	C0	C1	C2	C3		SS	TS	C0	C1	C2	C3		SS	TS	C0	C1	C2	C3		SS	TS	C0	C1	C2	C3
SS	50	17	28	25	22	21	SS	50	18	23	25	19	19	SS	50	42	57	52	49	51	SS	50	56	65	60	60	60
TS	83	50	50	48	43	44	TS	82	50	50	50	43	46	TS	58	50	59	55	52	51	TS	44	50	55	52	51	52
C0	72	50	50	53	50	49	C0	77	50	50	50	46	44	C0	43	41	50	49	47	48	C0	35	45	50	47	48	48
C1	75	52	47	50	52	50	C1	75	50	50	50	44	45	C1	48	45	51	50	49	50	C1	40	48	53	50	50	50
C2	78	57	50	48	50	52	C2	81	57	54	56	50	48	C2	51	48	53	51	50	51	C2	40	49	52	50	50	50
C3	79	56	51	50	48	50	C3	81	54	56	55	52	50	C3	49	49	52	50	49	50	C3	40	48	52	50	50	50
Comprehensiveness							Diversity							Empowerment							Directness						

#### News articles

	SS	TS	C0	C1	C2	C3		SS	TS	C0	C1	C2	C3		SS	TS	C0	C1	C2	C3		SS	TS	C0	C1	C2	C3
SS	50	20	28	25	21	21	SS	50	33	38	35	29	31	SS	50	47	57	49	50	50	SS	50	54	59	55	55	54
TS	80	50	44	41	38	36	TS	67	50	53	45	44	40	TS	53	50	58	50	50	48	TS	46	50	55	53	52	52
C0	72	56	50	52	54	52	C0	62	47	50	40	41	41	C0	43	42	50	42	45	44	C0	41	45	50	48	48	47
C1	75	59	48	50	58	55	C1	65	55	60	50	50	50	C1	51	50	58	50	52	51	C1	45	47	52	50	49	49
C2	79	62	46	42	50	59	C2	71	56	59	50	50	51	C2	50	50	55	48	50	50	C2	45	48	52	51	50	49
C3	79	64	48	45	41	50	C3	69	60	59	50	49	50	C3	50	52	56	49	50	50	C3	46	48	53	51	51	50
Comprehensiveness							Diversity							Empowerment							Directness						

Figure 2: Head-to-head win rate percentages of (row condition) over (column condition) across two datasets, four metrics, and 125 questions per comparison (each repeated five times and averaged). The overall winner per dataset and metric is shown in bold. Self-win rates were not computed but are shown as the expected 50% for reference. All Graph RAG conditions outperformed naïve RAG on comprehensiveness and diversity. Conditions C1-C3 also showed slight improvements in answer comprehensiveness and diversity over TS (global text summarization without a graph index).

图 2: 在两个数据集、四个指标以及每次比较 125 个问题 (每次重复五次并取平均) 中, (行条件) 相对于 (列条件) 的对决胜率百分比。每个数据集与指标的总获胜者以粗体显示。未计算自胜率, 但作为参考显示了预期的 50%。所有 Graph RAG 条件在全面性和多样性上均优于朴素 RAG。条件 C1-C3 在答案全面性和多样性上也较 TS(无图索引的全局文本摘要) 显示出轻微改进。

Table 2: Number of context units (community summaries for C0-C3 and text chunks for TS), corresponding token counts, and percentage of the maximum token count. Map-reduce summarization of source texts is the most resource-intensive approach requiring the highest number of context tokens. Root-level community summaries (C0) require dramatically fewer tokens per query (9x-43x).

表 2: 上下文单元数量 (C0-C3 的社区摘要和 TS 的文本块)、对应的 token 数以及占最大 token 数的百分比。对源文本进行 map-reduce 摘要是最耗资源的方法, 需要最多的上下文 token。根级社区摘要 (C0) 每次查询所需的 token 显著更少 (9 倍至 43 倍)。

Podcast Transcripts					News Articles					
	C0	C1	C2	C3	TS	C0	C1	C2	C3	TS
Units	34	367	969	1310	1669	55	555	1797	2142	3197
Tokens	26657	225756	565720	746100	1014611	39770	352641	980898	1140266	1707694
% Max	2.6	22.2	55.8	73.5	100	2.3	20.7	57.4	66.8	100

播客稿件					新闻文章					
	C0	C1	C2	C3	TS	C0	C1	C2	C3	TS
单位	34	367	969	1310	1669	55	555	1797	2142	3197
标记	26657	225756	565720	746100	1014611	39770	352641	980898	1140266	1707694
最大百分比	2.6	22.2	55.8	73.5	100	2.3	20.7	57.4	66.8	100

Community summaries vs. source texts. When comparing community summaries to source texts using GraphRAG, community summaries generally provided a small but consistent improvement in answer comprehensiveness and diversity, except for root-level summaries. Intermediate-level summaries in the Podcast dataset and low-level community summaries in the News dataset achieved comprehensiveness win rates of 57% ( $p < .001$ ) and 64% ( $p < .001$ ), respectively. Diversity win rates were 57% ( $p = .036$ ) for Podcast intermediate-level summaries and 60% ( $p < .001$ ) for News low-level community summaries. Table 2 also illustrates the scalability advantages of GraphRAG compared to source text summarization: for low-level community summaries (C3), GraphRAG required 26-33% fewer context tokens, while for root-level community summaries (C0), it required over 97% fewer tokens. For a modest drop in performance compared with other global methods, root-level GraphRAG offers a highly efficient method for the iterative question answering that characterizes sensemaking activity, while retaining advantages in comprehensiveness (72% win rate) and diversity (62% win rate) over vector RAG.

社区摘要与原文。使用 GraphRAG 将社区摘要与原文比较时，社区摘要通常在答案的全面性和多样性上带来小而稳定的提升，但根级摘要除外。Podcast 数据集中间级摘要和 News 数据集低级社区摘要分别在全面性上取得了 57%( $p<.001$ ) 和 64%( $p<.001$ ) 的胜率。多样性胜率为 Podcast 中间级摘要 57%( $p=.036$ ) 和 News 低级社区摘要 60%( $p<.001$ )。表 2 还展示了与原文摘要相比 GraphRAG 的可扩展性优势: 对于低级社区摘要 (C3)，GraphRAG 需要的上下文 token 少 26–33%，而对于根级社区摘要 (C0)，所需 token 则少于 97%。相比其他全局方法，根级 GraphRAG 在性能上仅有适度下降，但为表征意义构建活动的迭代问答提供了一种高效方法，同时在全面性 (72% 胜率) 和多样性 (62% 胜率) 上仍优于向量 RAG。

Table 3: Average number of extracted claims, reported by condition and dataset type. Bolded values represent the highest score in each column.

表 3: 按条件和数据集类型报告的平均提取主张数量。粗体值表示每列中的最高分。

Condition	Average Number of Claims	
	News Articles	Podcast Transcripts
C0	34.18	32.21
C1	32.50	32.20
C2	31.62	32.46
C3	33.14	32.28
TS	32.89	31.39
SS	25.23	26.50

状况	平均理赔次数	
	新闻文章	播客文字稿
C0	34.18	32.21
C1	32.50	32.20
C2	31.62	32.46
C3	33.14	32.28
TS	32.89	31.39
SS	25.23	26.50

## 5.2 Experiment 2

### 5.2 实验二

Table 3 shows the results for the average number of extracted claims (i.e., the claim-based measure of comprehensiveness) per condition. For both the News and Podcast datasets, all global search conditions (C0-C3) and source text summarization (TS) had greater comprehensiveness than vector RAG (SS). The differences were statistically significant ( $p<.05$ ) in all cases. These findings align with the LLM-based win rates from Experiment 1.

表 3 显示了每个条件下提取到的平均论点数 (即基于论点的全面性指标)。在新闻和播客数据集中, 所有全局检索条件 (C0-C3) 和源文本摘要 (TS) 的全面性均高于向量 RAG(SS)。在所有情况下差异在统计上均显著 ( $p < .05$ )。这些发现与实验一中基于大模型的胜出率一致。

Table 4 contains the results for the average number of clusters, the claim-based measure of diversity. For the Podcast dataset, all global search conditions had significantly greater diversity than SS across all distance thresholds ( $p < .05$ ), consistent with the win rates observed in Experiment 1. For the News dataset, however, only **C0** significantly outperformed SS across all distance thresholds ( $p < .05$ ). While C1-C3 also achieved higher average cluster counts than SS, the differences were statistically significant only at certain distance thresholds. In Experiment 1, all global search conditions significantly outperformed SS in the News dataset - not just C0. However, the differences in mean diversity scores between SS and the global search conditions were smaller for the News dataset than for the Podcast dataset, aligning directionally with the claim-based results.

表 4 列出了平均簇数 (基于论点的多样性指标) 的结果。对于播客数据集, 所有全局检索条件在所有距离阈值下相较于 SS 都具有显著更高的多样性 ( $p < .05$ ), 这与实验一中观察到的胜出率一致。然而, 对于新闻数据集, 只有 **C0** 在所有距离阈值下显著优于 SS ( $p < .05$ )。尽管 C1-C3 的平均簇数也高于 SS, 但这些差异仅在某些距离阈值上达到统计显著性。在实验一中, 新闻数据集中所有全局检索条件显著优于 SS——不仅仅是 C0。然而, SS 与全局检索条件之间的平均多样性得分差异在新闻数据集上比在播客数据集上更小, 这在方向上与基于论点的结果一致。

For both comprehensiveness and diversity, across both datasets, there were no statistically significant differences observed among the global search conditions or between global search and TS.

在两项指标 (全面性和多样性) 以及两个数据集中, 全局检索条件之间或全局检索与 TS 之间均未观察到统计显著差异。

Finally, for each pairwise comparison in Experiment 1, we tested whether the answer preferred by the LLM aligned with the winner based on the claim-based metrics. Since each pairwise comparison in Experiment 1 was performed five times, while the claim-based metrics provided only one outcome per comparison, we aggregated the Experiment 1 results into a single label using majority voting. For example, if **C0** won over SS in three out of five judgments for comprehensiveness on a given question, **C0** was labeled the winner and SS the loser. However, if **C0** won twice, SS won once, and they tied twice, then there was no majority outcome, so the final label was a tie.

最后, 对于实验一中的每一对比较, 我们测试了大模型偏好的答案是否与基于论点的胜者一致。由于实验一中的每对比较进行了五次, 而基于论点的指标每次比较只提供一个结果, 我们通过多数投票将实验一的结果合并为单一标签。例如, 如果在某个问题上 **C0** 在五次判断中有三次在全面性上胜过 SS, 则将 **C0** 标记为胜者, SS 为败者。然而, 如果 **C0** 赢了两次、SS 赢了 1 次、平局两次, 则没有多数结果, 最终标签为平局。

We found that exact ties were rare for the claim-based metrics. One possible solution is to define a tie based on a threshold (e.g., the absolute difference between the claim-based results for condition A and condition B must be less than or equal to  $x$ ). However, we observed that the results were sensitive to the choice of threshold. As a result, we focused on cases where the aggregated LLM label was not a tie, representing

33% and 39% of pairwise comparisons for comprehensiveness and diversity, respectively. In these cases, the aggregated LLM label matched the claim-based label in 78% of pairwise comparisons for comprehensiveness and 69-70% for diversity (across all distance thresholds), indicating moderately strong alignment.

我们发现基于论点的指标很少出现完全平局。一种可行的办法是基于阈值定义平局 (例如, 条件 A 与条件 B 的基于论点结果的绝对差必须小于或等于  $x$ )。不过, 我们观察到结果对阈值的选择很敏感。因此, 我们关注于聚合后大模型标签不是平局的情形, 这类情形在全面性和多样性的两两比较中分别占 33% 和 39%。在这些情形中, 聚合后大模型标签与基于论点的标签在全面性上匹配 78% 的两两比较, 在多样性上 (跨所有距离阈值) 匹配 69%-70%, 表明中等偏强的一致性。

## 6 Discussion

### 6 讨论

### 6.1 Limitations of evaluation approach

#### 6.1 评估方法的局限性

Our evaluation to date has focused on sensemaking questions specific to two corpora each containing approximately 1 million tokens. More work is needed to understand how performance generalizes to datasets from various domains with different use cases. Comparison of fabrication rates, e.g., using approaches like SelfCheckGPT (Manakul et al., 2023), would also strengthen the current analysis.

迄今为止, 我们的评估集中在针对两个语料库的理解性问题上, 每个语料库大约包含 100 万个标记。还需要更多工作以了解性能如何推广到来自不同领域、具有不同用例的数据集。比较虚构率, 例如使用 SelfCheckGPT(Manakul 等, 2023) 等方法, 也会增强当前分析。

Table 4: Average number of clusters across different distance thresholds, reported by condition and dataset type. Bolded values represent the highest score in each row.

表 4: 不同距离阈值下的平均簇数, 按条件和数据集类型报告。加粗值表示每行的最高得分。

Dataset	Distance Threshold	Average Number of Clusters					
		C0	C1	C2	C3	TS	SS
News Articles	0.5	23.42	21.85	21.90	22.13	21.80	17.92
	0.6	21.65	20.38	20.30	20.52	20.13	16.78
	0.7	20.19	19.06	19.03	19.13	18.62	15.80
	0.8	18.86	17.78	17.82	17.79	17.30	14.80
Podcast Transcripts	0.5	23.16	22.62	22.52	21.93	21.14	18.55
	0.6	21.65	21.33	21.21	20.62	19.70	17.39
	0.7	20.41	20.04	19.79	19.22	18.08	16.28
	0.8	19.26	18.77	18.46	17.89	16.66	15.07

数据集	距离阈值	平均簇数					
		C0	C1	C2	C3	TS	SS
新闻文章	0.5	23.42	21.85	21.90	22.13	21.80	17.92
	0.6	21.65	20.38	20.30	20.52	20.13	16.78
	0.7	20.19	19.06	19.03	19.13	18.62	15.80
	0.8	18.86	17.78	17.82	17.79	17.30	14.80
播客抄本	0.5	23.16	22.62	22.52	21.93	21.14	18.55
	0.6	21.65	21.33	21.21	20.62	19.70	17.39
	0.7	20.41	20.04	19.79	19.22	18.08	16.28
	0.8	19.26	18.77	18.46	17.89	16.66	15.07

## 6.2 Future work

### 6.2 未来工作

The graph index, rich text annotations, and hierarchical community structure supporting the current GraphRAG approach offer many possibilities for refinement and adaptation. This includes RAG approaches that operate in a more local manner, via embedding-based matching of user queries and graph annotations. In particular, we see potential in hybrid RAG schemes that combine embedding-based matching with just-in-time community report generation before employing our map-reduce summarization mechanisms. This "roll-up" approach could also be extended across multiple levels of the community hierarchy, as well as implemented as a more exploratory "drill down" mechanism that follows the information scent contained in higher-level community summaries.

图谱索引、富文本注释以及支持当前 GraphRAG 方法的层级社区结构提供了许多改进与适配的可能性。其中包括更本地化的 RAG 方法，通过基于嵌入的匹配来对用户查询与图注释进行比对。我们特别看好混合式 RAG 方案，将基于嵌入的匹配与在使用我们的 map-reduce 摘要机制前按需生成社区报告相结合。这种“汇总”方法也可以扩展到社区层级的多级，以及实现为一种更具探索性的“下钻”机制，沿着更高层社区摘要中包含的信息气味进行追踪。

Broader impacts. As a mechanism for question answering over large document collections, there are risks to downstream sensemaking and decision-making tasks if the generated answers do not accurately represent the source data. System use should be accompanied by clear disclosures of AI use and the potential for errors in outputs. Compared to vector RAG, however, GraphRAG shows promise as a way to mitigate these downstream risks for questions of a global nature, which might otherwise be answered by samples of retrieved facts falsely presented as global summaries.

更广泛的影响。作为针对大型文档集合的问答机制，如果生成的答案未能准确反映源数据，可能会对下游的理解与决策任务造成风险。系统使用应附带清晰的 AI 使用披露以及输出可能存在错误的提示。然而，与向量 RAG 相比，GraphRAG 在减缓此类全局性问题的下游风险方面展现出希望，否则这些问题可能通过被误呈为全局摘要的检索事实样本得到回答。

## 7 Conclusion

### 7 结论

We have presented GraphRAG, a RAG approach that combines knowledge graph generation and query-focused summarization (QFS) to support human sensemaking over entire text corpora. Initial evaluations show substantial improvements over a vector RAG baseline for both the comprehensiveness and diversity of answers, as well as favorable comparisons to a global but graph-free approach using map-reduce source text summarization. For situations requiring many global queries over the same dataset, summaries of root-level communities in the entity-based graph index provide a data index that is both superior to vector RAG and achieves competitive performance to other global methods at a fraction of the token cost.

我们提出了 GraphRAG，一种结合知识图生成与查询聚焦摘要 (QFS) 的 RAG 方法，以支持对整个文本语料库的人类理解。初步评估显示，相较于向量 RAG 基线，在答案的全面性与多样性上都有显著提升，并且与一种使用 map-reduce 源文本摘要的全局但无图方法相比也具有竞争力。对于需要对同一数据集进行大量全局查询的情况，基于实体的图谱索引中根级社区的摘要提供了一种既优于向量 RAG 又能以极低代价在 token 使用上达到与其他全局方法竞争的数据信息索引。

## Acknowledgements

### 致谢

We would also like to thank the following people who contributed to the work: Alonso Guevara Fernández, Amber Hoak, Andrés Morales Esquivel, Ben Cutler, Billie Rinaldi, Chris Sanchez, Chris Trevino, Christine Caggiano, David Tittsworth, Dayenne de Souza, Douglas Orbaker, Ed Clark, Gabriel Nieves-Ponce, Gaudy Blanco Meneses, Kate Lytvynets, Katy Smith, Mónica Carva-jal, Nathan Evans, Richard Ortega, Rodrigo Racanicci, Sarah Smith, and Shane Solomon.

我们还要感谢以下为本工作做出贡献的人:Alonso Guevara Fernández, Amber Hoak, Andrés Morales Esquivel, Ben Cutler, Billie Rinaldi, Chris Sanchez, Chris Trevino, Christine Caggiano, David Tittsworth, Dayenne de Souza, Douglas Orbaker, Ed Clark, Gabriel Nieves-Ponce, Gaudy Blanco Meneses, Kate Lytvynets, Katy Smith, Mónica Carva-jal, Nathan Evans, Richard Ortega, Rodrigo Racanicci, Sarah Smith 和 Shane Solomon。

## References

### 参考文献

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Al-tenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Al-tenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. (2023). Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. (2023). Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

Baek, J., Aji, A. F., and Saffari, A. (2023). Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. arXiv preprint arXiv:2306.04136.

Baek, J., Aji, A. F., and Saffari, A. (2023). Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. arXiv preprint arXiv:2306.04136.

Ban, T., Chen, L., Wang, X., and Chen, H. (2023). From query tools to causal architects: Harnessing large language models for advanced causal discovery from data.

Ban, T., Chen, L., Wang, X., and Chen, H. (2023). From query tools to causal architects: Harnessing large language models for advanced causal discovery from data.

Barlaug, N. and Gulla, J. A. (2021). Neural networks for entity matching: A survey. ACM Transactions on Knowledge Discovery from Data (TKDD), 15(3):1-37.

Barlaug, N. and Gulla, J. A. (2021). Neural networks for entity matching: A survey. ACM Transactions on Knowledge Discovery from Data (TKDD), 15(3):1-37.

Baumel, T., Eyal, M., and Elhadad, M. (2018). Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. arXiv preprint arXiv:1801.07704.

Baumel, T., Eyal, M., and Elhadad, M. (2018). Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. arXiv preprint arXiv:1801.07704.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10):P10008.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10):P10008.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33:1877-1901.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33:1877-1901.

Cheng, X., Luo, D., Chen, X., Liu, L., Zhao, D., and Yan, R. (2024). Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36.

Cheng, X., Luo, D., Chen, X., Liu, L., Zhao, D., and Yan, R. (2024). 提升自己: 带自我记忆的检索增强文本生成。 *Advances in Neural Information Processing Systems*, 36.

Christen, P. and Christen, P. (2012). *The data matching process*. Springer.

Christen, P. and Christen, P. (2012). 数据匹配过程。 Springer.

Chung, J., Pedigo, B. D., Bridgeford, E. W., Varjavand, B. K., Helm, H. S., and Vogelstein, J. T. (2019). Grasp: Graph statistics in python. *Journal of Machine Learning Research*, 20(158):1-7.

Chung, J., Pedigo, B. D., Bridgeford, E. W., Varjavand, B. K., Helm, H. S., and Vogelstein, J. T. (2019). Grasp: Python 中的图统计。 *Journal of Machine Learning Research*, 20(158):1-7.

Dang, H. T. (2006). DUC 2005: Evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, pages 48-55.

Dang, H. T. (2006). DUC 2005: 以问题为中心的摘要系统评估。收录于 *Task-Focused Summarization and Question Answering 研讨会论文集*, 页 48-55。

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2006). Duplicate record detection: A survey. *IEEE Transactions on knowledge and data engineering*, 19(1):1-16.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2006). 重复记录检测: 综述。 *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1-16.

Es, S., James, J., Espinosa-Anke, L., and Schockaert, S. (2023). Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Es, S., James, J., Espinosa-Anke, L., and Schockaert, S. (2023). RAGAS: 检索增强生成的自动化评估。 *arXiv preprint arXiv:2309.15217*.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004). Web-scale information extraction in knowitall: (preliminary results). In *Proceedings of the 13th International Conference on World Wide Web, WWW '04*, page 100-110, New York, NY, USA. Association for Computing Machinery.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004). KnowItAll 中的网络规模信息抽取:(初步结果)。收录于第 13 届万维网国际会议论文集, WWW '04, 页 100-110, 纽约, NY, USA。 Association for Computing Machinery.

Feng, Z., Feng, X., Zhao, D., Yang, M., and Qin, B. (2023). Retrieval-generation synergy augmented large language models. *arXiv preprint arXiv:2310.05149*.

Feng, Z., Feng, X., Zhao, D., Yang, M., and Qin, B. (2023). 检索-生成协同增强的大型语言模型。arXiv preprint arXiv:2310.05149.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3-5):75-174.

Fortunato, S. (2010). 图中的社区发现。 *Physics Reports*, 486(3-5):75-174.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. (2023). 面向大型语言模型的检索增强生成: 综述。arXiv preprint arXiv:2312.10997.

He, X., Tian, Y., Sun, Y., Chawla, N. V., Laurent, T., LeCun, Y., Bresson, X., and Hooi, B. (2024). G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. arXiv preprint arXiv:2402.07630.

He, X., Tian, Y., Sun, Y., Chawla, N. V., Laurent, T., LeCun, Y., Bresson, X., and Hooi, B. (2024). G-retriever: 用于文本图理解与问答的检索增强生成。arXiv preprint arXiv:2402.07630.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. (2023). Large language models cannot self-correct reasoning yet. arXiv preprint arXiv:2310.01798.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. (2023). 大型语言模型尚不能自我纠正推理。arXiv preprint arXiv:2310.01798.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONE* 9(6): e98679. <https://doi.org/10.1371/journal.pone.0098679>.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, 一种为 Gephi 软件设计的连续图布局算法, 便于网络可视化。 *PLoS ONE* 9(6): e98679. <https://doi.org/10.1371/journal.pone.0098679>.

Jin, D., Yu, Z., Jiao, P., Pan, S., He, D., Wu, J., Philip, S. Y., and Zhang, W. (2021). A survey of community detection approaches: From statistical modeling to deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1149-1170.

Jin, D., Yu, Z., Jiao, P., Pan, S., He, D., Wu, J., Philip, S. Y., and Zhang, W. (2021). 社区发现方法综述: 从统计建模到深度学习。 *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1149-1170.

Kang, M., Kwak, J. M., Baek, J., and Hwang, S. J. (2023). Knowledge graph-augmented language models for knowledge-grounded dialogue generation. arXiv preprint arXiv:2305.18846.

Kang, M., Kwak, J. M., Baek, J., and Hwang, S. J. (2023). 用于知识驱动对话生成的知识图谱增强语言模型。arXiv preprint arXiv:2305.18846.

Khattab, O., Santhanam, K., Li, X. L., Hall, D., Liang, P., Potts, C., and Zaharia, M. (2022). Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. arXiv preprint arXiv:2212.14024.

Khattab, O., Santhanam, K., Li, X. L., Hall, D., Liang, P., Potts, C., and Zaharia, M. (2022). Demonstrate-search-predict: 将检索与语言模型组合用于知识密集型 NLP。arXiv preprint arXiv:2212.14024.

Kim, D., Xie, L., and Ong, C. S. (2016). Probabilistic knowledge graph construction: Compositional and incremental approaches. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, page 2257-2262, New York, NY, USA. Association for Computing Machinery.

Kim, D., Xie, L., and Ong, C. S. (2016). 概率知识图构建: 组合与增量方法。见 Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, 页 2257-2262, 纽约, NY, 美国。Association for Computing Machinery.

Kim, G., Kim, S., Jeon, B., Park, J., and Kang, J. (2023). Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. arXiv preprint arXiv:2310.14696.

Kim, G., Kim, S., Jeon, B., Park, J., and Kang, J. (2023). 澄清树: 用检索增强的大型语言模型回答模糊问题。arXiv 预印本 arXiv:2310.14696.

Klein, G., Moon, B., and Hoffman, R. R. (2006). Making sense of sensemaking 1: Alternative perspectives. IEEE intelligent systems, 21(4):70-73.

Klein, G., Moon, B., and Hoffman, R. R. (2006). 理解 sensemaking 1: 替代视角。IEEE intelligent systems, 21(4):70-73。

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. Proceedings of the National Academy of Sciences, 121(45):e2405460121.

Kosinski, M. (2024). 在心智理论任务中评估大型语言模型。Proceedings of the National Academy of Sciences, 121(45):e2405460121。

Kuratov, Y., Bulatov, A., Anokhin, P., Sorokin, D., Sorokin, A., and Burtsev, M. (2024). In search of needles in a 11 m haystack: Recurrent memory finds what llms miss.

Kuratov, Y., Bulatov, A., Anokhin, P., Sorokin, D., Sorokin, A., and Burtsev, M. (2024). 在一堆 11 m 干草中寻找针: 循环记忆发现大型模型遗漏的内容。

LangChain (2024). Langchain graphs. <https://langchain-graphrag.readthedocs.io/en/latest/>.

LangChain (2024). Langchain graphs. <https://langchain-graphrag.readthedocs.io/en/latest/>。

Laskar, M. T. R., Hoque, E., and Huang, J. (2020). Query focused abstractive summarization via incorporating query relevance and transfer learning with transformer models. In *Advances in Artificial Intelligence: 33rd Canadian Conference on Artificial Intelligence, Canadian AI 2020, Ottawa, ON, Canada, May 13-15, 2020, Proceedings 33*, pages 342-348. Springer.

Laskar, M. T. R., Hoque, E., and Huang, J. (2020). 通过结合查询相关性和基于 Transformer 的迁移学习进行查询聚焦的摘要生成。见 *Advances in Artificial Intelligence: 33rd Canadian Conference on Artificial Intelligence, Canadian AI 2020, Ottawa, ON, Canada, May 13-15, 2020, Proceedings 33*, 页 342-348。Springer。

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459-9474.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). 用于知识密集型 NLP 任务的检索增强生成。 *Advances in Neural Information Processing Systems*, 33:9459-9474。

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2023). Lost in the middle: How language models use long contexts. *arXiv:2307.03172*.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2023). 迷失在中间: 语言模型如何使用长上下文。 *arXiv:2307.03172*。

LlamaIndex (2024). GraphRAG Implementation with LlamaIndex - V2. [https://github.com/run-llama/llama\\_index/blob/](https://github.com/run-llama/llama_index/blob/)

LlamaIndex (2024). 使用 LlamaIndex 的 GraphRAG 实现 - V2. [https://github.com/run-llama/llama\\_index/blob/main/docs/docs/examples/cookbooks/GraphRAG\\_v2.ipynb](https://github.com/run-llama/llama_index/blob/main/docs/docs/examples/cookbooks/GraphRAG_v2.ipynb)。

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhu-moye, S., Yang, Y., et al. (2024). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhu-moye, S., Yang, Y., et al. (2024). Self-refine: 基于自我反馈的迭代精炼。 *Advances in Neural Information Processing Systems*, 36。

Manakul, P., Liusie, A., and Gales, M. J. (2023). Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Manakul, P., Liusie, A., and Gales, M. J. (2023). Selfcheckgpt: 面向生成型大型语言模型的零资源黑盒幻觉检测。 *arXiv 预印本 arXiv:2303.08896*。

Mao, Y., He, P., Liu, X., Shen, Y., Gao, J., Han, J., and Chen, W. (2020). Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*.

Mao, Y., He, P., Liu, X., Shen, Y., Gao, J., Han, J., and Chen, W. (2020). 为开放域问答设计的生成增强检索。arXiv 预印本 arXiv:2009.08553。

Martin, S., Brown, W. M., Klavans, R., and Boyack, K. (2011). Openord: An open-source toolbox for large graph layout. SPIE Conference on Visualization and Data Analysis (VDA).

Martin, S., Brown, W. M., Klavans, R., and Boyack, K. (2011). Openord: 用于大图布局的开源工具箱。SPIE Conference on Visualization and Data Analysis (VDA)。

Melnyk, I., Dognin, P., and Das, P. (2022). Knowledge graph generation from text.

Melnyk, I., Dognin, P., and Das, P. (2022). 从文本生成知识图谱。

Metropolitansky, D. and Larson, J. (2025). Towards effective extraction and evaluation of factual claims.

Metropolitansky, D. and Larson, J. (2025). 朝着有效提取与评估事实性断言的方向。

Microsoft (2023). The impact of large language models on scientific discovery: a preliminary study using gpt-4.

Microsoft (2023). 大型语言模型对科学发现的影响: 使用 gpt-4 的初步研究。

Mooney, R. J. and Bunescu, R. (2005). Mining knowledge from text using information extraction. SIGKDD Explor. Newsl., 7(1):3-10.

Mooney, R. J. and Bunescu, R. (2005). 使用信息抽取从文本中挖掘知识。SIGKDD Explor. Newsl., 7(1):3-10。

NebulaGraph (2024). Nebulagraph launches industry-first graph rag: Retrieval-augmented generation with llm based on knowledge graphs. <https://www.nebula-graph.io/posts/graph-RAG>.

NebulaGraph (2024). Nebulagraph 发布业界首个图谱 RAG: 基于知识图谱的检索增强生成与 LLM。 <https://www.nebula-graph.io/posts/graph-RAG>。

Neo4J (2024). Get started with graphrag: Neo4j's ecosystem tools. <https://neo4j.com/developer-blog/graphrag-ecosystem-tools/>.

Neo4J (2024). 开始使用 GraphRAG:Neo4j 的生态工具。 <https://neo4j.com/developer-blog/graphrag-ecosystem-tools/>。

Newman, M. E. (2006). Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23):8577-8582.

Newman, M. E. (2006). 网络中的模块性与社区结构。美国国家科学院院刊, 103(23):8577-8582。

Ni, J., Shi, M., Stambach, D., Sachan, M., Ash, E., and Leippold, M. (2024). AFaCTA: Assisting the annotation of factual claim detection with reliable LLM annotators. In Ku, L.-W., Martins, A., and Srikumar, V., editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1890-1912, Bangkok, Thailand. Association for Computational Linguistics.

Ni, J., Shi, M., Stambach, D., Sachan, M., Ash, E., and Leippold, M. (2024). AFaCTA: 使用可靠的 LLM 注释者辅助事实性主张检测的标注。收入 Ku, L.-W., Martins, A., and Srikumar, V. 主编,《第 62 届计算语言学协会年会会议录 (第 1 卷: 长篇论文)》, 第 1890-1912 页, 曼谷, 泰国。计算语言学协会。

OpenAI (2023). Chatgpt: Gpt-4 language model.

OpenAI (2023). Chatgpt:GPT-4 语言模型。

Padmakumar, V. and He, H. (2024). Does writing with language models reduce content diversity? ICLR.

Padmakumar, V. and He, H. (2024). 使用语言模型写作是否降低内容多样性? ICLR。

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12:2825-2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn:Python 中的机器学习。Journal of Machine Learning Research, 12:2825-2830.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., and Shoham, Y. (2023). In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics, 11:1316-1331.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., and Shoham, Y. (2023). 上下文内检索增强语言模型。Transactions of the Association for Computational Linguistics, 11:1316-1331。

Ranade, P. and Joshi, A. (2023). Fabula: Intelligence report generation using retrieval-augmented narrative construction. arXiv preprint arXiv:2310.13848.

Ranade, P. and Joshi, A. (2023). Fabula: 使用检索增强叙事构建的情报报告生成。arXiv 预印本 arXiv:2310.13848。

Salminen, J., Liu, C., Pian, W., Chi, J., Häyhänen, E., and Jansen, B. J. (2024). Deus ex machina and personas from large language models: Investigating the composition of ai-generated persona descriptions. In Proceedings of the CHI Conference on Human Factors in Computing Systems, pages 1-20.

Salminen, J., Liu, C., Pian, W., Chi, J., Häyhänen, E., and Jansen, B. J. (2024). 机器之神与来自大型语言模型的人设: 调查 AI 生成人设描述的构成。收入《CHI 人机交互大会论文集》, 第 1-20 页。

Sarathi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., and Manning, C. D. (2024). Raptor: Recursive abstractive processing for tree-organized retrieval. arXiv preprint arXiv:2401.18059.

Sarathi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., and Manning, C. D. (2024). Raptor: 用于树状检索的递归抽象处理。arXiv 预印本 arXiv:2401.18059。

Scott, K. (2024). Behind the Tech. <https://www.microsoft.com/en-us/behind-the-tech>.

Scott, K. (2024). Behind the Tech. <https://www.microsoft.com/en-us/behind-the-tech>。

Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N., and Chen, W. (2023). Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. arXiv preprint arXiv:2305.15294.

Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N., and Chen, W. (2023). 通过迭代检索-生成协同增强检索增强大型语言模型。arXiv 预印本 arXiv:2305.15294。

Shin, J., Hedderich, M. A., Rey, B. J., Lucero, A., and Oulasvirta, A. (2024). Understanding human-ai workflows for generating personas. In Proceedings of the 2024 ACM Designing Interactive Systems Conference, pages 757-781.

Shin, J., Hedderich, M. A., Rey, B. J., Lucero, A., and Oulasvirta, A. (2024). 理解用于生成人设的人机协作 workflow。收入《2024 年 ACM 交互设计系统大会论文集》, 第 757-781 页。

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. (2024). Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 36.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. (2024). Reflexion: 具有口头强化学习的语言代理。Advances in Neural Information Processing Systems, 36.

Su, D., Xu, Y., Yu, T., Siddique, F. B., Barezi, E. J., and Fung, P. (2020). Caire-covid: A question answering and query-focused multi-document summarization system for covid-19 scholarly information management. arXiv preprint arXiv:2005.03975.

Su, D., Xu, Y., Yu, T., Siddique, F. B., Barezi, E. J., and Fung, P. (2020). Caire-covid: 用于 COVID-19 学术信息管理的问题回答与面向查询的多文档摘要系统。arXiv preprint arXiv:2005.03975.

Tan, Z., Zhao, X., and Wang, W. (2017). Representation learning of large-scale knowledge graphs via entity feature combinations. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, page 1777-1786, New York, NY, USA. Association for Computing Machinery.

Tan, Z., Zhao, X., and Wang, W. (2017). 通过实体特征组合进行大规模知识图表示学习。收录于 2017 年 ACM 信息与知识管理大会论文集, CIKM '17, 第 1777-1786 页, 纽约, NY, USA。美国计算机学会。

Tang, Y. and Yang, Y. (2024). MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. arXiv preprint arXiv:2401.15391.

Tang, Y. and Yang, Y. (2024). MultiHop-RAG: 对多跳查询的检索增强生成进行基准测试。arXiv preprint arXiv:2401.15391.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: 开放基础与微调对话模型。arXiv preprint arXiv:2307.09288.

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports, 9(1).

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). 从 Louvain 到 Leiden: 保证连通良好的社区。Scientific Reports, 9(1).

Trajanoska, M., Stojanov, R., and Trajanov, D. (2023). Enhancing knowledge graph construction using large language models. ArXiv, abs/2305.04676.

Trajanoska, M., Stojanov, R., and Trajanov, D. (2023). 利用大型语言模型增强知识图构建。ArXiv, abs/2305.04676.

Trivedi, H., Balasubramanian, N., Khot, T., and Sabharwal, A. (2022). Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. arXiv preprint arXiv:2212.10509.

Trivedi, H., Balasubramanian, N., Khot, T., and Sabharwal, A. (2022). 在链式思维推理中交织检索以解答知识密集型多步问题。arXiv preprint arXiv:2212.10509.

Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., Xu, J., Qu, J., and Zhou, J. (2023a). Is chatgpt a good nlg evaluator? a preliminary study. arXiv preprint arXiv:2303.04048.

Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., Xu, J., Qu, J., and Zhou, J. (2023a). ChatGPT 是一个好的自然语言生成评估者吗? 初步研究。arXiv preprint arXiv:2303.04048.

Wang, S., Khramtsova, E., Zhuang, S., and Zuccon, G. (2024). Feb4rag: Evaluating federated search in the context of retrieval augmented generation. arXiv preprint arXiv:2402.11891.

Wang, S., Khramtsova, E., Zhuang, S., and Zuccon, G. (2024). Feb4rag: 在检索增强生成背景下评估联邦搜索。arXiv preprint arXiv:2402.11891.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2022). 自洽性提升语言模型的链式思维推理. arXiv preprint arXiv:2203.11171.

Wang, Y., Lipka, N., Rossi, R. A., Siu, A., Zhang, R., and Derr, T. (2023b). Knowledge graph prompting for multi-document question answering.

Wang, Y., Lipka, N., Rossi, R. A., Siu, A., Zhang, R., and Derr, T. (2023b). 用于多文档问答的知识图提示。

Xu, Y. and Lapata, M. (2021). Text summarization with latent queries. arXiv preprint arXiv:2106.00104.

Xu, Y. and Lapata, M. (2021). 带潜在查询的文本摘要. arXiv preprint arXiv:2106.00104.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Conference on Empirical Methods in Natural Language Processing (EMNLP).

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. (2018). HotpotQA: 一个用于多样、可解释的多跳问答的数据集. 收录于自然语言处理实证方法会议 (EMNLP)。

Yao, J.-g., Wan, X., and Xiao, J. (2017). Recent advances in document summarization. Knowledge and Information Systems, 53:297-336.

Yao, J.-g., Wan, X., and Xiao, J. (2017). 文档摘要的最新进展. Knowledge and Information Systems, 53:297-336.

Yao, L., Peng, J., Mao, C., and Luo, Y. (2023). Exploring large language models for knowledge graph completion.

Yao, L., Peng, J., Mao, C., and Luo, Y. (2023). 探索用于知识图补全的大型语言模型。

Yates, A., Banko, M., Broadhead, M., Cafarella, M., Etzioni, O., and Soderland, S. (2007). Tex-tRunner: Open information extraction on the web. In Carpenter, B., Stent, A., and Williams, J. D., editors, Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pages 25-26, Rochester, New York, USA. Association for Computational Linguistics.

Yates, A., Banko, M., Broadhead, M., Cafarella, M., Etzioni, O., and Soderland, S. (2007). Tex-tRunner: Open information extraction on the web. In Carpenter, B., Stent, A., and Williams, J. D., editors, Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pages 25-26, Rochester, New York, USA. Association for Computational Linguistics.

Yuan, X., Li, J., Wang, D., Chen, Y., Mao, X., Huang, L., Xue, H., Wang, W., Ren, K., and Wang, J. (2024). S-eval: Automatic and adaptive test generation for benchmarking safety evaluation of large language models.

arXiv preprint arXiv:2405.14191.

Yuan, X., Li, J., Wang, D., Chen, Y., Mao, X., Huang, L., Xue, H., Wang, W., Ren, K., and Wang, J. (2024). S-eval: Automatic and adaptive test generation for benchmarking safety evaluation of large language models. arXiv preprint arXiv:2405.14191.

Zhang, J. (2023). Graph-toolformer: To empower llms with graph reasoning ability via prompt augmented by chatgpt. arXiv preprint arXiv:2304.11116.

Zhang, J. (2023). Graph-toolformer: To empower llms with graph reasoning ability via prompt augmented by chatgpt. arXiv preprint arXiv:2304.11116.

Zhang, Y., Zhang, Y., Gan, Y., Yao, L., and Wang, C. (2024a). Causal graph discovery with retrieval-augmented generation based large language models. arXiv preprint arXiv:2402.15301.

Zhang, Y., Zhang, Y., Gan, Y., Yao, L., and Wang, C. (2024a). Causal graph discovery with retrieval-augmented generation based large language models. arXiv preprint arXiv:2402.15301.

Zhang, Z., Chen, J., and Yang, D. (2024b). Darg: Dynamic evaluation of large language models via adaptive reasoning graph. arXiv preprint arXiv:2406.17271.

Zhang, Z., Chen, J., and Yang, D. (2024b). Darg: Dynamic evaluation of large language models via adaptive reasoning graph. arXiv preprint arXiv:2406.17271.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, D., Xing, E., et al. (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, D., Xing, E., et al. (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36.

Zhu, Y., Wang, X., Chen, J., Qiao, S., Ou, Y., Yao, Y., Deng, S., Chen, H., and Zhang, N. (2024). Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities.

Zhu, Y., Wang, X., Chen, J., Qiao, S., Ou, Y., Yao, Y., Deng, S., Chen, H., and Zhang, N. (2024). Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities.

## A Entity and Relationship Extraction Approach

### A 实体与关系抽取方法

The following prompts, designed for GPT-4, are used in the default GraphRAG initialization pipeline:

以下为在默认 GraphRAG 初始化流水线中为 GPT-4 设计使用的提示:

- Default Graph Extraction Prompt

- 默认图谱抽取提示

- Claim Extraction Prompt

- 论断抽取提示

## A.1 Entity Extraction

### A.1 实体抽取

We do this using a multipart LLM prompt that first identifies all entities in the text, including their name, type, and description, before identifying all relationships between clearly related entities, including the source and target entities and a description of their relationship. Both kinds of element instance are output in a single list of delimited tuples.

我们使用一个多部分的 LLM 提示，先识别文本中所有实体，包括其名称、类型和描述，然后识别明显相关实体之间的所有关系，包括源实体与目标实体及其关系描述。两类元素实例都以单一的分隔元组列表输出。

## A.2 Self-Reflection

### A.2 自我反思

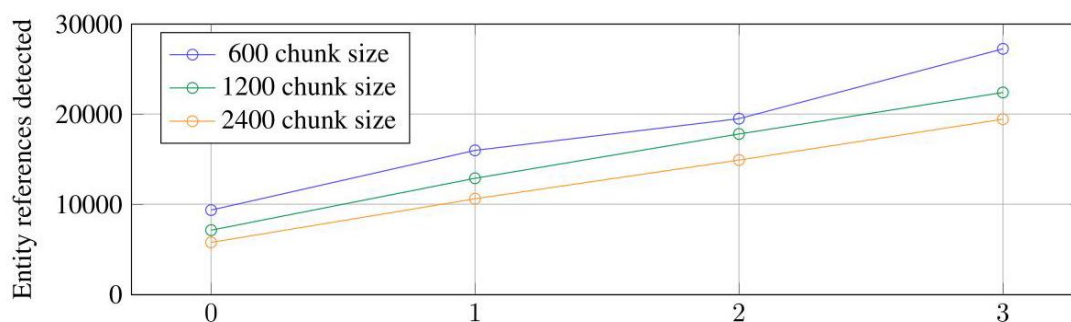
The choice of prompt engineering techniques has a strong impact on the quality of knowledge graph extraction (Zhu et al., 2024), and different techniques have different costs in terms of tokens consumed and generated by the model. Self-reflection is a prompt engineering technique where the LLM generates an answer, and is then prompted to evaluate its output for correctness, clarity, or completeness, then finally generate an improved response based on that evaluation (Huang et al., 2023; Madaan et al., 2024; Shinn et al., 2024; Wang et al., 2022). We leverage self-reflection in knowledge graph extraction, and explore ways how removing self-reflection affects performance and cost.

提示工程技术的选择对知识图谱抽取质量有重要影响 (Zhu et al., 2024)，不同技术在模型消耗与生成的 token 数上代价不同。自我反思是一种提示工程技术，LLM 先生成答案，然后被提示评估其输出的正确性、清晰度或完整性，最后基于该评估生成改进的回答 (Huang et al., 2023; Madaan et al., 2024; Shinn et al., 2024; Wang et al., 2022)。我们在知识图谱抽取中采用自我反思，并探讨去除自我反思如何影响性能与成本。

Using larger chunk size is less costly in terms of calls to the LLM. However, the LLM tends to extract few entities from chunks of larger size. For example, in a sample dataset (HotPotQA, Yang et al., 2018), GPT-4 extracted almost twice as many entity references when the chunk size was 600 tokens than when it was 2400. To address this issue, we deploy a self-reflection prompt engineering approach. After entities are extracted from a chunk, we provide the extracted entities back to the LLM, prompting it to "glean" any entities that it

may have missed. This is a multi-stage process in which we first ask the LLM to assess whether all entities were extracted, using a logit bias of 100 to force a yes/no decision. If the LLM responds that entities were missed, then a continuation indicating that "MANY entities were missed in the last extraction" encourages the LLM to detect these missing entities. This approach allows us to use larger chunk sizes without a drop in quality (Figure 3) or the forced introduction of noise. We iterate self-reflection steps up to a specified maximum number of times.

使用更大的 chunk 大小在调用 LLM 的次数上成本更低。然而，LLM 倾向于从较大 chunk 中抽取较少实体。例如，在一个样本数据集 (HotPotQA, Yang et al., 2018) 中，当 chunk 大小为 600 token 时，GPT-4 抽取的实体引用几乎是 2400 token 时的两倍。为了解决此问题，我们部署了自我反思的提示工程方法。在从一个 chunk 中抽取实体后，我们将抽取到的实体反馈给 LLM，提示其“揪出”可能遗漏的实体。这是一个多阶段过程：首先我们要求 LLM 评估是否已抽取所有实体，并使用 100 的 logit bias 强制其做出是/否决定。如果 LLM 回应说有实体被遗漏，则通过继续提示“在上次抽取中遗漏了大量实体”来鼓励 LLM 检测这些遗漏实体。该方法使我们能够在不降低质量 (图 3) 或强制引入噪声的情况下使用更大 chunk 大小。我们将自我反思步骤迭代到指定的最大次数。



Number of self-reflection iterations performed

执行的自我反思迭代次数

Figure 3: How the entity references detected in the HotPotQA dataset (Yang et al., 2018) varies with chunk size and self-reflection iterations for our generic entity extraction prompt with gpt-4-turbo.

图 3: 在 HotPotQA 数据集 (Yang et al., 2018) 中检测到的实体引用如何随块大小和自我反思迭代次数变化，基于我们使用 gpt-4-turbo 的通用实体提取提示。

## B Example Community Detection

### B 示例社群检测

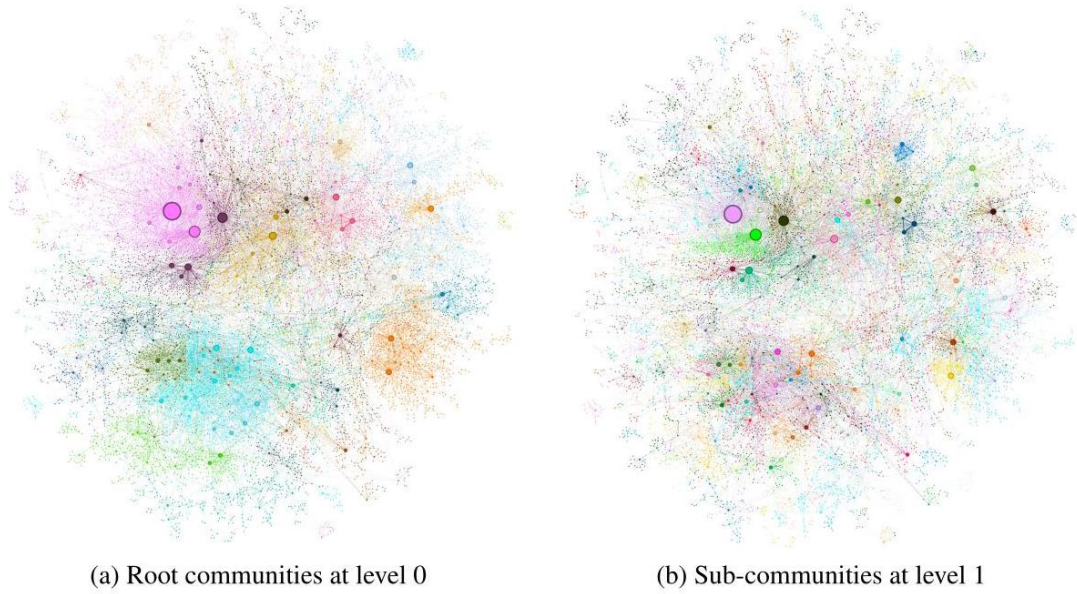


Figure 4: Graph communities detected using the Leiden algorithm (Traag et al., 2019) over the MultiHop-RAG (Tang and Yang, 2024) dataset as indexed. Circles represent entity nodes with size proportional to their degree. Node layout was performed via OpenORD (Martin et al., 2011) and Force Atlas 2 (Jacomy et al., 2014). Node colors represent entity communities, shown at two levels of hierarchical clustering: (a) Level 0, corresponding to the hierarchical partition with maximum modularity, and (b) Level 1, which reveals internal structure within these root-level communities.

图 4: 使用 Leiden 算法 (Traag et al., 2019) 在按索引的 MultiHop-RAG(Tang and Yang, 2024) 数据集上检测到的图社群。圆圈表示实体节点, 大小与节点度成比例。节点布局由 OpenORD(Martin et al., 2011) 和 Force Atlas 2(Jacomy et al., 2014) 完成。节点颜色表示实体社群, 展示两个层级的层次聚类:(a) 级别 0, 对应具有最大模块度的层次划分, 和 (b) 级别 1, 揭示这些根级社群内部结构。

## C Context Window Selection

### C 上下文窗口选择

The effect of context window size on any particular task is unclear, especially for models like gpt-4-turbo with a large context size of 128k tokens. Given the potential for information to be "lost in the middle" of longer contexts (Kuratov et al., 2024; Liu et al., 2023), we wanted to explore the effects of varying the context window size for our combinations of datasets, questions, and metrics. In particular, our goal was to determine the optimum context size for our baseline condition (SS) and then use this uniformly for all query-time LLM use. To that end, we tested four context window sizes: 8k, 16k, 32k and 64k. Surprisingly, the smallest context window size tested (8k) was universally better for all comparisons on comprehensiveness (average win rate of 58.1%), while performing comparably with larger context sizes on diversity (average win rate = 52.4%), and empowerment (average win rate = 51.3%). Given our preference for more comprehensive and diverse answers, we therefore used a fixed context window size of 8k tokens for the final evaluation.



—步骤—

1. Identify all entities. For each identified entity, extract the following information:

1. 识别所有实体。对于每个识别出的实体，提取以下信息：

- entity\_name: Name of the entity, capitalized

- entity\_name: 实体名称，首字母大写

- entity\_type: One of the following types: [{entity\_types}]

- entity\_type: 以下类型之一:[{entity\_types}]

- entity\_description: Comprehensive description of the entity's attributes and activities

- entity\_description: 对实体属性和活动的全面描述

Format each entity as ("entity"{tuple\_delimiter} <entity\_name>{tuple\_delimiter} <entity\_type>{tuple\_delimiter} <entity\_description>

将每个实体格式化为 ("entity"{tuple\_delimiter} <entity\_name>{tuple\_delimiter} <entity\_type>{tuple\_delimiter} <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

2. 在步骤 1 识别的实体中，识别所有彼此 \* 明确相关 \* 的 (source\_entity, target\_entity) 对

For each pair of related entities, extract the following information:

对于每一对相关实体，提取以下信息：

- source\_entity: name of the source entity, as identified in step 1

- source\_entity: 源实体的名称，如步骤 1 中所识别

- target\_entity: name of the target entity, as identified in step 1

- target\_entity: 目标实体的名称，如步骤 1 中所识别

- relationship\_description: explanation as to why you think the source entity and the target entity are related to each other

- relationship\_description: 解释为何你认为源实体与目标实体彼此相关

- `relationship_strength`: a numeric score indicating strength of the relationship between the source entity and target entity

- `relationship_strength`: 表示源实体与目标实体关系强度的数值评分

Format each relationship as ("`relationship`"`{tuple_delimiter}`<`source_entity`>`{tuple_delimiter}`<`target_entity`>`{tuple_delimiter}`<`relationship_description`>`{tuple_delimiter}`<`relationship_strength`>)

将每个关系格式化为 ("`relationship`"`{tuple_delimiter}`<`source_entity`>`{tuple_delimiter}`<`target_entity`>`{tuple_delimiter}`<`relationship_description`>`{tuple_delimiter}`<`relationship_strength`>)

3. Return output in English as a single list of all the entities and relationships identified in steps 1 and 2. Use `**{record_delimiter}` as the list delimiter.

3. 将输出以英文作为单一列表返回，包含步骤 1 和 2 中识别的所有实体和关系。使用 `**{record_delimiter}` 作为列表分隔符。

4. When finished, output `{completion_delimiter}`

4. 完成时，输出 `{completion_delimiter}`

—Examples—

—示例—

Entity\_types: ORGANIZATION, PERSON

Entity\_types: ORGANIZATION, PERSON

Input:

输入:

The Fed is scheduled to meet on Tuesday and Wednesday, with the central bank planning to release its latest policy decision on Wednesday at 2:00 p.m. ET, followed by a press conference where Fed Chair Jerome Powell will take questions. Investors expect the Federal Open Market Committee to hold its benchmark interest rate steady in a range of 5.25% – 5.5% .

美联储定于周二和周三召开会议,央行计划于周三美东时间下午 2:00 公布最新政策决定,随后召开新闻发布会,美联储主席杰罗姆·鲍威尔将回答提问。投资者预计联邦公开市场委员会将在 5.25%–5.5% 区间维持其基准利率不变。

Output:

输出:

("entity"{tuple\_delimiter}FED{tuple\_delimiter}ORGANIZATION{tuple\_delimiter}The Fed is the Federal Reserve, which is setting interest rates on Tuesday and Wednesday)

("entity"{tuple\_delimiter}FED{tuple\_delimiter}ORGANIZATION{tuple\_delimiter}The Fed is the Federal Reserve, which is setting interest rates on Tuesday and Wednesday)

{record\_delimiter}

{record\_delimiter}

("entity"{tuple\_delimiter} JEROME POWELL{tuple\_delimiter} PERSON{tuple\_delimiter} Jerome Powell is the chair of the Federal Reserve)

("entity"{tuple\_delimiter} JEROME POWELL{tuple\_delimiter} PERSON{tuple\_delimiter} Jerome Powell is the chair of the Federal Reserve)

{record\_delimiter}

{record\_delimiter}

("entity"{tuple\_delimiter}FEDERAL OPEN MARKET COMMITTEE{tuple\_delimiter}ORGANIZATION{tuple\_delimiter}

("entity"{tuple\_delimiter}FEDERAL OPEN MARKET COMMITTEE{tuple\_delimiter}ORGANIZATION{tuple\_delimiter}T

Federal Reserve committee makes key decisions about interest rates and the growth of the United States

Federal Reserve committee makes key decisions about interest rates and the growth of the United States

money supply)

money supply)

{record\_delimiter}

{record\_delimiter}

("relationship"{tuple\_delimiter}JEROME POWELL{{tuple\_delimiter}}FED{{tuple\_delimiter}}Or more Powell is the

("relationship"{tuple\_delimiter} 杰罗姆·鲍威尔 {{tuple\_delimiter}} 美联储 {{tuple\_delimiter}} 或者更确切地说, 鲍威尔是

Chair of the Federal Reserve and will answer questions at a press conference{tuple\_delimiter}9)

美联储主席, 并将在新闻发布会上回答问题 {tuple\_delimiter}9)

{completion\_delimiter}

{completion\_delimiter}

...More examples...

... 更多示例...

—Real Data—

—真实数据—

Entity\_types: {entity\_types}

实体类型: {entity\_types}

Input:

输入:

{input\_text}

{input\_text}

Output:

输出:

## E.2 Community Summary Generation

### E.2 社区摘要生成

—Role—

—角色—

You are an AI assistant that helps a human analyst to perform general information discovery. Information discovery is the process of identifying and assessing relevant information associated with certain entities (e.g., organizations and individuals) within a network.

你是一个帮助人类分析员进行一般信息发现的人工智能助理。信息发现是识别和评估与网络中某些实体 (例如组织和个人) 相关的相关信息的过程。

—Goal—

—目标—

Write a comprehensive report of a community, given a list of entities that belong to the community as well as their relationships and optional associated claims. The report will be used to inform decision-makers about information associated with the community and their potential impact. The content of this report includes an overview of the community's key entities, their legal compliance, technical capabilities, reputation, and noteworthy claims.

在给定属于某社区的实体列表及其关系和可选相关声明的情况下，撰写一份关于该社区的综合报告。该报告将用于向决策者通报与该社区相关的信息及其潜在影响。报告内容包括社区主要实体概况、其法律合规性、技术能力、声誉和值得注意的声明。

## —Report Structure—

### —报告结构—

The report should include the following sections:

报告应包括以下部分：

- **TITLE:** community's name that represents its key entities - title should be short but specific. When possible, include representative named entities in the title.

- **TITLE:** 社区名称，代表其关键实体 - 标题应简短但具体。若可能，包含具有代表性的命名实体。

- **SUMMARY:** An executive summary of the community's overall structure, how its entities are related to each other, and significant information associated with its entities.

- **SUMMARY:** 对社区整体结构、各实体之间的关系及与其实体相关的重要信息的执行摘要。

- **IMPACT SEVERITY RATING:** a float score between 0-10 that represents the severity of IMPACT posed by entities within the community. IMPACT is the scored importance of a community.

- **IMPACT SEVERITY RATING:** 一个介于 0-10 之间的浮点分数，表示社区内实体造成的影响严重程度。IMPACT 是对社区重要性的评分。

- **RATING EXPLANATION:** Give a single sentence explanation of the IMPACT severity rating.

- **RATING EXPLANATION:** 用一句话解释该 IMPACT 严重程度评分。

- **DETAILED FINDINGS:** A list of 5-10 key insights about the community. Each insight should have a short summary followed by multiple paragraphs of explanatory text grounded according to the grounding rules below. Be comprehensive.

- **DETAILED FINDINGS:** 列出 5-10 条关于该社区的关键洞见。每条洞见应有简短摘要，随后根据下述归因规则提供多段解释性文字并充分展开。

Return output as a well-formed JSON-formatted string with the following format:

Return output as a well-formed JSON-formatted string with the following format:

```
{{
```

```
"title": <report_title>,
```

```
"title": &lt;report_title&gt;,,
```

```
"summary": <executive_summary>,
```

```
"summary": &lt;executive_summary&gt;,,
```

```
"rating": <impact_severity_rating>,
```

```
"rating": &lt;impact_severity_rating&gt;,,
```

```
"rating_explanation": <rating_explanation>,
```

```
"rating_explanation": &lt;rating_explanation&gt;,,
```

```
"findings": [
```

```
"findings": [
```

```
{{
```

```
"summary":<insight_1_summary>,
```

```
"summary":&lt;insight_1_summary&gt;,,
```

```
"explanation": <insight_1_explanation>
```

```
"explanation": &lt;insight_1_explanation&gt;;
```

```
}},
```

```
{{
```

```
"summary":<insight_2_summary>,
```

```
"summary":&lt;insight_2_summary&gt;,,
```

```
"explanation": <insight_2_explanation>
```

```
"explanation": &lt;insight_2_explanation&gt;;
```

```
}}
```

```
]
```

```
]
```

}}

—Grounding Rules—

—基本准则—

Points supported by data should list their data references as follows:

有数据支持的论点应按如下方式列出其数据引用:

”This is an example sentence supported by multiple data references [Data: <dataset name> (record ids); <dataset name> (record ids)].”

“这是一个由多条数据引用支持的示例句子 [Data: <dataset name> (记录 id); <dataset name> (记录 id)].”

Do not list more than 5 record ids in a single reference. Instead, list the top 5 most relevant record ids and add “+more” to indicate that there are more.

单个引用中不要列出超过 5 个记录 id。应列出最相关的前 5 个记录 id，并添加 “+more” 以表明还有更多。

For example:

例如:

”Person X is the owner of Company Y and subject to many allegations of wrongdoing [Data: Reports (1),

“X 人是 Y 公司的所有者，并受到多项不当行为指控 [Data: Reports (1),

Entities (5, 7); Relationships (23); Claims (7, 2, 34, 64, 46, +more)].”

Entities (5, 7); Relationships (23); Claims (7, 2, 34, 64, 46, +more)].”

where 1, 5, 7, 23, 2, 34, 46, and 64 represent the id (not the index) of the relevant data record.

其中 1, 5, 7, 23, 2, 34, 46 和 64 表示相关数据记录的 id(不是索引)。

Do not include information where the supporting evidence for it is not provided.

不要包含没有提供支持证据的信息。

—Example—

—示例—

Input:

输入:

Entities

实体

id, entity, description

id, entity, description

5, VERDANT OASIS PLAZA, Verdant Oasis Plaza is the location of the Unity March

5, VERDANT OASIS PLAZA, Verdant Oasis Plaza 是 Unity March 的地点

6, HARMONY ASSEMBLY, Harmony Assembly is an organization that is holding a march at Verdant Oasis Plaza

6, HARMONY ASSEMBLY, Harmony Assembly 是一个在 Verdant Oasis Plaza 举行游行的组织

## Relationships

关系

id, source, target, description

id, source, target, description

37, VERDANT OASIS PLAZA, UNITY MARCH, Verdant Oasis Plaza is the location of the Unity March

37, VERDANT OASIS PLAZA, UNITY MARCH, Verdant Oasis Plaza 是 Unity March 的地点

38, VERDANT OASIS PLAZA, HARMONY ASSEMBLY, Harmony Assembly is holding a march at Verdant Oasis Plaza

38, VERDANT OASIS PLAZA, HARMONY ASSEMBLY, Harmony Assembly 正在 Verdant Oasis Plaza 举行一次游行

39, VERDANT OASIS PLAZA, UNITY MARCH, The Unity March is taking place at Verdant Oasis Plaza

39, VERDANT OASIS PLAZA, UNITY MARCH, The Unity March 正在 Verdant Oasis Plaza 举行

40, VERDANT OASIS PLAZA, TRIBUNE SPOTLIGHT, Tribune Spotlight is reporting on the Unity march taking place at Verdant Oasis Plaza

40, VERDANT OASIS PLAZA, TRIBUNE SPOTLIGHT, Tribune Spotlight 正在报道发生在 Verdant Oasis Plaza 的 Unity 游行

41, VERDANT OASIS PLAZA, BAILEY ASADI, Bailey Asadi is speaking at Verdant Oasis Plaza about the march

41, VERDANT OASIS PLAZA, BAILEY ASADI, Bailey Asadi 在 Verdant Oasis Plaza 就该游行发表演讲

43, HARMONY ASSEMBLY, UNITY MARCH, Harmony Assembly is organizing the Unity March

43, HARMONY ASSEMBLY, UNITY MARCH, Harmony Assembly 在组织 Unity March

Output:

Output:

```
{{
  "title": "Verdant Oasis Plaza and Unity March",
```

```
  "title": "Verdant Oasis Plaza and Unity March",
```

```
  "summary": "The community revolves around the Verdant Oasis Plaza, which is the location of the Unity March. The plaza has relationships with the Harmony Assembly, Unity March, and Tribune Spotlight, all of which are associated with the march event.",
```

```
  "summary": "社区围绕 Verdant Oasis Plaza 展开, 该广场是 Unity March 的地点。该广场与 Harmony Assembly、Unity March 及 Tribune Spotlight 有关联, 均与此次游行事件相关。",
```

```
  "rating": 5.0,
```

```
  "rating": 5.0,
```

```
  "rating_explanation": "The impact severity rating is moderate due to the potential for unrest or conflict during the Unity March.",
```

```
  "rating_explanation": "由于 Unity March 期间可能发生的不安或冲突, 影响严重性评估为中等。",
```

```
  "findings":
```

```
  "findings":
```

```
  {{
    "summary": "Verdant Oasis Plaza as the central location",
```

```
  "summary": "Verdant Oasis Plaza 作为中心地点",
```

"explanation": "Verdant Oasis Plaza is the central entity in this community, serving as the location for the Unity March. This plaza is the common link between all other entities, suggesting its significance in the community. The plaza's association with the march could potentially lead to issues such as public disorder or conflict, depending on the nature of the march and the reactions it provokes. [Data: Entities (5), Relationships (37, 38, 39, 40, 41,+more)]"

"explanation": "Verdant Oasis Plaza 是该社区的核心实体，为 Unity March 的举办地。该广场是所有其他实体的共同联系点，显示其在社区中的重要性。该广场与游行的关联可能会引发公共秩序或冲突等问题，具体取决于游行的性质及其引发的反应。[Data: Entities (5), Relationships (37, 38, 39, 40, 41,+more)]"

}},

{{

"summary": "Harmony Assembly's role in the community",

"summary": "Harmony Assembly's role in the community",

"explanation": "Harmony Assembly is another key entity in this community, being the organizer of the march at Verdant Oasis Plaza. The nature of Harmony Assembly and its march could be a potential source of threat, depending on their objectives and the reactions they provoke. The relationship between Harmony Assembly and the plaza is crucial in understanding the dynamics of this community. [Data: Entities(6), Relationships (38, 43)]"

"explanation": "和谐集会是该社区中的另一个关键实体，是在翠绿绿洲广场组织游行的发起者。和谐集会及其游行的性质可能构成潜在威胁，取决于其目标和引发的反应。理解和谐集会与广场之间的关系对于把握该社区的动态至关重要。[Data: Entities(6), Relationships (38, 43)]"

}},

{{

"summary": "Unity March as a significant event",

"summary": "统一游行作为一项重要事件",

"explanation": "The Unity March is a significant event taking place at Verdant Oasis Plaza. This event is a key factor in the community's dynamics and could be a potential source of threat, depending on the nature of the march and the reactions it provokes. The relationship between the march and the plaza is crucial in understanding the dynamics of this community. [Data: Relationships (39)]" }}

"explanation": "统一游行是在翠绿绿洲广场举行的一项重要活动。该事件是社区动态的关键因素，可能构成潜在威胁，取决于游行的性质及其引发的反应。理解游行与广场之间的关系对于把握该社区的动态至关重要。[Data: Relationships (39)]" }}

"summary": "Role of Tribune Spotlight", "explanation": "Tribune Spotlight is reporting on the Unity March taking place in Verdant Oasis Plaza. This suggests that the event has attracted media attention, which could amplify its impact on the community. The role of Tribune Spotlight could be significant in shaping public perception of the event and the entities involved. [Data: Relationships (40)]" }}

”summary”: ”论坛聚焦的角色”, ”explanation”: ”论坛聚焦正在报道发生在翠绿绿洲广场的统一游行。这表明该事件已吸引媒体关注, 可能放大其对社区的影响。论坛聚焦在塑造公众对该事件及相关实体看法方面可能具有重要作用。[Data: Relationships (40)]” }}

}}  
—Real Data—

—Real Data—

Use the following text for your answer. Do not make anything up in your answer.

使用以下文本作为你的回答。不要在回答中编造任何内容。

Input:

输入:

{input\_text}

{input\_text}

...Report Structure and Grounding Rules Repeated...

... 报告结构和依据规则重复中...

Output:

输出:

### E.3 Community Answer Generation

#### E.3 社区回答生成

—Role—

—角色—

You are a helpful assistant responding to questions about a dataset by synthesizing perspectives from multiple analysts.

你是一个通过综合多位分析师观点来回答关于数据集问题的有用助手。

—Goal—

—目标—

Generate a response of the target length and format that responds to the user's question, summarize all the reports from multiple analysts who focused on different parts of the dataset, and incorporate any relevant general knowledge.

生成符合目标长度和格式的回答，回应用户的问题，总结关注数据集不同部分的多位分析师的所有报告，并结合任何相关的一般知识。

Note that the analysts' reports provided below are ranked in the **\*\*descending order of helpfulness\*\***.

请注意，下方提供的分析师报告按有用性降序排列。

If you don't know the answer, just say so. Do not make anything up.

如果你不知道答案，就直说。不要编造任何内容。

The final response should remove all irrelevant information from the analysts' reports and merge the cleaned information into a comprehensive answer that provides explanations of all the key points and implications appropriate for the response length and format.

最终回复应当删去分析师报告中所有无关信息，并将清理后的信息合并成一份全面的答案，提供对所有关键点和影响的解释，且与回复的长度与格式相适应。

Add sections and commentary to the response as appropriate for the length and format. Style the response in markdown.

根据长度与格式的需要，为回复增加章节和评论。将回复以 markdown 风格呈现。

The response shall preserve the original meaning and use of modal verbs such as "shall", "may" or "will".

回复应当保留原意并保留诸如“shall”、“may”或“will”等情态动词的用法。

The response should also preserve all the data references previously included in the analysts' reports,

回复还应保留分析师报告中先前包含的所有数据引用，

but do not mention the roles of multiple analysts in the analysis process.

但不要提及多位分析师在分析过程中的角色。

Do not list more than 5 record ids in a single reference. Instead, list the top 5 most relevant record ids and add "+more" to indicate that there are more.

不要在单个引用中列出超过 5 个记录 id。相反，列出最相关的前 5 个记录 id，并添加“+more”以表示还有更多。

For example:

例如:

”Person X is the owner of Company Y and subject to many allegations of wrongdoing [Data: Reports (2, 7, 34, 46, 64, +more)]. He is also CEO of company X [Data: Reports (1, 3)]”

“Person X is the owner of Company Y and subject to many allegations of wrongdoing [Data: Reports (2, 7, 34, 46, 64, +more)]. He is also CEO of company X [Data: Reports (1, 3)]”

where 1, 2, 3, 7, 34, 46, and 64 represent the id (not the index) of the relevant data record.

其中 1, 2, 3, 7, 34, 46 和 64 表示相关数据记录的 id(不是索引)。

Do not include information where the supporting evidence for it is not provided.

不要包含支持证据未提供的信息。

—Target response length and format—

—目标回复长度与格式—

{response\_type}

{response\_type}

—Analyst Reports—

—分析师报告—

{report\_data}

{report\_data}

...Goal and Target response length and format repeated...

... 目标与目标回复长度和格式重复...

Add sections and commentary to the response as appropriate for the length and format. Style the response in markdown.

添加章节和评注到回复中，按长度和格式酌情调整。将回复样式设为 Markdown。

Output:

输出:

## E.4 Global Answer Generation

### E.4 全局答案生成

—Role—

—角色—

You are a helpful assistant responding to questions about data in the tables provided.

你是一名乐于助人的助手，回答关于所提供表格中数据的问题。

—Goal—

—目标—

Generate a response of the target length and format that responds to the user's question, summarize all relevant information in the input data tables appropriate for the response length and format, and incorporate any relevant general knowledge.

生成符合目标长度和格式的回复以回答用户问题，概括输入数据表中与回复长度和格式相称的所有相关信息，并结合任何相关的通用知识。

If you don't know the answer, just say so. Do not make anything up.

如果你不知道答案，就直接说明。不要编造任何内容。

The response shall preserve the original meaning and use of modal verbs such as "shall", "may" or "will".

回复应保留原意并保持情态动词的使用(如“shall”、“may”或“will”)。

Points supported by data should list the relevant reports as references as follows:

由数据支持的要点应列出相关报告作为参考，格式如下：

"This is an example sentence supported by data references [Data: Reports (report ids)]"

“这是一个由数据参考支持的示例句子 [Data: Reports (report ids)]”

Note: the prompts for SS (semantic search) and TS (text summarization) conditions use "Sources" in place of "Reports" above.

注意:SS(语义检索)和TS(文本摘要)条件的提示中用“Sources”代替上述的“Reports”。

Do not list more than 5 record ids in a single reference. Instead, list the top 5 most relevant record ids and add "+more" to indicate that there are more.

不要在单一参考中列出超过 5 个记录 id。应列出最相关的前 5 个记录 id 并加上 “+more” 以表示还有更多。

For example:

例如:

”Person X is the owner of Company Y and subject to many allegations of wrongdoing [Data: Reports (2, 7, 64, 46, 34, +more)]. He is also CEO of company X [Data: Reports (1, 3)]”

“某人 X 是公司 Y 的所有者，并受到多项不当行为指控 [Data: Reports (2, 7, 64, 46, 34, +more)]。他也是公司 X 的首席执行官 [Data: Reports (1, 3)]”

where 1,2,3,7,34,46, and 64 represent the id (not the index) of the relevant data report in the provided tables.

其中 1,2,3,7,34,46 和 64 代表所提供表格中相关数据报告的 id(不是索引)。

Do not include information where the supporting evidence for it is not provided.

不要包含没有提供支持证据的信息。

At the beginning of your response, generate an integer score between 0-100 that indicates how **helpful** is this response in answering the user’s question. Return the score in this format: <ANSWER\_HELPFULNESS> score\_value </ANSWER\_HELPFULNESS>.

在回答开始处生成一个介于 0 到 100 之间的整数分数，表示该回答在解答用户问题方面的有用程度。按此格式返回分数:<ANSWER\_HELPFULNESS> score\_value </ANSWER\_HELPFULNESS>。

—Target response length and format—

—目标回答长度与格式—

{response\_type}

{response\_type}

—Data tables—

—数据表—

{context\_data}

{context\_data}

...Goal and Target response length and format repeated...

... 目标与响应长度和格式重复...

Output:

输出:

## F Evaluation Prompts

### F 评估提示

#### F.1 Relative Assessment Prompt

##### F.1 相对评估提示

—Role—

—角色—

You are a helpful assistant responsible for grading two answers to a question that are provided by two different people.

你是一个负责为由两个人给出的两个答案评分的有用助手。

—Goal—

—目标—

Given a question and two answers (Answer 1 and Answer 2), assess which answer is better according to the following measure:

给定一个问题和两个答案 (答案 1 和答案 2)，根据以下准则评估哪个答案更好:

{criteria}

{criteria}

Your assessment should include two parts:

你的评估应包括两部分:

- Winner: either 1 (if Answer 1 is better) and 2 (if Answer 2 is better) or 0 if they are fundamentally similar and the differences are immaterial.

- 胜者: 若答案 1 更好则为 1, 若答案 2 更好则为 2, 若两者在本质上相同且差异无关紧要则为 0。

- Reasoning: a short explanation of why you chose the winner with respect to the measure described above.

- 理由: 简短说明你基于上述衡量标准为何选择该胜者。

Format your response as a JSON object with the following structure:

将你的回答格式化为具有以下结构的 JSON 对象:

```
{
  "winner": <1, 2, or 0>,
```

```
  "winner": <1, 2, or 0>,
```

```
  "reasoning": "Answer 1 is better because <your reasoning>."
```

```
  "reasoning": "Answer 1 is better because <your reasoning>."
```

```
}
—Question—
```

```
—问题—
```

```
{question}
```

```
{question}
```

```
—Answer 1—
```

```
—答案 1—
```

```
{answer1}
```

```
{answer1}
```

```
—Answer 2—
```

```
—答案 2—
```

```
{answer2}
```

```
{answer2}
```

Assess which answer is better according to the following measure:

根据以下衡量标准评估哪个答案更好:

{criteria}

{criteria}

Output:

输出:

## F.2 Relative Assessment Metrics

### F.2 相对评估指标

**CRITERIA = {**

**CRITERIA = {**

”comprehensiveness”: ”How much detail does the answer provide to cover all the aspects and details of the question? A comprehensive answer should be thorough and complete, without being redundant or irrelevant. For example, if the question is ’What are the benefits and drawbacks of nuclear energy?’, a comprehensive answer would provide both the positive and negative aspects of nuclear energy, such as its efficiency, environmental impact, safety, cost, etc. A comprehensive answer should not leave out any important points or provide irrelevant information. For example, an incomplete answer would only provide the benefits of nuclear energy without describing the drawbacks, or a redundant answer would repeat the same information multiple times.”,

”comprehensiveness”: ” 答案提供了多少细节以覆盖问题的各个方面和细节？一个全面的答案应当详尽完整，且不冗余或无关。例如，若问题是“核能的优点和缺点是什么？”，一个全面的答案会同时提供核能的正面和负面方面，如其效率、环境影响、安全性、成本等。全面的答案不应遗漏任何重要点或提供无关信息。例如，不完整的答案只给出核能的优点而不描述缺点，或冗余的答案会多次重复相同信息。”，

”diversity”: ”How varied and rich is the answer in providing different perspectives and insights on the question? A diverse answer should be multi-faceted and multi-dimensional, offering different viewpoints and angles on the question. For example, if the question is ’What are the causes and effects of climate change?’, a diverse answer would provide different causes and effects of climate change, such as greenhouse gas emissions, deforestation, natural disasters, biodiversity loss, etc. A diverse answer should also provide different sources and evidence to support the answer. For example, a single-source answer would only cite one source or evidence, or a biased answer would only provide one perspective or opinion.”,

”diversity”: ”答案在提供不同视角和见解方面有多丰富多样? 多样化的答案应当是多面且多维的, 从不同观点和角度阐述问题。例如, 若问题是 “气候变化的成因和影响有哪些?”, 多样的答案会提供气候变化的不同成因和影响, 如温室气体排放、森林砍伐、自然灾害、生物多样性丧失等。多样化的答案还应提供不同来源和证据来支持结论。例如, 单一来源的答案只引用一个来源或证据, 或有偏见的答案只提供单一观点或意见。”

”directness”: ”How specifically and clearly does the answer address the question? A direct answer should provide a clear and concise answer to the question. For example, if the question is ’What is the capital of France?’, a direct answer would be ’Paris’. A direct answer should not provide any irrelevant or unnecessary information that does not answer the question. For example, an indirect answer would be ’The capital of France is located on the river Seine.’”

”directness”: ”答案以多大程度上具体且清晰地回应问题? 直接的答案应当给出清晰简洁的答复。例如, 若问题是 “法国的首都是什么?”, 直接的答案就是 “巴黎”。直接的答案不应包含与回答无关或不必要的信息。例如, 间接的答案会说 “法国的首都位于塞纳河畔”。”

”empowerment”: ”How well does the answer help the reader understand and make informed judgements about the topic without being misled or making fallacious assumptions. Evaluate each answer on the quality of answer as it relates to clearly explaining and providing reasoning and sources behind the claims in the answer.”

”empowerment”: ”答案在多大程度上帮助读者理解并就该主题做出明智判断, 同时不被误导或做出谬误假设。根据答案在清晰解释、提供推理和列出论据来源方面的质量来评估每个回答。”

}

## G Statistical Analysis

### G 统计分析

Table 6: Pairwise comparisons of six conditions on four metrics across 125 questions and two datasets. For each question and metric, the winning condition received a score of 100, the losing condition received a score of 0, and in the event of a tie, each condition was scored 50 . These scores were then averaged over five evaluation runs for each condition. Results of Shapiro-Wilk tests indicated that the data did not follow a normal distribution. Thus, non-parametric tests (Wilcoxon signed-rank tests) were employed to assess the performance differences between pairs of conditions, with Holm-Bonferroni correction applied to account for multiple pairwise comparisons. The corrected p-values that indicated statistically significant differences are highlighted in bold.

表 6: 在两个数据集的 125 个问题上, 对六种条件在四个指标上的两两比较。对于每个问题和指标, 获胜条件得分为 100, 失败条件得分为 0, 若出现平局, 各条件各得 50 分。这些得分随后在每个条件下的五次评估运行中取平均。Shapiro-Wilk 检验结果表明数据不服从正态分布。因此, 采用非参数检验 (Wilcoxon 符号秩检验) 来评估条件对之间的性能差异, 并应用 Holm-Bonferroni 校正以调整多次两两比较。经校正后显示具有统计显著差异的 p 值以粗体标出。

				Podcast Transcripts		News Articles				
	Condition 1	Condition 2	Mean 1	Mean 2	Z-value	p-value	Mean 1	Mean 2	Z-value	p-value
Comprehensiveness	C0	TS	50.24	49.76	-0.06	1	55.52	44.48	-2.03	0.17
	C1	TS	51.92	48.08	-1.56	0.633	58.8	41.2	-3.62	0.002
	C2	TS	57.28	42.72	-4.1	<0.001	62.08	37.92	-5.07	<0.001
	C3	TS	56.48	43.52	-3.42	0.006	63.6	36.4	-5.63	<0.001
	C0	SS	71.92	28.08	-6.2	<0.001	71.76	28.24	-6.3	<0.001
	C1	SS	75.44	24.56	-7.45	<0.001	74.72	25.28	-7.78	<0.001
	C2	SS	77.76	22.24	-8.17	<0.001	79.2	20.8	-8.34	<0.001
	C3	SS	78.96	21.04	-8.12	<0.001	79.44	20.56	-8.44	<0.001
	TS	SS	83.12	16.88	-8.85	<0.001	79.6	20.4	-8.27	<0.001
	C0	C1	53.2	46.8	-1.96	0.389	51.92	48.08	-0.45	0.777
	C0	C2	50.24	49.76	-0.23	1	53.68	46.32	-1.54	0.371
	C1	C2	51.52	48.48	-1.62	0.633	57.76	42.24	-4.01	<0.001
	C0	C3	49.12	50.88	-0.56	1	52.16	47.84	-0.86	0.777
	C1	C3	50.32	49.68	-0.66	1	55.12	44.88	-2.94	0.016
	C2	C3	52.24	47.76	-1.97	0.389	58.64	41.36	-3.68	0.002
Diversity	C0	TS	50.24	49.76	-0.11	1	46.88	53.12	-1.38	0.676
	C1	TS	50.48	49.52	-0.12	1	54.64	45.36	-1.88	0.298
	C2	TS	57.12	42.88	-2.84	0.036	55.76	44.24	-2.16	0.184
	C3	TS	54.32	45.68	-2.39	0.1	60.16	39.84	-4.07	<0.001
	C0	SS	76.56	23.44	-7.12	<0.001	62.08	37.92	-3.57	0.003
	C1	SS	75.44	24.56	-7.33	<0.001	64.96	35.04	-4.92	<0.001
	C2	SS	80.56	19.44	-8.21	<0.001	70.56	29.44	-6.29	<0.001
	C3	SS	80.8	19.2	-8.3	<0.001	69.12	30.88	-5.53	<0.001
	TS	SS	82.08	17.92	-8.43	<0.001	67.2	32.8	-4.85	<0.001
	C0	C1	49.76	50.24	-0.13	1	39.68	60.32	-3.61	0.003
	C0	C2	46.32	53.68	-1.5	0.669	40.96	59.04	-3.14	0.012
	C1	C2	44.08	55.92	-3.27	0.011	50.24	49.76	-0.22	1
	C0	C3	44	56	-2.6	0.065	41.04	58.96	-3.47	0.004
	C1	C3	45.44	54.56	-2.98	0.026	49.52	50.48	-0.01	1
	C2	C3	48.48	51.52	-0.96	1	50.96	49.04	-0.39	1
Empowerment	C0	TS	40.96	59.04	-4.3	<0.001	42.24	57.76	-3.32	0.012
	C1	TS	45.2	54.8	-3.76	0.002	50	50	-0.12	1
	C2	TS	47.68	52.32	-2.2	0.281	49.52	50.48	-0.22	1
	C3	TS	48.72	51.28	-1.27	1	51.68	48.32	-1.2	1
	C0	SS	42.96	57.04	-3.71	0.003	42.72	57.28	-3.12	0.022
	C1	SS	47.68	52.32	-1.5	0.936	51.36	48.64	-0.84	1
	C2	SS	50.72	49.28	-0.55	1	49.84	50.16	-0.2	1
Empowerment	C3	SS	48.96	51.04	-0.57	1	49.52	50.48	-0.08	1
	TS	SS	57.52	42.48	-4.1	<0.001	52.88	47.12	-1.1	1
	C0	C1	48.72	51.28	-1.23	1	42.4	57.6	-3.9	0.001
	C0	C2	46.64	53.36	-2.54	0.12	44.8	55.2	-2.16	0.336
	C1	C2	49.28	50.72	-1.73	0.682	52	48	-1.45	1
	C0	C3	47.6	52.4	-1.78	0.682	44.32	55.68	-3.45	0.008
	C1	C3	50	50	0	1	51.44	48.56	-1.02	1
	C2	C3	50.72	49.28	-0.86	1	50.4	49.6	-0.22	1
Directness	C0	TS	44.96	55.04	-4.09	<0.001	45.2	54.8	-3.68	0.003
	C1	TS	47.92	52.08	-2.41	0.126	46.64	53.36	-2.91	0.04
	C2	TS	48.8	51.2	-2.23	0.179	48.32	51.68	-2.12	0.179
	C3	TS	48.08	51.92	-2.23	0.179	48.32	51.68	-2.56	0.074
	C0	SS	35.12	64.88	-6.17	<0.001	41.44	58.56	-4.82	<0.001
	C1	SS	40.32	59.68	-4.83	<0.001	45.2	54.8	-3.19	0.017
	C2	SS	40.4	59.6	-4.67	<0.001	44.88	55.12	-3.65	0.003
	C3	SS	40.48	59.52	-4.69	<0.001	45.6	54.4	-2.86	0.043
	TS	SS	43.6	56.4	-3.96	<0.001	46	54	-2.68	0.066
	C0	C1	46.96	53.04	-2.87	0.037	47.6	52.4	-2.17	0.179
	C0	C2	48.4	51.6	-2.06	0.197	48.48	51.52	-1.61	0.321
	C1	C2	49.84	50.16	-1	0.952	49.28	50.72	-1.6	0.321
	C0	C3	48.4	51.6	-1.8	0.29	47.2	52.8	-2.62	0.071
	C1	C3	49.76	50.24	0	1	48.8	51.2	-1.29	0.321
	C2	C3	50	50	0	1	48.8	51.2	-1.84	0.262

				播客文字记录		新闻文章				
	条件 1	条件 2	均值 1	均值 2	Z 值	p 值	均值 1	均值 2	Z 值	p 值
全面性	C0	TS	50.24	49.76	-0.06	1	55.52	44.48	-2.03	0.17
	C1	TS	51.92	48.08	-1.56	0.633	58.8	41.2	-3.62	0.002
	C2	TS	57.28	42.72	-4.1	<0.001	62.08	37.92	-5.07	<0.001
	C3	TS	56.48	43.52	-3.42	0.006	63.6	36.4	-5.63	<0.001
	C0	SS	71.92	28.08	-6.2	<0.001	71.76	28.24	-6.3	<0.001
	C1	SS	75.44	24.56	-7.45	<0.001	74.72	25.28	-7.78	<0.001
	C2	SS	77.76	22.24	-8.17	<0.001	79.2	20.8	-8.34	<0.001
	C3	SS	78.96	21.04	-8.12	<0.001	79.44	20.56	-8.44	<0.001
	TS	SS	83.12	16.88	-8.85	<0.001	79.6	20.4	-8.27	<0.001
	C0	C1	53.2	46.8	-1.96	0.389	51.92	48.08	-0.45	0.777
	C0	C2	50.24	49.76	-0.23	1	53.68	46.32	-1.54	0.371
	C1	C2	51.52	48.48	-1.62	0.633	57.76	42.24	-4.01	<0.001
	C0	C3	49.12	50.88	-0.56	1	52.16	47.84	-0.86	0.777
	C1	C3	50.32	49.68	-0.66	1	55.12	44.88	-2.94	0.016
	C2	C3	52.24	47.76	-1.97	0.389	58.64	41.36	-3.68	0.002
多样性	C0	TS	50.24	49.76	-0.11	1	46.88	53.12	-1.38	0.676
	C1	TS	50.48	49.52	-0.12	1	54.64	45.36	-1.88	0.298
	C2	TS	57.12	42.88	-2.84	0.036	55.76	44.24	-2.16	0.184
	C3	TS	54.32	45.68	-2.39	0.1	60.16	39.84	-4.07	<0.001
	C0	SS	76.56	23.44	-7.12	<0.001	62.08	37.92	-3.57	0.003
	C1	SS	75.44	24.56	-7.33	<0.001	64.96	35.04	-4.92	<0.001
	C2	SS	80.56	19.44	-8.21	<0.001	70.56	29.44	-6.29	<0.001
	C3	SS	80.8	19.2	-8.3	<0.001	69.12	30.88	-5.53	<0.001
	TS	SS	82.08	17.92	-8.43	<0.001	67.2	32.8	-4.85	<0.001
	C0	C1	49.76	50.24	-0.13	1	39.68	60.32	-3.61	0.003
	C0	C2	46.32	53.68	-1.5	0.669	40.96	59.04	-3.14	0.012
	C1	C2	44.08	55.92	-3.27	0.011	50.24	49.76	-0.22	1
	C0	C3	44	56	-2.6	0.065	41.04	58.96	-3.47	0.004
	C1	C3	45.44	54.56	-2.98	0.026	49.52	50.48	-0.01	1
	C2	C3	48.48	51.52	-0.96	1	50.96	49.04	-0.39	1
	C0	TS	40.96	59.04	-4.3	<0.001	42.24	57.76	-3.32	0.012
	C1	TS	45.2	54.8	-3.76	0.002	50	50	-0.12	1
	C2	TS	47.68	52.32	-2.2	0.281	49.52	50.48	-0.22	1
	C3	TS	48.72	51.28	-1.27	1	51.68	48.32	-1.2	1
	C0	SS	42.96	57.04	-3.71	0.003	42.72	57.28	-3.12	0.022
	C1	SS	47.68	52.32	-1.5	0.936	51.36	48.64	-0.84	1
	C2	SS	50.72	49.28	-0.55	1	49.84	50.16	-0.2	1
赋权	C3	SS	48.96	51.04	-0.57	1	49.52	50.48	-0.08	1
	TS	SS	57.52	42.48	-4.1	<0.001	52.88	47.12	-1.1	1
	C0	C1	48.72	51.28	-1.23	1	42.4	57.6	-3.9	0.001
	C0	C2	46.64	53.36	-2.54	0.12	44.8	55.2	-2.16	0.336
	C1	C2	49.28	50.72	-1.73	0.682	52	48	-1.45	1
	C0	C3	47.6	52.4	-1.78	0.682	44.32	55.68	-3.45	0.008
	C1	C3	50	50	0	1	51.44	48.56	-1.02	1
	C2	C3	50.72	49.28	-0.86	1	50.4	49.6	-0.22	1
	C0	TS	44.96	55.04	-4.09	<0.001	45.2	54.8	-3.68	0.003
	C1	TS	47.92	52.08	-2.41	0.126	46.64	53.36	-2.91	0.04
	C2	TS	48.8	51.2	-2.23	0.179	48.32	51.68	-2.12	0.179
	C3	TS	48.08	51.92	-2.23	0.179	48.32	51.68	-2.56	0.074
	C0	SS	35.12	64.88	-6.17	<0.001	41.44	58.56	-4.82	<0.001
	C1	SS	40.32	59.68	-4.83	<0.001	45.2	54.8	-3.19	0.017
	C2	SS	40.4	59.6	-4.67	<0.001	44.88	55.12	-3.65	0.003