MultiHop-RAG： 对多跳查询的检索增强生成进行基准测试

Yixuan Tang and **Yi Yang**
Hong Kong University of Science and Technology
{ yixuantang,imyiyang } @ust.hk

Abstract

检索增强生成（RAG）通过检索相关知识来增强大型语言模型（LLM），在减轻 LLM 幻觉和提高响应质量方面显示出巨大的潜力，从而促进 LLM 在实践中的广泛采用。然而，我们发现现有的 RAG 系统不足以回答多跳查询，这需要对多个支持证据进行检索和推理。此外，据我们所知，现有的 RAG 基准测试数据集都没有专注于多跳查询。在本文中，我们开发了一个新的数据集 MultiHop-RAG ，它由知识库、大量多跳查询、它们的真值答案以及相关的支持证据组成。我们详细介绍了构建数据集的过程，利用英文新闻文章数据集作为底层 RAG 知识库。我们在两个实验中证明了 MultiHop-RAG 的基准测试效用。第一个实验比较了用于检索多跃点查询证据的不同嵌入模型。在第二个实验中，我们研究了各种最先进的 LLM，包括 GPT-4、PaLM 和 Llama2-70B，在给定证据的情况下推理和回答多跳查询的能力。这两个实验都表明，现有的 RAG 方法在检索和应答多跳查询方面的表现并不理想。我们希望 MultiHop-RAG 能成为社区开发有效 RAG 系统的宝贵资源，从而促进 LLM 在实践中的更多采用。MultiHop-RAG 和实施的 RAG 系统可在 https://github.com/yixuantt/MultiHop-RAG/ 上公开获得。

1 介绍

ChatGPT 等大型语言模型（LLM）的出现促进了广泛的创新，为智能聊天机器人和其他自然语言处理（NLP）应用程序提供了动力 (**?**) 。一个很有前途的用例是检索增强生成（RAG）(**?**) ，它通过在生成响应之前引用 LLM 训练数据源之外的外部知识库来优化大型语言模型的输出。RAG 提高了 LLM 的反应 (**?**) ，也减轻了幻觉的发生，从而提高了模型的可信度 (**?**) 。基于 LLM 的框架，例如 LlamaIndex (**?**) 和 LangChain (**?**) ，专门支持 RAG 管道。

在实际的检索增强生成（RAG）应用程序中，用户的查询通常需要从多个文档中检索和推理证据，这一过程称为多跳查询。例如，考虑使用财务报告数据库进行财务分析。金融分析师

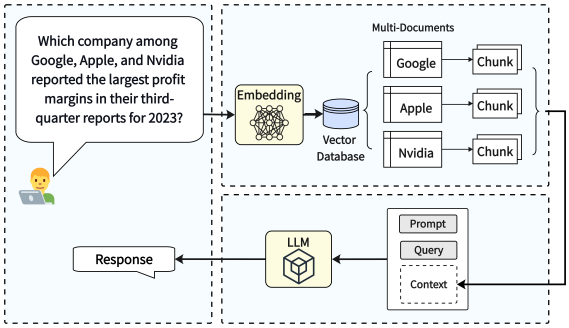

Figure 1: 具有多跳查询的 RAG。

可能会问，谷歌、苹果和英伟达中哪家公司在 2023 年第三季度报告中报告了最大的利润率？或询问特定公司随时间推移的表现，例如苹果过去三年的销售趋势如何？这些查询需要来自多个文档的证据来制定答案。由于此类查询的多面性，涉及来自各种来源的信息，传统的相似性匹配方法（如查询和财务报表块嵌入之间的余弦相似性）可能无法产生最佳结果。我们在图 **??** 中演示了这种多跳检索过程。

然而，现有的 RAG 基准测试，例如 RGB (**?**) 和 RECALL (**?**) ，主要评估一个简单的案例，其中可以使用单个证据检索和解决查询的答案。这些基准测试都没有评估 LLM 对复杂多跳查询的检索和推理能力。为了弥补这一差距并使 RAG 基准测试更接近真实世界的场景，在本文中，我们介绍了 MultiHop-RAG. 据我们所知，MultiHop-RAG 是首批专门针对多跳查询的 RAG 数据集之一。

基于实际场景中常见的 RAG 查询，我们首先将多跳查询分为四种类型：推理查询、比较查询、时态查询和 Null 查询。前三种类型（推理、比较和时间）需要从多个来源检索和分析证据，包括推断关系、比较数据点和随时间推移对事件进行排序等任务。Null 查询表示无法从知识库派生查询的方案。此类别对于评估 LLM 在检索到的文本缺乏相关性时是否可能产生多跳查询答案的幻觉至关重要。

我们使用一系列新闻文章构建我们的 RAG 知识库。使用 GPT-4 作为数据生成器，然后我

News source	Fortune Magazine	The Sydney Morning Herald
Evidence	Back then, just like today, home prices had boomed for years before Fed officials were ultimately forced to hike interest rates aggressively in an attempt to fight inflation.	Postponements of such reports could complicate things for the Fed, which has insisted it will make upcoming decisions on interest rates based on what incoming data say about the economy.
Claim	Federal Reserve officials were forced to aggressively hike interest rates to combat inflation after years of booming home prices.	The Federal Reserve has insisted that it will base its upcoming decisions on interest rates on the incoming economic data.
Bridge-Topic	Interest rate hikes to combat inflation	Interest rate decisions based on economic data
Bridge-Entity	Federal Reserve	Federal Reserve
Query	Does the article from Fortune suggest that the Federal Reserve's interest rate hikes are a response to past conditions, such as booming home prices, while The Sydney Morning Herald article indicates that the Federal Reserve's future interest rate decisions will be based on incoming economic data?	
Answer	Yes	

Table 1: 多跃点查询的示例，包括来自两篇新闻文章的支持证据、释义声明、bridge-topic 和 bridge-entity，以及相应的答案。

们采取广泛的过程来构建一组不同的多跳查询，每个查询都需要对多个文档进行检索和推理。查询构造的示例如表 **??** 所示。首先，我们首先从每篇新闻文章中提取事实句子作为证据。例如，从一篇文章中提取的证据可能会说："当时，就像今天一样，在美联储官员最终被迫大幅加息以对抗通胀之前，房价已经上涨了多年。其次，我们将每个证据片段输入到 GPT-4 中，促使它将证据重新表述为主张。这一说法通过一个消除歧义的主题和实体进行了澄清。例如，GPT-4 可能会将上述证据改写为："在房价多年飙升后，美联储官员被迫积极加息以对抗通胀"，将 "加息对抗通胀" 确定为主题，将 "美联储" 作为实体。这些主题和实体充当用于构造多跃点查询的桥梁，称为桥接主题或桥接实体。接下来，我们使用 GPT-4 生成与同一桥主题或桥实体相关的特定多跳查询，并附有正确答案。最后，我们进行验证步骤以确保数据质量。

我们使用两个实验来演示 MultiHop-RAG 的基准测试能力，利用 LlamaIndex (**?**) 实现的 RAG 系统。第一个实验涉及比较不同的嵌入模型，以检索多跳查询的相关证据。在第二个实验中，我们评估了各种最先进的 LLM 的推理和回答能力，包括 GPT-4、GPT-3.5、PaLM、Claude-2、Llama2-70B 和 Mixtral-8x7B，用于提供检索到的文本时进行多跳查询。两个实验的结果表明，当前的 RAG 实现不足以有效地检索和应答多跳查询。我们公开发布这个具有挑战性的 MultiHop-RAG 数据集，并希望它能成为社区开发和基准测试 RAG 系统的宝贵资源，从而释放生成式 AI 在实践中的巨大潜力。

2 具有多跳查询的 RAG

2.1 检索增强生成（RAG）

在 RAG 应用程序中，我们使用一个外部语料库，表示为 $\mathcal{D}$ ，它由多个文档组成并用作知识库。此语料库中的每个文档（表示为 $d_i \in \mathcal{D}$ ）都被分割成一组块。然后，使用嵌入模型将这些块转换为向量表示，并存储在嵌入数据库中。给定用户查询 q ，系统通常会检索与查询最匹配的前 K 个块。这些块构成了查询 q 的检索集，表示为 $\mathcal{R}_q = \{r_1, r_2, ..., r_K\}$ 。然后，将检索到的块与查询和可选提示相结合，输入到 LLM 中以生成最终答案，格式如下：$\texttt{LLM}(q, \mathcal{R}_q, \texttt{prompt}) \rightarrow \texttt{answer}$.

2.2 多跳查询

我们将多跳查询定义为需要检索和推理多个支持证据以提供答案的查询。换句话说，对于多跳查询 q ，检索集中的块 $\mathcal{R}_q$ 共同提供对 q 的答案。例如，查询 "谷歌、苹果和英伟达中哪家公司在其 2023 年第三季度报告中报告了最大的利润率?" 需要 1）从这三家公司的报告中检索与利润率相关的相关证据;2）通过对检索到的多个证据进行比较和推理来产生答案。这与单跳查询不同，例如 "谷歌在 2023 年第三季度报告中的利润率是多少"，在单跳查询中，答案可以直接从单个证据中得出。

根据实际 RAG 系统中常用的查询，我们确定了四种类型的多跳查询。对于每种类型，我们在财务 RAG 系统的上下文中提出一个假设查询，其中知识库由年度报告的集合组成。

推理查询：对于这样的查询 q ，从检索集 $\mathcal{R}_q$ 通过推理推导出答案。推理查询的一个示例可能是：哪个报告讨论了 Apple 的供应链风险，是 2019 年年度报告还是 2020 年年度报告?

比较查询：对于这样的查询 q ，答案需要对检

索集内的证据进行比较 $\mathcal{R}_q$ 。例如，比较查询可能会问：Netflix 或 Google 是否报告了 2023 年的收入增加？"
时态查询：对于这样的查询 q ，答案需要分析检索到的块的时态信息。例如，临时查询可能会问：Apple 是在第 5 代 iPad Pro 推出之前还是之后推出了 AirTag 跟踪设备?
空查询：对于如查询 q ，无法从检索到的集合 $\mathcal{R}_q$ 中派生答案。我们包括 null 查询来评估生成质量，尤其是关于幻觉问题。对于 null 查询，即使提供了检索到的集合，LLM 也应生成 null 响应，而不是产生幻觉答案。例如，假设 ABCD 是一家不存在的公司，则 null 查询可能会询问：ABCD 公司 2022 年和 2023 年年度报告中报告的销售额是多少?

2.3 评估指标

处理多跳查询的 RAG 系统可以从检索评估和生成评估两个关键方面进行评估。

检索评估：显然，检索集的质量 $\mathcal{R}_q$ 决定了最终的生成质量。我们将检索到的集合与与每个查询关联的真实证据进行比较，但空查询除外，因为它们没有证据可派生。假设检索到前 K 个块，即 $|\mathcal{R}_q| = K$ ，我们使用检索评估指标，包括 K 处的平均精度（MAP@K）、K 处的平均倒数秩（MRR@K）和 K 处的命中率（Hit@K）。MAP@K 测量所有查询的平均 top-K 检索精度。MRR@K 计算每个查询的第一个相关块的倒数排名的平均值，同时考虑检索到的前 K 个集合。Hit@K 指标衡量在检索到的前 K 集中出现的证据比例。

响应评估：由于多跳查询需要对检索到的多个块进行推理，因此我们还可以通过将 LLM 响应与查询的真值答案进行比较来评估 LLM 的推理能力。

3 基准数据集：MultiHop-RAG

在本节中，我们将提供有关 MultiHop-RAG 数据集构建的详细信息。具体来说，我们描述了创建一组多跳查询的过程，以及相应的真实证据集和从一系列新闻文章中派生的答案。

3.1 MultiHop-RAG 施工

第 1 步：数据集采集。我们使用 mediastack API [1] 下载新闻数据集，这是一个提供全球新闻数据的 REST API 接口。新闻数据源包括各种英语网站，涵盖一系列新闻类别：娱乐、商业、体育、技术、健康和科学。为了模拟真实世界的 RAG 场景，其中知识库数据（例如企业的内部数据）可能与 LLM 的训练数据不同，我们选择了 2023 年 9 月 26 日至 2023 年 12 月 26 日发布的新闻文章。截至撰写本文时，这个时间框架超出了一些广泛使用的 LLM 的知识截止时间，包括 ChatGPT 和 LLaMA。这种选择还有助于梳理出潜在的 LLM 接触过这些新闻文章的可能性。我们只保留令牌长度大于或等于 1,024 的文章。每篇新闻文章都与元数据配对，包括标题、发布日期、作者、类别、URL 和新闻来源。

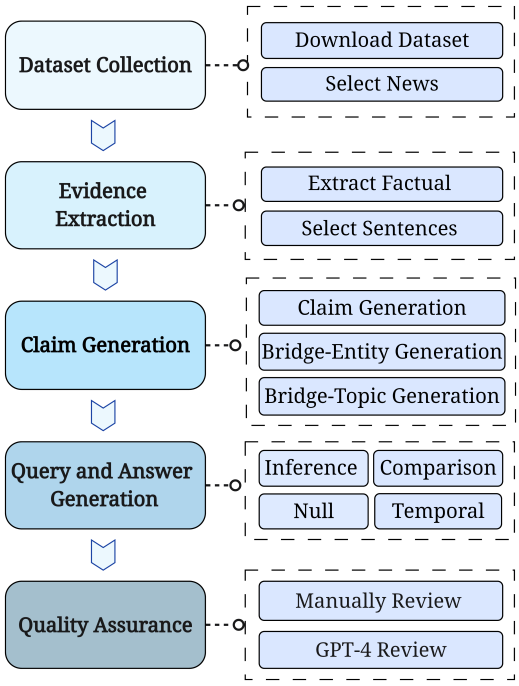

Figure 2: MultiHop-RAG 施工管道。

第 2 步：证据提取。对于每篇文章，我们使用经过训练的语言模型提取事实或观点句子 [2] 。这些事实句稍后被用作回答多跃点查询的证据。我们只保留那些包含可能与其他新闻文章有重叠关键词的证据的新闻文章。这允许我们稍后创建多跃点查询，其中答案的证据来自多个来源。

第 3 步：声明、桥接实体、桥接主题生成。我们的目标是使用 GPT-4 使用证据集自动生成高质量的多跳查询。但是，由于语言结构不一致，从步骤 2 获得的原始证据对于查询生成并不理想。例如，一些证据使用代词来指代主体，而文本中缺乏实际实体。为了解决这个问题，我们使用 GPT-4 来解释证据，我们称之为主张，给定原始证据及其上下文。为确保生成的声明与证据之间的一致性，我们进一步使用 UniEval (**?**) 框架进行事实核查，以验证证据与声明之间的一致性。附录 **??** 介绍了用于 GPT-4

[1] https://mediastack.com/

[2] https://huggingface.co/lighteternal/fact-or-opinion-xlmr-el

生成声明的提示。

Bridge-Entity 和 Bridge-Topic：跨证据的共享实体或主题称为 bridge-entity 或 bridge-topic。这些桥接实体或桥接主题可用于链接从中派生多跳查询答案的不同证据。例如，在诸如“谷歌报告其 2023 年第三季度业绩，展示其财务业绩的详细概述，包括收入增长、利润率”之类的声明中，利润率一词可以被视为一个桥梁话题，而谷歌一词可以被视为连接不同证据的桥梁实体。我们提示 GPT-4 识别每个声明的桥接实体和桥接主题。附录 **??** 还介绍了用于 GPT-4 生成桥接的提示。

第 4 步：查询和答案生成。在此步骤中，我们利用 bridge-entity 或 bridge-topic 来生成多跳查询。具体来说，我们首先将具有相同桥接实体或桥接主题的声明分组到声明集中。我们将声明集限制为至少包含两个声明，但不超过四个声明。对于每种类型的查询，我们将声明集提供给 GPT-4，并提示它使用指令生成包含每个声明信息的查询。下面，我们将介绍不同多跳查询类型的规范。在每个查询的构建中，我们还包括与支持证据相关的新闻文章的来源，以模拟真实世界的 RAG 场景。附录 **??** 介绍了用于 GPT-4 生成查询的提示。

推理查询：这些查询是通过综合跨多个声明的桥接实体的各种特征来制定的，最终答案是实体本身的标识。

比较查询：这些查询的结构用于比较与桥接实体或主题相关的相似性和差异性。根据比较，此类查询的结果通常是明确的“是”或“否”。

时态查询：这些查询探索事件在不同时间点的时间顺序。此类查询的答案通常是“是”或“否”，或者是“之前”或“之后”等单个时间指示词。

Null 查询：Null 查询是其答案无法从检索到的集合中派生的查询。为了创建空查询，我们使用现有桥接实体中不存在的实体生成多跳查询。为了增加复杂性，我们在制定这些问题时还包括虚构的新闻源元数据，确保问题不引用知识库中任何上下文相关的内容。空查询的答案应为“信息不足”或类似。

第 5 步：质量保证。最后，我们使用两种方法来保证数据集的质量。首先，我们手动查看生成的多跃点查询的子集样本、其相应的证据集和最终答案。人工检查的结果表明具有高度的准确性和数据质量。其次，我们利用 GPT-4 根据以下标准评估数据集中的每个示例：1）生成的查询必须利用所有提供的证据来制定响应;2）查询应仅根据提供的证据进行回答;3）对生成的查询的响应应为单个单词或特定实体;4）查询必须符合其指定的查询类型。

Category	Avg. Tokens	Entry Count
technology	2262.3	172
entertainment	2084.3	114
sports	2030.6	211
science	1745.5	21
business	1723.8	81
health	1481.1	10
total	2046.5	609

Table 2: MultiHop-RAG 中新闻文章知识库的描述性统计。

Query Category	Entry Count	Percentage
Inference Query	816	31.92 %
Comparison Query	856	33.49 %
Temporal Query	583	22.81 %
Null Query	301	11.78 %
Total	2,556	100.00 %

Table 3: 查询类型在 MultiHop-RAG 中的分布。

3.2 描述统计学

MultiHop-RAG 数据集包含六种不同类型的新闻文章，涵盖 609 条不同的新闻，平均有 2,046 个标记。新闻类别的分布如表 **??** 所示。MultiHop-RAG 包含四种类型的多跃点查询，这些查询的分布如表 **??** 所示。数据集中总共有大约 88 个% 查询是非空查询，可以从知识库中检索和推理答案。此外，查询的形式表现出相当大的多样性。大约 27 个% 疑问以“does”开头，大约 15 个% 以“what”开头，类似比例的% 以“which”开头，14 个% 以“who”开头，其余的包含一小部分其他疑问词，例如“when”。此外，回答多跳查询所需的证据数量各不相同。表 **??** 显示了数据集中每个查询的证据编号的分布情况。大约 42 个% 的查询可以使用两个证据来回答，而大约 30 个% 和 15 个% 的查询可以分别使用三个或四个证据来回答。

Num. of Evidence Needed	Count	Percentage
0 (Null Query)	301	11.78 %
2	1078	42.18 %
3	779	30.48 %
4	398	15.56 %
Total	2,556	100.00 %

Table 4: 在 MultiHop-RAG 中回答多跳查询所需的证据数的分布。

4 使用 MultiHop-RAG 对 RAG 系统进行基准测试

MultiHop-RAG 可用作各种 RAG 相关任务的基准。从广义上讲，RAG 相关任务可分为检索相关任务和生成相关任务。与检索相关的任务侧重于从知识库中检索相关文本，而与生成相关的任务侧重于根据检索到的文本生成高质量的响应。在本节中，我们将展示每个任务的两个用例，其中可以使用 MultiHop-RAG 。

4.1 检索相关任务

RAG 系统中一个重要的设计选择是嵌入模型的选择。嵌入模型将数据转换为数值向量，然后将这些向量存储在嵌入数据库中。在本实验中，我们通过检查其检索质量来评估不同的嵌入模型。

实验设置：我们使用 LlamaIndex 框架实现 RAG 系统 (**?**) 。我们将 MultiHop-RAG 知识库中的文档划分为多个块，每个块由 256 个标记组成。然后，我们使用嵌入模型转换块，并将嵌入保存到向量数据库中。同样，在检索步骤中，我们使用相同的嵌入模型转换查询，并检索与查询嵌入具有最高余弦相似度的前 K 个最相关的块。在这个实验中，我们测试了各种嵌入模型，包括 OpenAI 的 ada-embeddings（text-embedding-ada-002，text-search-ada-query-001）、voyage-02 [3] 、llm-embedder (**?**) 、bge-large-en-v1.5 (**?**) 、jina-embeddings-v2-base-en (**?**) 、e5-base-v2 (**?**) 和 instructor-large (**?**) 。在此试验中排除了 NULL 查询，因为没有与查询匹配的证据。此外，我们还包含一个 Reranker 模块，用于使用 bge-reranker-large (**?**) 检查检索性能。在使用嵌入模型检索 20 个相关块后，我们使用 Reranker 进一步选择前 K 个块。

实验结果：表 **??** 显示了使用不同嵌入模型的检索结果。这表明在检索多跳查询的相关证据方面仍然存在很大差距。虽然 Rerank 可以有效提高检索相关性，但在使用 Reranker 技术时，最高 Hits@10 仅为 0.7467。此外，最高 Hits@4 跌至 0.6625 令人担忧。在实际的 RAG 系统中，底层 LLM 通常具有上下文窗口限制。因此，检索到的块的数量通常限制为少量。检索指标的低值凸显了在多跳查询和文本块之间使用直接相似性匹配时检索多跃点查询的相关证据的挑战。

4.2 与生成相关的任务

底层 LLM 在 RAG 系统中生成响应中起着至关重要的作用。在这个实验中，我们评估了在两种不同设置下生成的响应的质量。在第一种设置中，我们采用性能最佳的检索模型，即带有 bge-reranker-large 的 voyage-02，如表 **??** 所示，检索 top-K 文本，然后将它们输入到 LLM 中。在第二种设置中，我们使用与每个查询相关的真实证据作为 LLM 的检索文本。此设置表示测试 LLM 响应能力的上限性能，因为它利用了实际证据。

实验设置：在第一个实验中，我们检索前 6 个块，以便检索到的文本的总长度不超过 2,048。MultiHop-RAG 中的所有查询都在实验中进行测试。在第二个实验中，由于 null 查询没有关联的证据，因此我们在实验中排除了这种类型的查询。对于实验中使用的 LLM，我们考虑了最先进的商业模型，包括 GPT-4 (**?**) 、GPT-3.5、Claude-2 (**?**) 和 Google-PaLM (**?**) 。我们使用相应模型提供的 API 获得答案。我们还评估了一些开源模型，包括 Mixtral-8x7b-instruct (**?**) 和 Llama-2-70b-chat-hf (**?**) 。

实验结果：表 **??** 显示了不同 LLM 的响应精度。首先，我们可以看到，使用检索到的块的响应准确率并不令人满意，最先进的 GPT-4 模型仅实现了 0.56 的准确率。这是意料之中的，因为检索组件在从知识库中检索相关证据方面存在不足。其次，即使我们向 LLM 提供真实证据，我们也可以看到响应的准确性远非完美。Llama02-70B 和 Mixtral-8x7B 等开源 LLM 分别只能达到 0.32 和 0.36 的精度。GPT-4 以 0.89 的准确率实现了强大的推理能力，其次是基于第二种的 LLM Google-PaLM，准确率为 0.74。

图 **??** 显示了 GPT-4 和 Mixtral-8x7B-instruct 不同查询类型的详细结果。这两种模型在空查询上都表现出相对较高的鲁棒性，这意味着它们通常擅长根据检索到的文本确定何时无法回答查询。这是令人鼓舞的，因为 RAG 的一个好处是通过用检索知识增强 LLM 来缓解 LLM 幻觉问题。然而，Mixtral-8x7B 模型在比较和时间查询方面的表现明显差于 GPT-4。在查看不正确的响应后，我们发现 Mixtral-8x7B 无法准确处理逻辑否定，导致对语句的误解，从而在比较查询中性能低下。此外，Mixtral-8x7B 通常无法正确识别事件的时间顺序，这对于回答时间是关键因素的时间查询至关重要。综上所述，该实验表明，LLM 的推理能力仍有改进的空间，尤其是那些开源的 LLM，用于多跳查询。

4.3 其他用例

除了嵌入模型和 LLM 生成之外，还有其他值得探索的领域。例如，查询分解是 RAG 框架（如 LLamaIndex）中广泛使用的技术。此过程

[3] https://www.voyageai.com/

Embedding	Without Reranker				With bge-reranker-large			
	MRR@10	MAP@10	Hits@10	Hits@4	MRR@10	MAP@10	Hits@10	Hits@4
text-embedding-ada-002	0.4203	0.3431	0.6381	0.504	0.5477	0.4625	0.7059	0.6169
text-search-ada-query-001	0.4203	0.3431	0.6399	0.5031	0.5483	0.4625	0.7064	0.6174
llm-embedder	0.2558	0.1725	0.4499	0.3189	0.425	0.3059	0.5478	0.4756
bge-large-en-v1.5	0.4298	0.3423	0.6718	0.5221	0.563	0.4759	0.7183	0.6364
jina-embeddings-v2-base-en	0.0621	0.031	0.1479	0.0802	0.1412	0.0772	0.1909	0.1639
intfloat/e5-base-v2	0.1843	0.1161	0.3556	0.2334	0.3237	0.2165	0.4176	0.3716
voyage-02	0.3934	0.3143	0.6506	0.4619	0.586	0.4795	0.7467	0.6625
hkunlp/instructor-large	0.3458	0.265	0.5717	0.4229	0.5115	0.4118	0.659	0.5775

Table 5: 不同嵌入模型的检索性能。

Models	Accuracy	
	Retrieved Chunk	Ground-truth Chunk
GPT-4	0.56	0.89
ChatGPT	0.44	0.57
Llama-2-70b-chat-hf	0.28	0.32
Mixtral-8x7B-Instruct	0.32	0.36
Claude-2.1	0.52	0.56
Google-PaLM	0.47	0.74

Table 6: LLM 的生成精度。

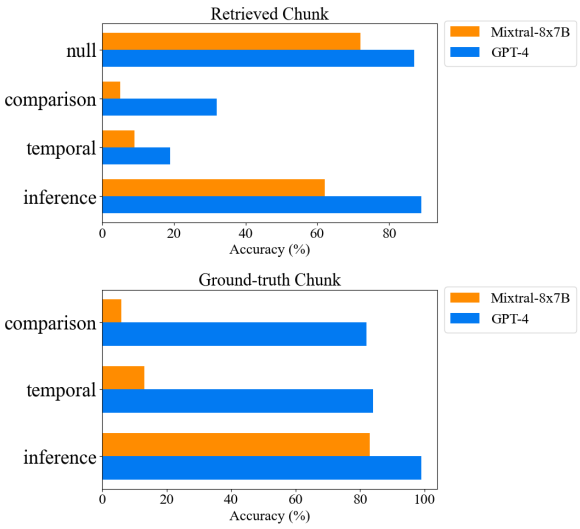

Figure 3: 不同查询类型的生成准确性。

涉及将查询分解为更小的段; 它以单个文档为检索目标，随后整合信息，从而有可能提高检索的准确性。另一种先进且有前途的方法是构建基于 LLM 的代理，这些代理可以自动计划和执行多跳查询，例如 AutoGPT (**?**) 。另一个感兴趣的领域是混合检索方法，它结合了关键字和嵌入匹配技术。我们相信，有许多潜在的领域可以提高 RAG 在多跳查询上的性能，而精心策划的数据集 MultiHop-RAG 可以成为社区的宝贵资源。

5 相关工作

RAG 评估：随着 RAG 系统的日益普及，各种 RAG 基准数据集和评估工具应运而生。例如，RGB (**?**) 和 RECALL (**?**) 评估 LLM 在涉及噪声、集成和反事实查询的条件下为 RAG 系统生成响应的性能。然而，这两个数据集主要侧重于评估 RAG 系统的生成方面，而没有专门解决其检索准确性。此外，自动化 RAG 评估工具（如 ARES (**?**) 和 RAGAS (**?**) . 这些工具利用 LLM 自动评估 RAG 生成的质量，但它们没有引入基准数据集。我们的工作引入了首批 RAG 基准数据集之一，该数据集由知识库、大量多跳查询、其真实答案以及相关的支持证据组成，从而补充了现有的 RAG 评估。

检索数据集：除了 RAG 的上下文之外，还存在几个用于信息检索评估的基准数据集。例如，FEVER（事实提取和验证）数据集包含被给定的维基百科文章分类为支持、驳斥或 NotEnoughInfo 的声明 (**?**) 。同样，SciFact 数据集包括科学声明和包含证据的摘要 (**?**) 。然而，这两个数据集中的声明都是单跳语句，支持证据来自一篇文章，与本文讨论的多跳查询形成鲜明对比。另一个数据集 HoVer 涉及需要从多篇维基百科文章中提取和推理的声明 (**?**) 。然而，与我们的数据集不同，HoVer 只专注于将声明分类为文章支持或不支持，而不评估 LLM 生成步骤。此外，在 HoVer 中，从中提取证据的维基百科文章被用于声明验证，这与我

们的设置有很大不同，在 HoVer 中，需要从大型知识库中提取相关证据。另外，(**?**) 评估了一系列用于信息检索的商业嵌入 API，但这种评估也没有在 RAG 系统的框架内进行上下文化。

多文档 QA 数据集：问答（QA）是 NLP 中的一项基本任务，一些流行的基准测试，如 HotpotQA (**?**) 、MultiRC (**?**) 和 2WikiMultiHopQA (**?**) ，旨在从多个文档源实现 QA。此任务类似于我们的多跃点查询 RAG 任务，因为两者都涉及来自多个信息源的推理。然而，这些数据集主要侧重于评估模型的推理技能，并不强调从知识库中检索证据。此外，它们的主要数据源维基百科与大多数现有 LLM 的训练数据显着重叠。如果我们使用这些资源对 RAG 系统进行基准测试，则存在一个潜在的担忧，即 LLM 响应可能依赖于训练知识，而不是从检索到的知识库中进行推理。

6 结论

在这项工作中，我们引入了 MultiHop-RAG ，这是一种新颖而独特的数据集，专为需要从多个支持证据中检索和推理的查询而设计。这些类型的多跃点查询表示实际方案中常见的用户查询。MultiHop-RAG 由知识库、大量多跃点查询、其真实答案以及相关的支持证据组成。本文详细介绍了 MultiHop-RAG 的创建过程，采用了一种将人类努力与 GPT-4 相结合的混合方法。此外，我们还探讨了 RAG 系统基准测试中 MultiHop-RAG 的两个用例，从而突出了该数据集的潜在应用。通过公开发布 MultiHop-RAG ，我们的目标是为社区提供宝贵的资源，为 RAG 系统的进步和基准测试做出贡献。

7

局限性 这项工作有几个局限性，可以在未来的研究中加以改进。首先，我们的真值答案仅限于简单的响应，如 “是”、“否”、实体名称或时间指标，如 “之前” 或 “之后”，以便于使用简单的准确性指标来评估生成性能。未来的工作可以考虑允许自由文本作为答案，并采用更复杂的指标来评估生成质量。其次，当前的数据集将查询的支持证据限制为最多四条。未来的工作可以通过包括需要从更多证据中检索和推理的查询来扩展数据集。最后，虽然我们的实验利用了使用 LlamaIndex 的基本 RAG 框架，但未来的工作可能涉及使用更高级的 RAG 框架或 LLM 代理框架评估多跳查询的答案。

A 附录 A：用于数据生成的 GPT-4 提示

我们介绍了用于指导 GPT-4 生成数据的提示。表 **??** 显示了用于生成声明的提示，以及这些声明中的相应主题和实体。表 **??** 、表 **??** 和表 **??** 分别显示了用于生成推理、比较和时态类型的多跳查询的提示。

B 附录 B：数据集示例

在本附录中，我们将提供 MultiHop-RAG 数据集中包含的每种类型的多跳查询的示例。这些示例在相应的表中进行了说明：用于推理查询的表 **??** 、用于比较查询的表 **??** 、用于临时查询的表 **??** 和用于 Null 查询的表 **??** 。每个查询都与一个真实答案配对，用于评估生成准确性，同时包含多个支持证据以评估检索性能。此外，还提供新闻文章的标题、来源和出版时间等元数据作为参考。

A "claim" is a statement or assertion made within a text expressing a belief, opinion, or fact. Given evidence from the original context, please extract one claim and its associated topics.

Note: The claim should not contain ambiguous references, such as 'he',' she,' and' it', and should use complete names. If there are multiple topics, give the most dominant one. The target of the claim (one entity)is the specific individual, group, or organization that the statement or assertion within a text is directed towards or about which it is making a case. The topic of the claim should be a simple phrase representing the claim's central argument concept. If there is no claim, please leave it blank. Please generate a claim based on the given evidence. Don't generate the evidence yourself.

Please give the response following this format:
Evidence: [original context]
Claims: [extract claim]
Claim Target: [target]
Claim Topic: [topic]

Here are examples:
<examples>
Now, it's your turn.
<News>
<evidence>

Table 7: 索赔生成提示

A multi-hop question is a query requiring multiple inferential leaps or accessing several pieces of information from different locations or sources to arrive at an answer. The following are news articles' metadata and claims come from the articles. All the claims from the article are related to a similar target. Your task is to generate one multi-hop inference question based on the claims. Here are some instructions:
1. Find the Connection: The connection between claims is <target> , which is how these key pieces of information are related or how they can be combined to form a more complex idea.
2. Formulate the Question: Create a question that cannot be answered by relying on just one of the sentences but instead requires understanding and linking the information from all of the sources. The answer is <target> .
3. Ensure Coherence: Make sure the question flows logically from the combined information and is clear and unambiguous.
4. Use the keywords: <key set>

<examples>
Context:
<Context>

Table 8: 推理查询生成提示

<Context>

The above are news articles' metadata and claims come from the articles. All the claims from the articles are related to a similar target. Your task is to generate one comparison question based on all the claims from different sources. This question needs to compare some factual elements of the claims that are explicitly stated to find where they agree or differ. The correct answer to this question is expressed as a comparative adjective, a statement of alignment, a simple yes or no. To generate a comparative question from claims, you need to use the following keywords: <key set>

The Good Comparison Questions:
<examples>
Your Comparison Question:

Table 9: 比较查询生成提示

<Context>

Please create a time-sensitive comparison question using metadata and excerpts from multiple news articles. That is to compare the consistency or sequence of reports on similar topics at multiple different time points. If it is to compare the consistency, please clearly mention the news source and time in the question using <time frame> . If it is to compare sequences of reports, just clearly mention the news source and do not mention the timeline. Utilize the following keywords provided in the <key set> to construct the question. The correct answer should based on the factual excerpts and is only one word.

<examples>
Your time-sensitive comparison question:

Table 10: 时态查询生成提示

A multi-hop question is a query requiring multiple inferential leaps or accessing several pieces of information from different locations or sources to arrive at an answer. Considering you have read at least two news articles on <entity> , construct a multi-hop question that incorporates all the news sources. The source of the news should be stated in the question. Also, ensure that the answer to the question is a single word/entity. Do not answer this question directly. Just give me the question:

Table 11: Null 查询生成提示

Query: Which platform is at the center of discussions in articles from Music Business Worldwide, Polygon, and FOX News - Health, concerning the policing of AI-driven voice replication, the debate over "reaction" content, and being the most used app overnight by young people?
Answer: YouTube

Evidence List:
Title: Sony Music's artists aren't involved in YouTube's new voice-cloning AI experiment.
Source: Music Business Worldwide
Published Time: 2023-11-23T18:48:48+00:00
Fact: During this period of discussion, YouTube has made a number of positive announcements regarding the biggest issue for any rightsholder regarding AI-driven voice replication of artists: their ability to police it.

Title: YouTube demonetizes popular content creator SSSniperwolf after doxxing accusations
Source: Polygon
Published Time: 2023-10-25T18:18:06+00:00
Fact: The debate over "reaction" content on YouTube has been brewing for years, but a recent incident between two creators has refueled the urgency of the conversation.

Title: Cell phone shocker as 97 % of kids use their device during school hours and beyond, says study
Source: FOX News - Health
Published Time: 2023-10-01T09:05:26+00:00
Fact: Overnight phone use was primarily spent engaging with the same media, although YouTube appeared to be the longest-running app because videos were often left playing during the night.

Table 12: 推理问题示例

Query: Did the Cnbc | World Business News Leader report on Nike's net income and the article from The Age on the 10-year Treasury yield both report a decrease in their respective financial metrics?
Answer: Yes

Evidence List:
Title: Nike misses revenue expectations for the first time in two years, beats on earnings and gross margin
Source: Cnbc | World Business News Leader
Published Time: 2023-09-28T20:31:00+00:00
Fact: The company's reported net income for the three-month period that ended August 31 was $ 1.45 billion, or 94 cents per share, compared with $ 1.47 billion, or 93 cents per share, a year earlier.

Title: ASX set to open higher as Wall Street rebounds; $ A rises
Source: The Age
Published Time: 2023-10-04T21:01:01+00:00
Fact: The yield on the 10-year Treasury, which is the centrepiece of the bond market, pulled back from its highest level since 2007, down to 4.73 per cent from 4.80 per cent late on Tuesday.

Table 13: 比较问题的例子

Query: Was the performance of the Chicago Bears' defense reported as improved by Yardbarker after Sporting News highlighted a sack by the Bears' defense on Joshua Dobbs during the NFL 'Monday Night Football' game? Answer: Yes
Evidence List: Title: Bears vs. Vikings live score, updates, highlights from NFL 'Monday Night Football' game Source: Sporting News Published Time: 2023-11-27T23:32:04+00:00 Fact: The Bears answer right back and sack Dobbs, with Sweat and Brisker in there to take him down. Title: Hottest seat on each NFC team: Buns burning for these four head coaches Source: Yardbarker Published Time: 2023-11-30T22:29:33+00:00 Fact: In his second season as HC, the defense has improved, but positive results are hard to come by behind a lackluster offense ranked 19th in yards (323.2) and 21st in points per game (20.2).

Table 14: 时效性问题的例子

Query: What is the first letter of the CEO's last name in the news article from Bloomberg on TomTom, and what is the first letter of the city where the company's headquarters is located in the news article from Reuters? Answer: Insufficient information.

Table 15: 否定拒绝问题的例子