MINTQA: A Multi-Hop Question Answering Benchmark for Evaluating LLMs on New and Long-tail Knowledge

Jie He[1*] **Nan Hu**[2*] **Wanqiu Long**[4] **Jiaoyan Chen**[3] **Jeff Z. Pan**[1]
[1] School of Informatics, University of Edinburgh, UK
[2] Southeast University, Nanjing, Jiangsu, China
[3] University of Manchester, UK
[4] Amazon
{j.he, j.z.pan}@ed.ac.uk, nanhu@seu.edu.cn, wanqlong@amazon.com
jiaoyan.chen.manchester.ac.uk

Abstract

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) by integrating external knowledge. However, limited research has explored how LLMs effectively leverage RAG techniques for multi-hop Question Answering (QA), particularly when handling knowledge with varying degrees of familiarity. In this paper, we introduce **MINTQA** (**M**ulti-hop **Q**uestion **A**nswering on **N**ew and long**T**ail Knowledge), a benchmark designed to evaluate multi-hop QA performance, including 10,479 question-answer pairs for evaluating old/new knowledge and 17,887 pairs for assessing popular/unpopular knowledge, with each question associated with its sub-questions and answers. This benchmark primarily evaluates the multi-hop reasoning ability of LLMs and their capacity to handle knowledge with varying levels of familiarity during the reasoning process. We evaluate 22 state-of-the-art LLMs using three distinct QA strategies: LLM-based parameterized knowledge QA, direct RAG-enhanced QA, and multi-hop RAG-enhanced QA. Our experiments reveal key challenges in how LLMs handle knowledge with different familiarity and offer insights into improving their multi-hop reasoning capabilities when combined with RAG techniques.[1]

1 Introduction

Large **L**anguage **M**odels (LLMs) have demonstrated remarkable capabilities in Question Answering (QA) (Kamalloo et al., 2023; Wang and Qin, 2024). However, they face significant challenges when handling questions requiring domain specific knowledge or up-to-date information (Pan et al., 2023). Although **R**etrieval-**A**ugmented **G**eneration (RAG) provides an effective strategy for generating responses by incorporating external knowledge

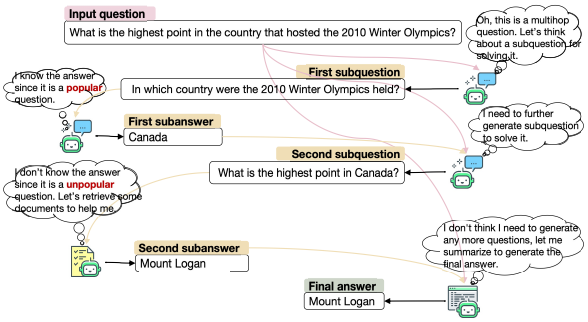


Figure 1: **An example for our benchmark:** Given a complex question, the model must decide whether to decompose it into sub-questions and determine if external knowledge retrieval is required.

(Soudani et al., 2024; Islam et al., 2024), optimizing the combination of LLMs and RAG remains a critical challenge, particularly for multi-hop QA tasks involving knowledge with varying familiarity for models.

Consider a complex question: "What is the highest point in the country that hosted the 2010 Winter Olympics?" As illustrated in Figure 1, answering such questions necessitates decomposing them into sub-questions, such as: (1) In which country were the 2010 Winter Olympics held? and (2) What is the highest point in Canada? Each sub-question may require different knowledge sources. For instance, the first can leverage parametric knowledge, while the second needs related knowledge retrieval.

Current benchmarks for evaluating LLMs on such multi-hop QA scenario have several limitations. First, current studies such as (Sun et al., 2023; Maekawa et al., 2024; Zhang et al., 2024) primarily focus on single-hop queries, leaving complex multi-hop questions largely unexplored. Second, while multi-hop benchmarks such as MultiHop-RAG (Tang and Yang, 2024) assess retrieval effectiveness, they do not systematically evaluate the interaction between question decomposition and retrieval, a capability essential for real-world applications. Furthermore, existing works

[*] Equal Contribution.
[1]The MINTQA benchmark is available at https://github.com/probe2/multi-hop/.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 445–479
July 2-7, 2026

such as FanoutQA (Zhu et al., 2024a) and HotpotQA (Yang et al., 2018) lack assessment of how models handle queries containing new or unpopular knowledge, which presents unique challenges in both decomposition and retrieval.

To address these gaps and facilitate the effective integration of LLMs and RAG techniques to handle different knowledge during multi-hop QA reasoning, we propose MINTQA, a benchmark for evaluating LLMs on complex multi-hop questions across two critical dimensions: **Unpopular knowledge** (information appearing infrequently in training corpora) and **New Knowledge** (recently emerged entities or relationships). Figure 2 outlines our benchmark construction and the evaluation framework based on the benchmark.

We construct MINTQA by systematically collecting knowledge triplets based on Wikipedia and Wikidata and using GPT-4o to generate multi-hop questions spanning one to four hops. The benchmark comprises two sub-datasets: **MINTQA-POP** (17,887 examples) focusing on unpopular/popular knowledge, and **MINTQA-TI** (10,479 examples) examining new/old knowledge, with each example including sub-questions and answers for fine-grained analysis of models' reasoning processes. Table 1 presents a comparison of MINTQA with existing benchmarks, highlighting its unique contributions to multi-hop QA evaluation.

Our framework evaluates LLMs across five aspects: using parametric knowledge, retrieval-augmented generation, sub-question generation, and direct or dynamic decomposition-retrieval. Our comprehensive evaluation of 22 state-of-the-art LLMs reveals: **(1) LLMs face compounding challenges from both reasoning depth and knowledge familiarity.** Even the strongest models see sharp performance drops as the number of reasoning hops increases, and they perform significantly worse on new knowledge than on unpopular but known facts—highlighting that multi-hop reasoning over unfamiliar content remains a fundamental limitation. **(2) While retrieval generally improves performance, it remains difficult to coordinate effectively across reasoning steps.** Models often succeed when all facts involved are either familiar (e.g., popular or old) or unfamiliar (e.g., rare or new), but they struggle substantially when questions require reasoning over mixed knowledge—such as starting from well-known facts and bridging to novel ones—due to inconsistent retrieval behavior and unstable confidence in when and what to retrieve. **(3) Naïve decomposition may hurt performance, especially when dealing with unfamiliar knowledge.** Although providing gold sub-questions leads to significant gains (e.g., a 33% accuracy increase on MINTQA-POP), models often perform worse when generating sub-questions themselves—particularly for new knowledge—suggesting that sub-question generation is a key bottleneck in multi-hop QA. **(4) Only large models benefit from integrating decomposition and retrieval.** While decomposition-then-retrieval or dynamic retrieval strategies improve performance for models with more than 14B parameters, smaller models tend to fail both in decomposing questions effectively and in making retrieval decisions, resulting in degraded performance. **(5) Even with gold sub-questions and gold documents, models still struggle with synthesizing and reasoning across multiple steps.** These findings highlight MINTQA's value in exposing new challenges at the intersection of reasoning depth and knowledge familiarity, offering a more fine-grained and realistic evaluation of multi-hop QA capabilities.

2 Related Work

2.1 Multi-hop Question Answering (QA)

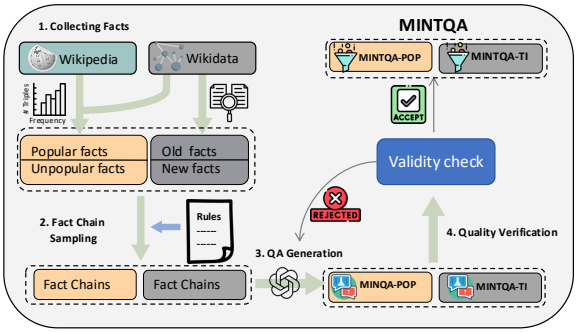
Multi-hop QA challenges LLMs by requiring synthesis and reasoning across multiple sources (Huang and Chang, 2023; Feng et al., 2020; Khashabi et al., 2019). While researchers have proposed decomposing complex questions into sub-questions (Min et al., 2019; Wang et al., 2022, 2023; Liu et al., 2024), generating relevant sub-questions and reasoning chains remains challenging. Existing benchmarks (Zhang et al., 2024; Zhu et al., 2024a; Tang and Yang, 2024) assess retrieval and multi-hop reasoning, but overlook when and how to retrieve, interactions between decomposition and retrieval, or queries with new and unpopular knowledge. Our MINTQA fills these gaps by systematically evaluating LLMs on multi-hop QA.

2.2 Retrieval Augmented Generation (RAG)

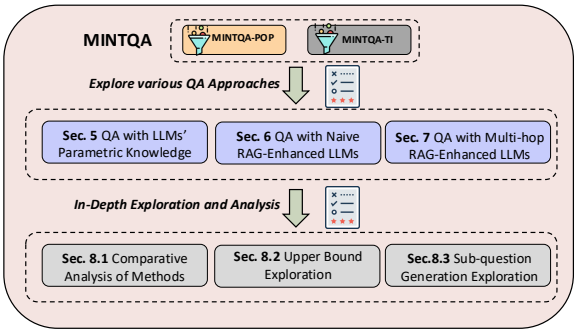
RAG enhances LLMs' performance in multi-question answering by providing access to external documents (Lewis et al., 2020; Xiong et al., 2020; He et al., 2025), particularly for knowledge-intensive tasks (Yu et al., 2020; Zhu et al., 2023). In sub-question generation, RAG can verify and correct LLMs' outputs (Zhao et al., 2023; Shi et al.,

Dataset	# Questions	Multi-hop	Time	Popularity	Sub-questions	Original of Generated Questions
PopQA (Mallen et al., 2022)	14,268	×	×	✓	×	Templates
WiTQA (Maekawa et al., 2024)	14,837	×	×	✓	×	Machine
Head-to-tail (Sun et al., 2023)	18,171	×	×	✓	×	Template
RetrievalQA (Zhang et al., 2024)	1,271	×	✓	✓	×	Mixed
FreshQA (Vu et al., 2024)	600	✓	✓	×	×	Human
MultihopQA-RAG (Tang and Yang, 2024)	2,556	✓	×	×	×	Machine
HotpotQA (Yang et al., 2018)	112,779	✓	×	×	×	Human
MuSiQue (Trivedi et al., 2021)	24,814	✓	×	×	×	Human
2WikiMultiHopQA (Ho et al., 2020)	192,606	✓	×	×	×	Templates
Mintaka (Sen et al., 2022)	20,000	✓	×	×	×	Human
FanoutQA (Zhu et al., 2024a)	1,034	✓	×	×	✓	Human
MINTQA (Ours)	28,366	✓	✓	✓	✓	Machine

Table 1: Comparison between MINTQA and other datasets.



(a) Benchmark Construction



(b) Benchmark Evaluation

Figure 2: Illustration of the MINTQA construction process (a) and the evaluation framework leveraging MINTQA (b).

2024a). However, irrelevant retrievals can introduce noise (Yoran et al., 2024; Joren et al., 2024; He et al., 2026), and external knowledge may override model's inherent knowledge (Xu, 2023; Li et al., 2022), while adding computational overhead (Zhu et al., 2024b). While Jeong et al. (2024) propose using a classifier to determine retrieval necessity, our research investigates LLMs' inherent ability to recognize when retrieval is needed for sub-questions.

2.3 Evaluation of LLMs-based QA

Existing QA datasets for evaluating retrieval-augmented LLMs fall into two main types: reasoning-focused (e.g., MuSiQue (Trivedi et al., 2021), FanOutQA (Zhu et al., 2024a), MultiHop-RAG (Tang and Yang, 2024)) and long-tail (e.g., WiTQA (Maekawa et al., 2024), Head-to-Tail (Sun et al., 2023)). Our work extends these by jointly addressing long-tail and new-fact multi-hop QA, while analyzing sub-question generation and retrieval. Unlike RetrievalQA (Zhang et al., 2024), which integrates existing datasets and is limited to short-form QA, we generate questions with language models for scalable construction. Compared to FreshQA (Vu et al., 2024), which introduces new knowledge but remains small and manually created, our benchmark offers larger-scale coverage and explicit sub-questions, enabling a more comprehensive evaluation of retrieval-augmented LLMs.

3 Benchmark Construction

This section presents our comprehensive methodology for constructing two multi-hop QA benchmarks: MINTQA-POP and MINTQA-TI, designed to evaluate LLM across two critical dimensions: knowledge popularity (popular versus unpopular) and knowledge freshness (new versus old). We first present the data construction methodology for MINTQA-POP (Section 3.1), then detail the construction process of MINTQA-TI, following a similar procedure but focusing on new/old knowledge (Section 3.2), and finally describe our QA generation process (Section 3.3) and present comprehensive statistics of our constructed datasets (Section 3.4).

3.1 Data Construction of MINTQA-POP

Collecting Facts We gather a collection of facts with popularity, denoted as $\mathcal{G}_{pop} = \{(s,r,o),p|(s,r,o) \in \mathcal{G}, p \in \mathbb{Z}^+)\}$, where $\mathcal{G}$ refers to Wikidata, (s,r,o) represents a triple, p indicates the popularity as the positive integer set $\mathbb{Z}^+$. The triples are extracted from Wikipedia (version 2024-05-01). Specifically, we extract raw triples in the format of (Head Span, Relation Span, Tail

MINTQA-POP	1-hop	2-hop	3-hop	4-hop	Total
#Samples	5,894	4,428	4,664	2,901	17,887
Avg. In. Len.	8.91	13.48	15.94	19.85	13.65
Avg. Out. Len.	1.96	1.41	1.33	1.93	1.65
Avg. Ctx. Len.	32.93	375.93	549.18	706.12	361.63
#Relations	124	84	85	98	140
#Entities	7,482	4,357	5,191	3,180	18,501
MINTQA-TI					
#Samples	3,949	2,198	2,057	2,275	10,479
Avg. In. Len.	8.61	13.86	16.58	19.76	13.70
Avg. Out. Len.	2.04	2.61	2.12	2.24	2.22
Avg. Ctx. Len.	56.44	321.32	487.22	645.61	324.47
#Relations	123	147	149	147	189
#Entities	4,096	2,484	2,250	2,346	9,616

Table 2: Data statistics of MINTQA.

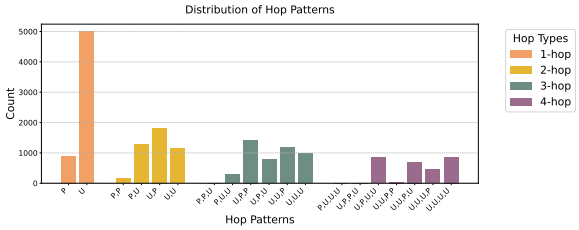


Figure 3: Popularity Related Data Distribution

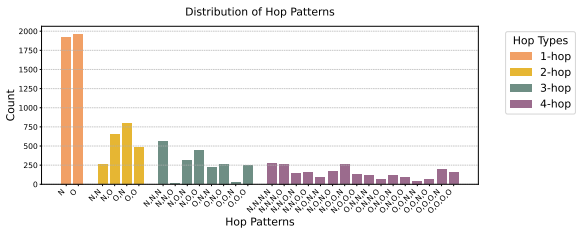


Figure 4: Time Related Data Distribution

Span) from Wikipedia passages using an existing information extraction tool[2]. The three spans of each of these raw triples are linked to entities and relations of Wikidata (version 2024-04-22) using WikiMapper[3], producing corresponding Wikidata triples in form of (s, r, o) where s, r and o are all Wikidata IDs. The popularity p of each triple is calculated as the occurrence frequency of its original raw triple across the entire Wikipedia corpus.

Sampling fact chains We sample facts from $\mathcal{G}_{pop}$ and concatenate them into a chain $\mathcal{FC} = \{(s_1, r_1, o_1), \ldots, (s_n, r_n, o_n)\}$ as the grounded facts of a multi-hop question. We label facts in $\mathcal{G}_{pop}$ based on their popularity scores with **unpopular** ($p \in [1, 10)$), **popular** ($p \in [50, \infty)$), and **other** ($p \in [10, 50)$). A fact chain $\mathcal{FC}$ is constructed as an ordered sequence of connected triples: $\mathcal{FC} = \{(s_1, r_1, o_1), \ldots, (s_n, r_n, o_n)\}$, where $n \leq 4$ and each triple is of either **popular** or **unpopular**. This construction follows five key constraints:

1. **Connectivity**: $o_i = s_{i+1}$ for all $i \in \{1, \ldots, n-1\}$.
2. **Acyclicity**: $o_i \neq s_j$ for all $i, j \in \{1, \ldots, n\}$.
3. **Uniqueness**: No fact chain $\mathcal{FC}$ can be a subchain of another fact chain.
4. **No Shortcuts**: For each fact chain $\mathcal{FC}$, there does not exist a triple (s_i, r, o_j) of **popular** such that $j > i+1$, where $i \in \{1, \ldots, n-1\}$ and $j \in \{2, \ldots, n\}$.
5. **Single Object**: For each triple (s_i, r_i, o_i) in the fact chain, there does not exist another triple (s_i, r_i, o_j) in Wikidata with the same subject and relation but a different object.

3.2 Data Construction of MINTQA-TI

Building on the methodology established for MINTQA-POP, we construct MINTQA-TI, focusing on old and new knowledge. To construct the dataset, we extract two versions of Wikidata: 2021-06-21 and 2024-06-05. We identify triples that exist in both versions or in only one version. These triples form a knowledge graph $\mathcal{G}_{ti}$. We label triples present in both Wikidata versions as **old**, and triples present in the newer version alone as **new**, characterized by a new subject, relation, or object. We create fact chains consisting of new and old knowledge from $\mathcal{G}_{ti}$ by reusing the fact chain construction constraints and method in Section 3.1, except that the original labels of **popular** and **unpopular** are replaced by **old** and **new**, respectively).

3.3 QA Generation and Verification

Following WiTQA (Maekawa et al., 2024), we employ GPT-4o to automatically generate questions from the extracted fact chains, overcoming the diversity and scalability issues of template-based methods like PopQA (Mallen et al., 2022) and the high costs of manual annotation. Given a fact chain $\mathcal{FC} = \{(s_1, r_1, o_1), \ldots, (s_n, r_n, o_n)\}$, we aim to generate a question about s_1 that yields o_n as the answer. To enhance generation quality, we provide one demonstration example per hop. And to ensure validity, we verify questions by having the model answer them using source contexts; only questions yielding o_n are retained. Invalid questions are regenerated up to three times, and unsatisfactory examples are discarded. For multi-hop questions (hop count $\geq$ 2), sub-questions for each intermediate fact are also generated and validated. Examples are included in the dataset only if the main question and all sub-questions pass validation. This process filtered out 138 and 67 examples from MINTQA-

[2]https://github.com/Babelscape/rebel
[3]https://github.com/jcklie/wikimapper

POP and MINTQA-TI, respectively. Prompts and examples are in Appendices C and D.

3.4 Dataset Statistics

Table 2 summarizes the statistics of the MINTQA-POP and MINTQA-TI datasets, which exhibit diverse coverage across multiple dimensions. MINTQA-POP contains 17,887 examples and MINTQA-TI 10,479, with over 2,000 examples per hop category, ensuring robust evaluation. The datasets include 18,501 and 9,616 entities, and 140 and 198 relationships, respectively, demonstrating their diversity. As the number of hops increases, the average context length grows, requiring models to retrieve more documents and face greater challenges.

3.5 Data Type Distribution

Figures 3 and 4 show the distributions of popular versus unpopular facts and old versus new facts within each hop category. Due to our focus on model performance with unpopular and new facts, we sampled more of these fact types. Certain fact combinations such as "P,P,P", three-hop chains composed entirely of popular facts, do not occur in our dataset, so they are not shown in the figures. For more details about the data distribution, see the App. A.

4 Experimental Setup

4.1 Language Models and Configurations

Models We evaluate state-of-the-art LLMs across various architectures and model sizes: GPT-3.5, GPT-4o, GPT-4o mini, LLaMA-3.1/3.2 (Grattafiori et al., 2024), Gemma-2 (Team et al., 2024), Mistral (Jiang et al., 2023), Phi-3 (Abdin et al., 2024), and Qwen2.5 (Hui et al., 2024). All models are instruct versions. For simplicity, we omitted the "instruct" name in the result presentation. To ensure reproducibility, we set the temperature parameter to 0 across all models and accelerated inference using vLLM (Kwon et al., 2023). For more details, please refer to Appendix E.

4.2 Evaluation Metrics

We adopt *Accuracy* (Acc) as the evaluation metric across all experiments, where a prediction is considered correct if the ground-truth answer appears in the model's prediction after converting it to lowercase, following established benchmarks for factual knowledge assessment (Ren et al., 2023; Maekawa et al., 2024; Mallen et al., 2022).

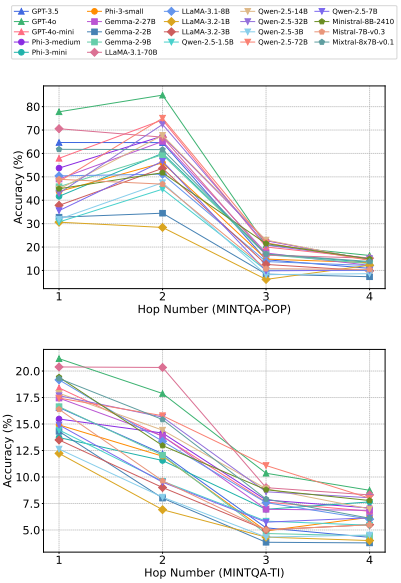


Figure 5: Zero-shot accuracy of different LLMs across various hops.

4.3 Research Questions

Based on the two proposed multi-hop QA sub-datasets, MINTQA-POP, which involves popular/unpopular knowledge, and MINTQA-TI, involving old/new knowledge, we investigate the following research questions:

RQ_1: How do various LLMs perform in the two multi-hop QA scenarios relying solely on their internal knowledge? (Section 5) RQ_2: How do LLMs perform, when enhanced with a direct retrieval approach, in the multi-hop QA scenarios, and how do different retrievers perform? (Section 6) RQ_3: How do LLMs enhanced with a multi-hop RAG strategy (decomposition-then-retrieval approach) perform in the multi-hop QA scenarios, particularly when handling popular/unpopular and old/new knowledge during the decomposition and retrieval processes? (Section 7.1) RQ_4: Whether decomposition-dynamic retrieval can help achieve an optimal balance between performance and efficiency, given the different knowledge types (popular/unpopular, old/new knowledge) involved in multi-hop QA? (Section 7.2)

5 LLMs' Performance on MINTQA with Parametric Knowledge

We evaluate LLMs on MINTQA using their parametric knowledge to understand intrinsic model capabilities and dataset challenges. Specifically, we prompt LLM using its own knowledge to directly answer questions, as shown in Table 16. The

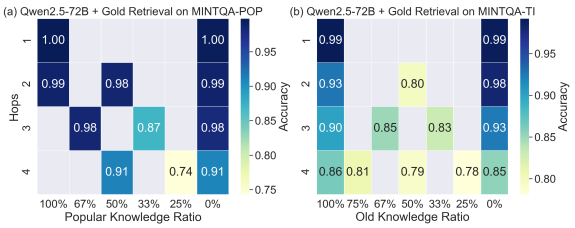


Figure 6: The performance of Qwen2.5-72B with gold retrieval across two datasets. The X-axis represents the proportion of popular knowledge required in the question, and the Y-axis indicates question hops.

results are shown in Figure 5 and further elaborated in App. F.1. Our findings reveal **significant performance gaps between the MINTQA-POP and MINTQA-TI**. Models perform reasonably on MINTQA-POP (e.g., GPT-4o: 77.79%, LLaMA3.1-8B: 50.42% for single-hop questions) but struggle on MINTQA-TI, with GPT-4o's accuracy dropping to 21.17% for single-hop questions. This confirms MINTQA-TI's effectiveness in evaluating knowledge beyond training data, and low performance across models from LLaMA-3.2-1B (7.78%) to GPT-4o (21.17%) demonstrates scaling model size alone doesn't address this. Moreover, **increased reasoning complexity further highlights these limitations**. On MINTQA-POP, performance drops sharply for three-hop (20.03%) and four-hop (16.41%) questions, while on MINTQA-TI, accuracy consistently declines with complexity.

6 Effectiveness of Direct Retrieval

After analyzing the performance of LLMs using only their parametric knowledge on MINTQA in Section 5, we find that LLMs struggle to answer MINTQA questions independently. In this section, we follow prior work (Mallen et al., 2022; Maekawa et al., 2024) to assess the effectiveness of the naive RAG approach, direct retrieval, when applied to LLMs for handling our complex multi-hop questions. For each retrieval, we select the top-5 passages that are relevant to the question and input them as context. The prompt is shown in Table 17.

6.1 Retrieval Model Setup

Retrieval Models We evaluate seven retrieval approaches across three categories: 1) Sparse retriever: **BM25** (Robertson and Zaragoza, 2009). 2) Vector retrievers pre-trained on large unlabeled corpora: **Contriever** (Izacard et al., 2021), **GTR-LARGE/XL** (Ni et al., 2021) and **BGE** (Xiao et al., 2023). 3) Instruction-tuned text embedding retrievers: **Instructor-XL** (Su et al., 2022) and **Promptriever** (Weller et al., 2024)..

Configuration We follow the approach of Yu et al. (2023) to construct the retrieval corpus by linearizing the knowledge graph $\mathcal{G}$ into text. $\mathcal{G}$ consists of $\mathcal{G}_{pop}$ (Section 3.1) and $\mathcal{G}_{ti}$ (Section 3.2). See Appendix E.2 for details.

6.2 Performance Analysis

Figure 7 demonstrates that **retrieval significantly enhances performance**, especially on MINTQA-TI, with an average 30% accuracy gain over the Vanilla setting (no retrieval). Similar trends are observed on MINTQA-POP (refer to Figure 14). Notably, in the Oracle setting, where gold-standard passages are used, even small models like Llama-3.2-1B achieve a 25% accuracy improvement compared to the average performance of all retrievers we used, emphasizing the potential for better retrievers. Appendix F.3 provides more analysis of the retriever.

We analyze the impact of knowledge popularity and newness on QA performance. **Models with different retrievers show inconsistent patterns when varying proportions of new and popular knowledge**. To isolate retrieval quality, we pair models with gold retrieval. Figures 6(a) and (b) show that with Qwen2.5-72B with gold retrieval, performance initially declines and then improves as the proportion of popular or old knowledge decreases. This occurs likely because the model effectively determines whether using parametric knowledge and retrieval for fully familiar (100% popular/old) or unfamiliar (100% unpopular/new) questions but struggles with mixed knowledge, leading to errors. Further analyses are in Appendix F.3.

7 Enhancing Multi-hop QA through Integrating Decomposition and Retrieval

Li and Peng (2023) and Shi et al. (2024b) highlight the importance of effectively combining question decomposition and retrieval for solving multi-hop questions. In this section, we explore two advanced RAG strategies (Decomposition-then-Retrieval and Decomposition-Dynamic Retrieval).

7.1 Decomposition-then-Retrieval

Follow the experimental methodology of IRCOT (Trivedi et al., 2023), we implement an iterative **decomposition-then-retrieval (DTR)** approach

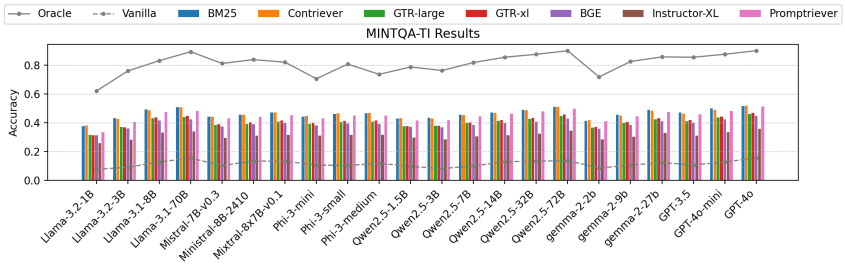


Figure 7: Performance comparison of LLMs on MINTQA-TI using different retrieval methods: "Oracle" uses gold-standard retrieval passages, while "Vanilla" involves models answering without retrieval content.

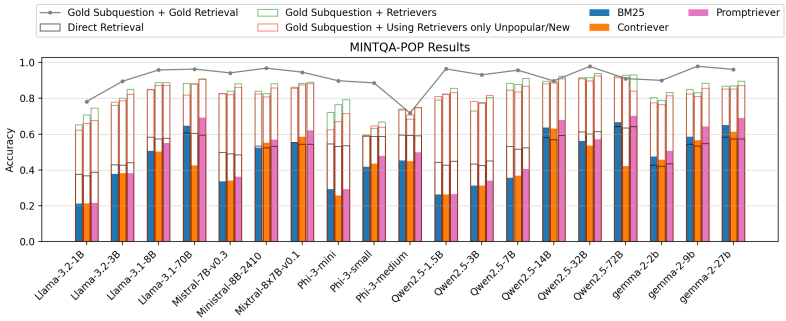


Figure 8: Performance comparison of all models using three retrievers under the decomposition-then-retrieval (DTR) approach on the MINTQA-POP dataset (represented by bars with three colors). **Gold Subquestion + Gold Retriever** indicates that the model utilizes gold subquestions and gold retrieval results. **Gold Subquestion + Retrievers** indicates that the model uses gold subquestions and employs different retrievers for retrieval. **Gold Subquestion + Using Retrievers Only Unpopular/New** denotes that the model uses gold subquestions and retrieves only for subquestions involving unpopular or new knowledge, while relying on the model itself to directly answer subquestions involving popular or old knowledge. **Direct Retrieval** refers to the use of different retrievers for direct retrieval (see Section 6) instead of adopting the DTR approach. See full results in Figure 13.

for multi-hop QA. At each step, the LLM determines one of two options: (1) further decomposing the question into sub-questions, or (2) summarizing a final answer based on the results of previously resolved sub-questions. If option (1) is selected, the LLM uses the full history of sub-questions and their corresponding answers as context to generate a new sub-question. Subsequently, five relevant documents are retrieved to assist in answering the new sub-question. If option (2) is chosen, the LLM summarizes the entire history, including all previously resolved sub-questions and their results, to produce a final answer. This iterative process concludes when either option (2) is selected or a maximum of five iterations is reached. The prompts used for both decision-making and summarization are shown in Tables 21 and 22. We evaluate this approach using BM25, Contriever, and Promptriever[4].

Figure 8 shows partial results and Figure 13 shows the full results. On MINTQA-POP, larger models (>14B) benefit from decomposition and retrieval compared to direct retrieval, while smaller models (<8B) perform worse due to decomposition errors. On MINTQA-TI, direct retrieval outperforms decomposition-then-retrieval for most models, suggesting new knowledge poses greater challenges than question decomposition.

7.2 Decomposition-Dynamic Retrieval

The iterative DTR strategy in Section 7.1 faces two key challenges: high computational overhead from repeated retrievals (Zhuang et al., 2024) and performance degradation from unnecessary retrievals (Mallen et al., 2022; Maekawa et al., 2024). To address this, we explore the **decomposition-dynamic retrieval (DDR)** approach that whether LLMs can dynamically determine retrieval necessity. Following (Ni et al., 2024), we implement a confidence-guided retrieval mechanism, where LLM determines whether to directly answer a sub-question or adopt the retrieval action when it has low con-

[4]GPT models were excluded due to high cost and limited performance advantages over open-source LLMs (70B+).

Model	BM25		
	Acc (%) ↑	Avg. Sub ↓	Avg. Ret ↓
MINTQA-POP			
Qwen2.5-1.5B	25.86 (-0.5)	**1.13 (1.0)**	0.32 (0.19)
Qwen2.5-3B	31.16 (-0.13)	1.78 (0.95)	1.54 (0.71)
Qwen2.5-7B	32.98 (-2.68)	2.18 (0.97)	1.14 (-0.07)
Qwen2.5-14B	53.77 (-10.02)	3.44 (1.1)	1.22 (-1.12)
Qwen2.5-32B	50.33 (-5.86)	2.79 (1.02)	1.18 (-0.59)
Qwen2.5-72B	58.63 (-7.93)	3.01 (1.1)	1.34 (-0.57)
LLaMA-3.2-1B	20.53 (-0.75)	1.79 (1.0)	1.36 (0.57)
LLaMA-3.2-3B	37.23 (-0.64)	3.48 (0.72)	3.26 (0.5)
LLaMA-3.1-8B	50.01 (-0.72)	3.88 (0.56)	3.79 (0.47)
LLaMA-3.1-70B	**64.80 (0.17)**	3.41 (0.89)	3.39 (0.87)
Mistral-7B-v0.3	29.47 (-4.23)	3.13 (0.76)	1.84 (-0.53)
Ministral-8B-2410	35.91 (-16.94)	2.95 (0.96)	**0.02 (-1.97)**
Mixtral-8x7B-v0.1	48.28 (-7.08)	3.61 (0.77)	1.29 (-1.55)
Phi-3-mini	26.81 (-2.54)	4.75 (0.04)	2.23 (-2.58)
Phi-3-small	37.09 (-4.58)	2.92 (0.44)	0.78 (-1.7)
Phi-3-medium	40.16 (-5.24)	2.98 (0.87)	1.08 (-1.03)
Gemma-2-2B	34.20 (-13.31)	4.96 (0.13)	0.98 (-3.85)
Gemma-2-9B	39.99 (-18.52)	3.93 (0.18)	0.32 (-3.43)
Gemma-2-27B	64.64 (-0.37)	4.05 (0.74)	4.01 (0.7)

Table 3: The results for Decomposition-Dynamic Retrieval approach. **Acc** represents the accuracy of the model in answering questions, **Avg. Sub** indicates the average number of sub-questions generated by the model, **Avg. Ret** refers to the average number of sub-questions that are deemed necessary for retrieval by the model. The value in brackets indicates the value of DDR minus that of DTR.

fidence for answering the sub-question (details in App. E). Table 3 shows some results, with complete results in App. F.4. The prompt used for this experiment is shown in Table 23.

Our analysis reveals two key findings. **First, reducing retrievals while maintaining performance proves challenging, with only the largest models (LLaMA-3.1-70B and Gemma-2-27B) maintaining accuracy despite high retrieval rates (>98%)**. Other models show significant performance drops, reflecting our datasets' emphasis on rare and new information. **Second, models exhibit varying retrieval dependencies.** Mistral and Phi models show high self-confidence (55% retrieval rate), LLaMA variants consistently trigger retrieval (>90%), while Gemma models exhibit size-dependent behavior, with retrieval rates ranging from <10% (2-9B) to >98% (2-27B) on MINTQA-POP.

8 More Investigations

8.1 Comparison of DTR and DDR

In Tables 3 and 13, we can compare the performance of the DTR strategy and the DDR strategy in terms of accuracy (Acc), average number of sub-questions (Avg.Sub), and average number of retrieved sub-questions (Avg.Ret). In terms of accuracy, most models perform worse with the DDR strategy compared to the DTR strategy. Under DDR, models generate more sub-questions than with DTR, but retrieve for fewer sub-questions on average. This suggests that while DDR improves the retrieval efficiency of the model, it comes at the cost of accuracy. The observed accuracy drop indicates that, **for current models on the MINTQA dataset, there is a trade-off between accuracy and efficiency that needs further improvement**.

8.2 Oracle Analysis with Gold Component

We evaluate system limitations using gold-standard sub-questions and their retrieved documents. Figure 8 shows notable gains across all models and retrievers when using gold sub-questions (i.e. Gold Subquestion + Retrievers), especially for smaller LLMs, highlighting their difficulties in generating accurate sub-questions independently. Additionally, previous work (Mallen et al., 2022; Maekawa et al., 2024) has shown that retrieval introduces noise that hurts QA performance when answering popular single-hop questions, and we verify this conclusion in the complex multi-hop question scenario by having LLM perform retrieval only on unpopular or new questions. The results (i.e. Gold Subquestion + Using Retrievers only Unpop/New v.s. Gold Subquestion + Retrievers) show that answering complex multi-hop questions generally yields better QA performance when using a common retriever that always performs the retrieval operation.

Notably, even with perfect decomposition and relevant documents (i.e. Gold Subquestion + Gold Retrieval), the accuracy of various LLM remains below 100%. This reveals two challenges: **extracting relevant information from documents containing multiple facts and synthesizing information across sub-questions, suggesting areas for future improvement beyond retrieval and decomposition**.

9 Discussion on Future Directions

Based on the findings and limitations highlighted in this study, we propose several promising directions for future research:

▷ *Enhanced Sub-Question Generation and Planning* Smaller models (<14B) struggle with sub-question generation, suggesting a need for specialized training or architectural improvements. Future work could explore fine-tuning on decomposed reasoning chains or integrating reinforcement learning to optimize decomposition strategies. Hybrid approaches combining rule-based methods with LLM-driven planning may also help.

▷ *Adaptive Retrieval-Augmented Strategies* Models often over-rely on retrieval, leading to inefficiencies. Future research could focus on confidence calibration, uncertainty-aware retrieval triggers, or lightweight classifiers to reduce unnecessary retrievals while maintaining accuracy. Integrating retrieval necessity prediction into training could further optimize performance.

▷ *Knowledge-Type-Aware Reasoning* Models struggle to distinguish between popular/unpopular and old/new knowledge, leading to suboptimal strategy selection. Future frameworks could incorporate explicit knowledge-type classifiers or metadata to guide retrieval and parametric knowledge usage, improving decision-making during inference.

▷ *Cross-Hop Information Synthesis* Even with perfect retrieval and decomposition, models fail to fully synthesize information across sub-questions. This calls for the exploration of novel techniques and architectures that can effectively integrate knowledge from multiple sub-questions, leading to more comprehensive and accurate answers.

10 Conclusion

In this work, we introduce **MINTQA**, a multi-hop QA benchmark that requires reasoning across popular/unpopular and old/new knowledge. MINTQA spans reasoning chains from one to four hops, enabling systematic assessment of LLMs' complex reasoning abilities. We also propose a comprehensive evaluation framework to assess key aspects of multi-hop QA, including the effectiveness of different question decomposition and retrieval strategies, which allows detailed analysis of models' reasoning capabilities. Extensive evaluation on state-of-the-art LLMs reveals that even the best LLM with retrieval still struggle on our benchmark.

Limitation

This work has several key limitations. **First**, our definition of long-tail and new facts relies solely on Wikidata distribution patterns, which may not very accurately reflect knowledge representation in LLMs' diverse pre-training corpora. **Second**, our simplified approach to constructing the retrieval corpus by concatenating entity-related facts into sequential sentences—differs from the complexity of real-world documents and might potentially overestimate the performance of retrieval-augmented methods. **Third**, budget constraints limited our evaluation of powerful closed-source models like GPT-4, though preliminary results suggest our benchmark remains challenging even for these advanced systems. **Fourth**, using only GPT-4o for eval data generation may cause bias. However, we counter that this bias is negligible. We chose GPT-4o for its superior quality. The data-gen task is simple, and our focus is on diverse knowledge and retrieval. GPT-4o mainly adds aux words, not affecting factual QA. Also, single-LLM dataset construction is common in related work (Maekawa et al., 2024; Zhong et al., 2023). Regarding methodology, while our prompting strategy proved effective on the sampled data, we did not explore advanced techniques such as iterative prompt optimization. However, we hypothesize that such optimizations would yield limited improvements, as the core challenge lies in models' knowledge gaps.

Acknowledgments

We gratefully acknowledge the Edinburgh International Data Facility (EIDF), and the Data-Driven Innovation Programme at the University of Edinburgh for providing computational resources and support for the numerical experiments conducted in this paper.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, and Martin Cai etc. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Yufei Feng, Mo Yu, Wenhan Xiong, Xiaoxiao Guo, Junjie Huang, Shiyu Chang, Murray Campbell, Michael Greenspan, and Xiaodan Zhu. 2020. Learning to recover reasoning chains for multi-hop question answering via cooperative games. *Preprint*, arXiv:2004.02393.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, and . Christian Keller etc. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jie He, Richard He Bai, Sinead Williamson, Jeff Z. Pan, Navdeep Jaitly, and Yizhe Zhang. 2026. Clara: Bridging retrieval and generation with continuous latent reasoning. *Preprint*, arXiv:2511.18659.

Jie He, Victor Gutiérrez-Basulto, and Jeff Z. Pan. 2025. From sufficiency to reflection: Reinforcement-guided thinking quality in retrieval-augmented reasoning for llms. *Preprint*, arXiv:2507.22716.

Xanh Ho, A. Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *ArXiv*, abs/2011.01060.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. 2024. Qwen2.5-coder technical report. *Preprint*, arXiv:2409.12186.

Shayekh Bin Islam, Md Asib Rahman, K S M Tozammel Hossain, Enamul Hoque, Shafiq R. Joty, and Md. Rizwan Parvez. 2024. Open-rag: Enhanced retrieval-augmented reasoning with open-source large language models. *ArXiv*, abs/2410.01782.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *North American Chapter of the Association for Computational Linguistics*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2024. Sufficient context: A new lens on retrieval augmented generation systems. *CoRR*, abs/2411.06037.

Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. *ArXiv*, abs/2305.06984.

Daniel Khashabi, Erfan Sadeqi Azer, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2019. On the possibilities and limitations of multi-hop reasoning under linguistic imperfections. *arXiv: Computation and Language*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Haotong Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Proceedings of the 29th Symposium on Operating Systems Principles*.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix X. Yu, and Surinder Kumar. 2022. Large language models with controllable working memory. *ArXiv*, abs/2211.05110.

Zekai Li and Wei Peng. 2023. Self-adaptive reasoning on sub-questions for multi-hop question answering. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yanming Liu, Xinyue Peng, Xuhong Zhang, Weihao Liu, Jianwei Yin, Jiannan Cao, and Tianyu Du. 2024. Ra-isf: Learning to answer and understand from retrieval augmentation via iterative self-feedback. In *Annual Meeting of the Association for Computational Linguistics*.

Seiji Maekawa, Hayate Iso, Sairam Gurajada, and Nikita Bhutani. 2024. Retrieval helps or hurts? a deeper dive into the efficacy of retrieval augmentation to language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5506–5521, Mexico City, Mexico. Association for Computational Linguistics.

Alex Troy Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Annual Meeting of the Association for Computational Linguistics*.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Annual Meeting of the Association for Computational Linguistics*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers. *ArXiv*, abs/2112.07899.

Shiyu Ni, Keping Bi, J. Guo, and Xueqi Cheng. 2024. When do llms need retrieval augmentation? mitigating llms' overconfidence helps retrieval augmentation. In *Annual Meeting of the Association for Computational Linguistics*.

Jeff Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, et al. 2023. Large language models and knowledge graphs: Opportunities and challenges. *Transactions on Graph Data and Knowledge*.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, J. Liu, Hao Tian, Huaqin Wu, Ji rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *ArXiv*, abs/2307.11019.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. *ArXiv*, abs/2210.01613.

Yucheng Shi, Qiaoyu Tan, Xuansheng Wu, Shaochen Zhong, Kaixiong Zhou, and Ninghao Liu. 2024a. Retrieval-enhanced knowledge editing in language models for multi-hop question answering. In *International Conference on Information and Knowledge Management*.

Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2024b. Generate-then-ground in retrieval-augmented generation for multi-hop question answering. In *Annual Meeting of the Association for Computational Linguistics*.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Fine tuning vs. retrieval augmented generation for less popular knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, page 12–22, New York, NY, USA. Association for Computing Machinery.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. One embedder, any task: Instruction-finetuned text embeddings.

Kai Sun, Y. Xu, Hanwen Zha, Yue Liu, and Xinhsuai Dong. 2023. Head-to-tail: How knowledgeable are large language models (llms)? a.k.a. will llms replace knowledge graphs? *ArXiv*, abs/2308.10168.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *ArXiv*, abs/2401.15391.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, and Sammy Jerome etc. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

H. Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2021. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Jinyuan Wang, Junlong Li, and Hai Zhao. 2023. Self-prompted chain-of-thought on large language models for open-domain multi-hop reasoning. *ArXiv*, abs/2310.13552.

Shouhui Wang and Biao Qin. 2024. No need for large-scale search: Exploring large language models in complex knowledge base question answering. In *International Conference on Language Resources and Evaluation*.

Siyuan Wang, Zhongyu Wei, Zhihao Fan, Qi Zhang, and Xuanjing Huang. 2022. Locate then ask: Interpretable stepwise reasoning for multi-hop question answering. In *International Conference on Computational Linguistics*.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, et al. 2024.

Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736.

Orion Weller, Benjamin Van Durme, Dawn Lawrie, Ashwin Paranjape, Yuhao Zhang, and Jack Hessel. 2024. Promptriever: Instruction-trained retrievers can be prompted like language models. *ArXiv*, abs/2409.11136.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Xingrun Xing. 2023. Lm-cocktail: Resilient tuning of language models via model merging. *ArXiv*, abs/2311.13534.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oğuz. 2020. Answering complex open-domain questions with multi-hop dense retrieval. *ArXiv*, abs/2009.12756.

Shicheng Xu. 2023. Search-in-the-chain: Towards accurate, credible and traceable large language models for knowledge-intensive tasks.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Donghan Yu, Sheng Zhang, Patrick Ng, Henghui Zhu, Alexander Hanbo Li, Jun Wang, Yiqun Hu, William Yang Wang, Zhiguo Wang, and Bing Xiang. 2023. Decaf: Joint decoding of answers and logical forms for question answering over knowledge bases. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

W. Yu, Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2020. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54:1 – 38.

Zihan Zhang, Meng Fang, and Ling Chen. 2024. Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. *ArXiv*, abs/2402.16457.

Ruochen Zhao, Xingxuan Li, Shafiq R. Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *ArXiv*, abs/2305.03268.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, Singapore. Association for Computational Linguistics.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024a. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Yun Zhu, Jia-Chen Gu, Caitlin Sikora, Ho Ko, Yinxiao Liu, Chu-Cheng Lin, Lei Shu, Liangchen Luo, Lei Meng, Bang Liu, and Jindong Chen. 2024b. Accelerating inference of retrieval-augmented generation via sparse context selection. *ArXiv*, abs/2405.16178.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji rong Wen. 2023. Large language models for information retrieval: A survey. *ArXiv*, abs/2308.07107.

Ziyuan Zhuang, Zhiyang Zhang, Sitao Cheng, Fangkai Yang, Jia Liu, Shujian Huang, Qingwei Lin, S. Rajmohan, Dongmei Zhang, and Qi Zhang. 2024. Efficientrag: Efficient retriever for multi-hop question answering. *ArXiv*, abs/2408.04259.

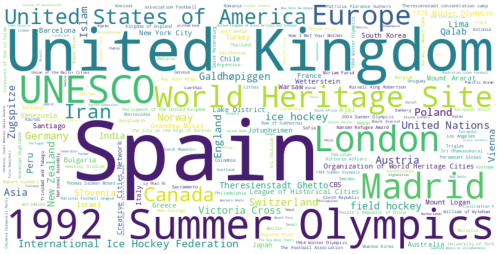


Figure 9: The word cloud of the entities.

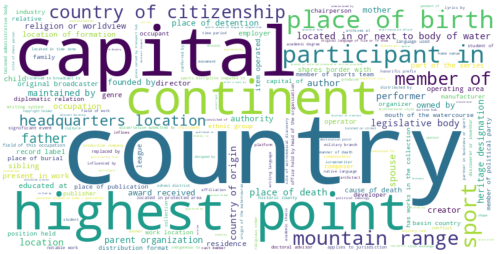


Figure 10: The word cloud of the relations.

A Data Analysis

A.1 Word Cloud Distribution

We conducted a word cloud analysis on the triplets used in our dataset. We can observe from Figure 9 and 10 that the most frequent entities in our constructed dataset are related to geographical information, with "United Kingdom" appearing the most frequently. Following that, there are entities related to events, such as "1992 Summer Olympics".

Regarding relationships, the word "Country" appears the most, followed closely by "Capital". These relationships are also related to geographical information. However, we can observe other noticeably frequent relationships, such as those related to "Place of Birth" and "Participant In", which are associated with individuals and events.

A.2 Analysis of Knowledge Familiarity in MINTQA

To better understand the dataset composition, we analyze the distribution of questions with different proportions of unfamiliar knowledge. During the construction of MINTQA, we aimed to maintain a balanced distribution across both the number of reasoning hops and the types of knowledge involved (e.g., popular vs. unpopular, old vs. new). Table 4 provides detailed statistics, showing the number of instances and their corresponding proportions under different unfamiliar knowledge ratios for both MINTQA-POP and MINTQA-TI. The results demonstrate that the dataset covers a broad spectrum of familiarity levels, which is crucial for evaluating models' robustness under varying degrees of knowledge unfamiliarity.

B Human Sampling Check

Despite our data filtering strategy's success in improving data quality compared to the initial dataset, some failure cases still exist. Our analysis revealed that certain samples in the MINTQAQA dataset share a common limitation: they contain reasoning questions that lack sufficient context for definitive answers. For instance, questions like "Which event did the Iberian Revolutionary Liberation Directory participate in?" demonstrate this issue.

Given MINTQA's role as an evaluation benchmark, we took measures to understand the effects of such cases. We implemented a rigorous human validation process. Two specialists who are English native speakers were hired to conduct a systematic evaluation of our randomly selected 500 instances from MINTQA. Our primary objective was to verify whether each question could be answered definitively with the provided information. The annotators were tasked with assessing whether each question could be answered unambiguously based on the available context. The results were highly encouraging: only 2% of the evaluated samples exhibited contextual insufficiency, and no other significant issues were identified. These findings validate MINTQA's overall quality while also confirming the effectiveness of our sample filtering methodology. This low error rate demonstrates that our quality control pipeline successfully maintains the dataset's integrity and reliability for evaluation purposes.

C Details of Benchmark

C.1 Details of Benchmark Curation

In Section 3, we present a comprehensive description of our benchmark construction methodology. Our approach includes carefully designed prompts for both question generation and validation processes. The complete specifications of these prompts are illustrated in figures 9 through 15.

C.2 License

Our benchmark data are released under the MIT License, which is detailed in https:// open-

Unfamiliar Knowledge Ratio	MINTQA-POP		MINTQA-TI	
	Count	Proportion (%)	Count	Proportion (%)
0%	7988	44.66	2851	27.21
25%	2000	11.18	484	4.62
33%	2264	12.66	728	6.95
50%	3147	17.59	2253	21.50
67%	1423	7.96	545	5.20
75%	–	–	608	5.80
100%	1065	5.95	3010	28.72

Table 4: Distribution of unfamiliar knowledge ratios in MINTQA datasets.

source.org/licenses/MIT.

D Qualitative Analysis

Table 24 to Table 32 present representative examples of multi-hop questions and their corresponding sub-questions generated by GPT-4o for both MINTQA-POP and MINTQA-TI datasets. We have selected three representative instances for each hop level, ranging from single-hop to four-hop questions. As demonstrated in the table, GPT-4o effectively converted the triplets into well-structured, coherent questions. The high quality of these generated questions makes them suitable for evaluating retrieval-augmented LLMs' capabilities in handling multi-hop questions that involve rare and new knowledge.

E Additional Experimental Details

E.1 Implementation Details

In our experiments, we utilized the following state-of-the-art LLMs, with detailed version specifications: GPT-3.5 (gpt-3.5-turbo-1106), GPT4o-min (gpt-4o-mini-2024-07-18), GPT4o (gpt-4o-2024-08-06), LLaMA-3.1-8B (LLaMA-3.1-8B-instruct), LLaMA-3.1-70B (LLaMA-3.1-70B-instruct), LLaMA-3.2-1B (LLaMA-3.2-1B-instruct), LLaMA-3.2-3B (LLaMA-3.2-3B-Instruct), Qwen-2.5-1.5B (Qwen-2.5-1.5B-Instruct), Qwen-2.5-3B (Qwen-2.5-3B-Instruct), Qwen-2.5-7B (Qwen-2.5-7B-Instruct), Qwen-2.5-14B (Qwen-2.5-14B-Instruct), Qwen-2.5-32B (Qwen-2.5-32B-Instruct), Qwen-2.5-72B (Qwen-2.5-72B-Instruct), Gemma-2-2B (Gemma-2-2B-it), Gemma-2-9B (Gemma-2-9B-it), Gemma-2-27B (Gemma-2-27B-it), Phi-3-mini (Phi-3-mini-4k), Phi-3-small (Phi-3-small-8k), Phi-3-medium (Phi-3-medium-4k), Mistral-7B (mistral-7B-instruct-v0.3), Mixtral-8x7B (Mixtral-8x7B-instruct-v0.1), and Ministral-8B (Ministral-8B-instruct-2410). All experiments were conducted using 4 A100 (80GB) GPUs. From Table 16 to 23, we provide the prompts used to instruct these models in completing their respective tasks.

E.2 Retrievers and KG Linearization Details

We evaluate seven retrieval approaches across three categories: 1) Sparse retriever: **BM25** (Robertson and Zaragoza, 2009). 2) Vector retrievers pre-trained on large unlabeled corpora: **Contriever** (Izacard et al., 2021): Fine-tuned on MS-MARCO, **GTR-LARGE/XL** (Ni et al., 2021) and **BGE** (Xiao et al., 2023): Further fine-tuned on NQ (Kwiatkowski et al., 2019) and HotpotQA (Yang et al., 2018). 3) Instruction-tuned text embedding retrievers: **Instructor-XL** (Su et al., 2022): Multi-task trained on 330 tasks for instruction robustness. **Promptriever** (Weller et al., 2024): Uses LLaMA backbone, trained on curated instance-level instruction sets from MS-MARCO, demonstrating superior retrieval performance compared to Instructor-XL.

We linearise the knowledge graph (KG) $\mathcal{G}$ as a source of text retrieval in the corpus, with reference to the work in Yu et al. (2023). Specifically, for each entity in $\mathcal{G}$, we extract a 1-hop subgraph centered on the entity and convert it into linearized text, treating it as a passage. Since $\mathcal{G}$ includes both old and new versions of the Wikidata dump, knowledge conflicts may arise due to updates. Conflicting triples are separated into different passages. Each passage is split into chunks of 512 tokens, a size shown to be effective for practical applications (Wang et al., 2024).

Model	POP						TI						
	1	0.67	0.5	0.33	0.25	0	1	0.75	0.67	0.5	0.33	0.25	0
GPT													
GPT-3.5	83.94	2.74	69.59	18.46	9.40	49.34	9.68	4.34	8.79	8.88	2.02	2.30	17.97
GPT-4o	89.11	3.23	87.83	29.95	14.65	63.42	14.73	7.85	15.80	14.03	3.85	6.91	22.89
GPT-4o-mini	84.32	2.88	78.55	26.24	12.30	47.96	12.31	6.20	12.09	11.32	3.49	4.93	18.90
Llama													
Llama-3.2-1B	64.51	0.49	29.30	4.24	5.45	23.59	6.35	2.69	5.36	5.33	2.57	2.96	14.29
Llama-3.2-3B	75.87	1.41	58.91	14.22	7.05	29.17	7.40	5.37	6.18	6.84	3.12	4.44	15.91
Llama-3.1-8B	75.87	2.18	54.91	19.88	7.70	38.29	11.29	5.37	8.65	9.59	5.50	5.43	21.73
Llama-3.1-70B	90.42	2.32	68.99	22.84	12.15	55.23	13.47	7.44	11.95	15.80	4.22	7.57	23.06
Qwen													
Qwen2.5-1.5B	62.44	3.79	49.13	14.84	7.00	22.68	8.49	5.17	8.10	7.32	2.39	4.44	16.41
Qwen2.5-3B	62.82	3.65	51.57	12.06	6.40	23.40	7.89	4.96	5.22	5.68	3.30	4.28	13.62
Qwen2.5-7B	73.62	5.34	61.49	12.81	6.35	26.87	9.40	4.55	7.83	7.81	3.12	4.93	16.15
Qwen2.5-14B	78.59	5.20	73.66	37.68	11.75	35.18	13.29	6.82	12.36	10.39	4.77	5.43	18.97
Qwen2.5-32B	81.13	6.04	76.87	30.17	12.25	35.69	12.56	7.23	13.46	11.45	3.49	7.89	19.47
Qwen2.5-72B	80.56	5.41	78.74	31.85	10.60	40.55	12.63	8.47	14.29	11.90	3.67	5.26	20.63
Gemma													
Gemma-2-2B	57.18	1.05	36.83	10.73	4.40	24.39	7.30	4.13	5.08	5.73	1.47	2.80	16.18
Gemma-2-9B	80.28	1.62	65.33	26.63	11.75	33.99	10.73	4.55	6.04	7.19	2.39	4.28	18.14
Gemma-2-27B	80.56	2.74	70.07	27.83	9.85	37.74	12.17	6.40	9.20	9.50	3.85	7.40	18.87
Phi													
Phi-3-mini	79.53	2.67	64.41	23.06	10.70	33.17	9.61	5.58	11.13	9.19	2.39	7.57	15.45
Phi-3-small	74.74	2.18	60.06	21.33	9.60	34.34	9.93	5.17	7.42	9.41	2.02	5.26	15.85
Phi-3-medium	84.79	2.60	70.35	25.57	11.40	45.97	10.87	6.82	12.50	11.10	2.20	4.28	17.34
Mistral													
Mistral-7B-v0.3	81.31	1.48	48.78	12.81	8.00	36.20	9.12	3.93	7.14	7.55	2.39	4.11	18.21
Ministral-8B-2410	76.06	4.50	56.34	28.75	11.75	35.57	10.84	5.58	11.26	9.28	5.32	8.72	23.29
Mixtral-8x7B-v0.1	84.69	2.74	66.03	24.16	10.95	47.30	13.43	5.58	10.85	10.92	4.22	5.10	20.40

Table 5: The model's accuracy (%) in the zero-shot setting is analyzed within MINTQA-POP and MINTQA-TI, categorized based on the proportion of popular facts and old facts. A value of 0 indicates that the questions are entirely composed of unpopular facts or new facts, with other numbers increasing proportionally.

F Additional Experiments and Result Analysis

F.1 Zero-shot: Performance Across Retrieval Categories

In Table 5, we present the performance of LLMs in a zero-shot evaluation setting across different proportions of unpopular/popular and old/new facts. As observed, the accuracy is highest when questions are composed solely of popular or old facts. For example, LLaMA-3.1-70B achieves an accuracy of 90.42% on MINTQA-POP and 13.47% on MINTQA-TI.

However, as the proportion of unpopular or new facts increases, the accuracy of the models shows a declining trend. Interestingly, when this proportion reaches 1, the accuracy tends to rise compared to lower ratios. This is likely because the proportion of 1 often includes many 1-hop questions, which are comparatively easier for the models to resolve.

F.2 Sub-question Generation Analysis

From Figures 15, we illustrate the relationship between the number of sub-questions generated by models and the corresponding gold sub-question counts. This analysis considers scenarios where models are required to independently generate and answer sub-questions.

We observe substantial differences among models of similar sizes. For instance, the Qwen2.5-7B model tends to generate fewer sub-questions, with most counts falling in the range of 1 or 2. In contrast, the Mistral-7B model produces sub-questions with a more uniform distribution, primarily ranging from 2 to 5. Despite these differences, smaller models, such as Qwen2.5-1.5B and LLaMA-3.2-1B, exhibit similar trends. Both predominantly generate only 1 sub-question, reflecting the limited capability of these smaller LLMs to generate sub-questions as part of their answering process. Examining the distributions of larger models on the MINTQA-POP and MINTQA-TI datasets reveals that, despite differences in the datasets, large models exhibit similar distributions in terms of actual step counts and the number of sub-questions generated by the models.

F.3 More Analysis of Direct Retrieval

Direct retrieval strategy have limitations when handling multi-hop questions. Figure 11 reveals significant limitations in current direct retrieval approaches when handling multi-hop questions. First, the retrieval effectiveness decreases markedly with increased hop count. We observe a consistent decline in recall rates across all retrieval methods as question complexity increases, indicating fundamental limitations in the direct retrieval approach. Second, among the retrieval methods evaluated, BM25 demonstrated the best performance. This can be explained by the highly structured nature of our KG-linearized corpus. While dense retrieval methods excel at capturing semantic similarities in

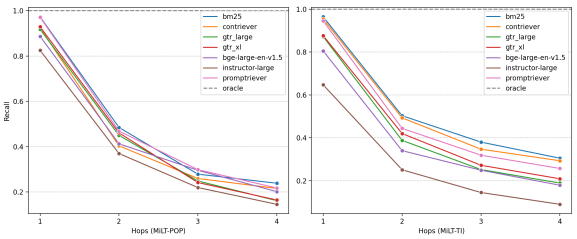


Figure 11: Recall performance of retrieval methods across two datasets for varying question hops.

natural text, BM25's lexical matching approach is well-suited for knowledge graph-derived text.

We demonstrate the influence of knowledge newness and popularity on direct retrieval scenarios, using Qwen2.5-72B paired with BM25 as a representative example. As shown in Figure 12 (a) and (c), QA performance declines with an increasing proportion of unpopular or new knowledge in questions. However, performance improves when the proportion of new knowledge reaches 100% (i.e., no old knowledge), as higher new knowledge presence boosts recall rates (Figure 12 (d)), ultimately enhancing QA accuracy on MINTQA-TI. This highlights the retriever's effectiveness in handling new knowledge.

F.4 Complete Results for Decomposition-Dynamic Retrieval

Table 13 presents the complete results on MINTQA-POP and MINTQA-TI using large models to output confidence scores for sub-questions and determine whether retrieval is needed based on the confidence values. We conducted experiments with three retrievers: BM25, Contrieve, and PromptRetrieve.

G Evaluating LLMs' Decision-Making Capabilities in Multi-hop QA

When evaluating the ability of LLMs to answer questions using their parametric knowledge, we frequently observed "I don't know" responses or no answers at all. This clearly indicates the difficulties LLMs encounter in solving complex multi-hop questions relying solely on their internal knowledge or limited single-step reasoning capabilities. To overcome these challenges, LLMs often employ sub-question decomposition and retrieval strategies. Nevertheless, the effectiveness of these strategies highly depends on the model's capacity to determine when to utilize them. We analyze this from three crucial aspects. Our objective here is to assess the degree to which the model's decisions are consistent with the heuristic labels derived from Wikidata. We regard this as an evaluation of **preference**.

Model	MINTQA-POP		MINTQA-TI	
	Acc	F1	Acc	F1
GPT-3.5	54.82	41.21	52.04	26.67
GPT-4o-mini	65.61	48.67	59.56	28.72
GPT-4o	68.84	51.35	65.46	38.18
LLaMA-3.2-1B	67.05	26.76	62.32	25.59
LLaMA-3.2-3B	56.40	32.91	47.37	33.65
LLaMA-3.1-8B	59.93	25.83	55.55	26.80
LLaMA-3.1-70B	70.03	44.88	63.26	29.54
Qwen2.5-1.5B	64.01	26.47	58.42	24.92
Qwen2.5-3B	47.08	38.89	48.09	41.86
Qwen2.5-7B	70.90	48.23	62.22	34.47
Qwen2.5-14B	69.31	42.08	62.39	26.28
Qwen2.5-32B	28.74	29.80	18.69	11.04
Qwen2.5-72B	65.39	55.33	51.11	37.00
Gemma-2-2B	5.39	3.84	21.07	14.53
Gemma-2-9B	70.90	46.18	63.05	30.39
Gemma-2-27B	73.22	55.67	63.64	36.15
Phi-3-mini	64.47	26.44	56.60	24.72
Phi-3-small	75.79	55.56	62.53	40.50
Phi-3-medium	28.43	21.69	18.50	10.64
Mistral-7B-v0.3	37.58	30.94	26.28	18.88
Ministral-8B-2401	67.07	27.35	62.27	25.66
Mixtral-8x7B-v0.1	28.00	15.14	18.48	10.51

Table 6: The model's accuracy and F1 score for the task of determining question retrieval, sub-question generation, or direct answering.

Model	MINTQA-POP		MINTQA-TI	
	Acc	F1	Acc	F1
GPT-3.5	49.32	49.15	47.32	47.27
GPT-4o-mini	47.77	46.98	47.48	47.14
GPT-4o	37.13	33.19	44.67	43.83
LLaMA-3.2-1B	31.40	24.71	43.07	30.74
LLaMA-3.2-3B	30.82	23.56	43.05	30.10
LLaMA-3.1-8B	30.88	23.66	43.05	30.15
LLaMA-3.1-70B	54.60	54.53	48.46	47.70
Qwen2.5-1.5B	41.76	41.55	51.03	51.03
Qwen2.5-3B	33.42	28.22	43.62	33.43
Qwen2.5-7B	32.11	25.63	43.12	31.24
Qwen2.5-14B	65.33	63.42	53.25	42.62
Qwen2.5-32B	68.13	63.53	53.87	43.39
Qwen2.5-72B	32.25	26.34	43.57	34.20
Gemma-2-2B	30.82	23.56	43.06	30.11
Gemma-2-9B	54.37	54.37	50.99	50.65
Gemma-2-27B	69.39	63.95	55.58	41.72
Phi-3-mini	32.28	26.05	43.23	32.96
Phi-3-small	35.28	30.68	44.35	41.48
Phi-3-medium	40.77	38.12	44.19	42.75
Mistral-7B-v0.3	38.13	36.23	46.37	45.53
Ministral-8B-2410	30.86	23.63	43.06	30.13
Mixtral-8x7B-v0.1	68.29	43.32	56.96	36.97

Table 7: The accuracy and F1 scores of different models in determining whether sub-questions should be retrieved or directly answered.

G.1 Direct Answer vs. Decompositions vs. Retrieval

When encountering multi-hop questions, models must choose between direct answering, sub-question generation, or retrieval. This decision significantly impacts system efficiency and accuracy. Specifically, simple factual questions are often answered directly, while multi-hop or rare fact queries benefit from decomposition or retrieval. The evaluation is conducted on the main question only and involves a three - class classification: decomposition, retrieval, and direct answer. The labels are assigned as follows:

- All queries with $hop\_num \geq 2$ are labeled as decomposition.
- Queries with $hop\_num == 1$ constructed from unpopular or new knowledge are labeled as retrieval.
- Queries with $hop\_num == 1$ constructed from popular or old knowledge are labeled as direct answer.

The purpose of this evaluation is to analyze the model's initial decision - making when presented with a question.

As shown in Table 6, Phi-3-small-8k performs best on MINTQA-POP (Accuracy: 75.79%, F1: 55.56%), while GPT-4o leads on MINTQA-TI (Accuracy: 65.46%, F1: 38.18%). However, model size doesn't always predict performance; Qwen2.5-32B underperforms its 14B variant. Lower-performing models, like Gemma-2-2B, favor direct answering (92.59% on MINTQA-TI), likely due to their limited ability to assess question complexity.

G.2 Direct Answer vs. Retrieval for Sub-questions

When handling sub-questions, models must decide between direct answering and retrieval based on the required knowledge. Popular facts might be answered directly, while tail knowledge or recent information often requires retrieval. This evaluation focuses on queries with $hop\_num \geq 2$. Given the main question and sub-questions, the model must decide whether to retrieve or directly answer each sub-question. The labels are assigned as follows:

- Class A: Sub-questions constructed from popular or old knowledge are labeled as direct answer.
- Class B: Sub-questions constructed from unpopular or new knowledge are labeled as retrieval.

Model	MINTQA-POP		MINTQA-TI	
	Acc	F1	Acc	F1
GPT-3.5	37.74	30.78	43.74	38.20
GPT-4o-mini	59.31	58.96	56.25	56.02
GPT-4o	71.30	71.22	59.30	57.91
LLaMA-3.2-1B	23.14	20.09	33.86	23.47
LLaMA-3.2-3B	28.62	22.63	37.52	26.98
LLaMA-3.1-8B	34.81	25.82	40.83	28.99
LLaMA-3.1-70B	58.19	57.67	52.43	52.43
Qwen2.5-1.5B	34.81	25.82	40.83	28.99
Qwen2.5-3B	77.49	72.19	59.52	49.39
Qwen2.5-7B	62.16	62.13	53.25	52.85
Qwen2.5-14B	79.05	78.53	60.02	57.83
Qwen2.5-32B	95.94	95.52	62.53	58.74
Qwen2.5-72B	83.68	83.06	62.03	59.54
Gemma-2-2B	34.81	25.82	40.83	28.99
Gemma-2-9B	40.05	34.31	43.79	38.25
Gemma-2-27B	65.83	65.82	55.32	54.93
Phi-3-mini	34.87	25.92	41.11	30.07
Phi-3-small	47.19	44.44	46.71	44.78
Phi-3-medium	35.36	26.77	42.65	35.16
Mistral-7B-v0.3	34.91	25.99	40.87	29.31
Ministral-8B-2410	34.81	25.82	40.88	29.14
Mixtral-8x7B-v0.1	35.96	28.18	41.73	32.52

Table 8: The accuracy and F1 scores of the model in determining whether the main question has been answered based on the given sub-question-answer pair.

Our experiments results in Table 7 reveal a general correlation between model size and decision quality, with some exceptions. LLaMA-3.1-70B outperforms other LLaMA variants, achieving 54.60% and 48.46% accuracy on MINTQA-POP and MINTQA-TI, respectively. However, GPT-4o underperforms GPT-3.5, likely due to overconfidence in its parametric knowledge, as it selects direct answering on 93.48% of MINTQA-POP and 69.20% of MINTQA-TI questions. Additionally, models perform better on MINTQA-TI, indicating new knowledge provides a clearer signal for retrieval compared to knowledge of varying popularity, where the decision boundary is less distinct.

G.3 Decomposition vs. Synthesis

For multi-hop questions (hop count ≥ 2), we evaluate models' ability to decide whether to decompose further or synthesize the final answer from intermediate results. This evaluation is also conducted on queries with $hop\_num \geq 2$. Given the main question, sub-questions, and corresponding sub-answers, the model must decide whether to continue decomposition or synthesize an answer. The labels are assigned as follows:

- Class A: If the number matches the hop number, the model should synthesize an answer.
- Class B: If the number of sub-questions and

Model	Class A		Class B	
	Acc.	F1	Acc.	F1
GPT				
GPT-3.5	89.4	52.1	31.5	46.2
GPT-4o-mini	97.3	53.5	25.7	40.5
GPT-4o	99.7	49.4	9.3	16.9
Llama				
Llama-3.2-1B	99.3	47.1	1.2	2.3
Llama-3.2-3B	100.0	47.1	0.0	0.0
Llama-3.1-8B	100.0	47.1	0.1	0.2
Llama-3.1-70B	94.9	56.3	36.6	52.8
Qwen				
Qwen2.5-1.5B	57.9	38.0	34.6	45.1
Qwen2.5-3B	97.9	47.5	4.7	8.9
Qwen2.5-7B	100.0	47.6	1.9	3.7
Qwen2.5-14B	68.9	55.1	63.7	71.8
Qwen2.5-32B	52.9	50.6	74.9	76.5
Qwen2.5-72B	98.2	47.2	2.8	5.5
Gemma				
Gemma-2-2B	100.0	47.1	0.0	0.0
Gemma-2-9B	88.7	54.5	39.1	54.2
Gemma-2-27B	49.6	50.0	78.2	78.0
Phi				
Phi-3-mini	99.5	47.5	2.4	4.6
Phi-3-small	99.0	48.5	6.9	12.8
Phi-3-medium	99.7	50.9	14.5	25.3
Mistral				
Mistral-7B-v0.3	89.8	47.2	15.1	25.2
Ministral-8B-2410	100.0	47.1	0.1	0.1
Mixtral-8x7B-v0.1	3.1	5.7	97.3	80.9

Table 9: The per-label accuracy and F1 scores for the tasks of sub-question judgment, retrieval, or direct answer generation.

Model	Class A		Class B	
	Acc.	F1	Acc.	F1
GPT				
GPT-3.5	99.7	52.7	4.6	8.8
GPT-4o-mini	98.5	62.8	38.4	55.1
GPT-4o	95.2	69.8	58.5	72.7
Llama				
Llama-3.2-1B	100.0	51.6	0.0	0.0
Llama-3.2-3B	100.0	51.6	0.0	0.0
Llama-3.1-8B	100.0	51.6	0.0	0.0
Llama-3.1-70B	99.6	62.4	36.1	53.0
Qwen				
Qwen2.5-1.5B	100.0	51.6	0.0	0.0
Qwen2.5-3B	48.6	60.1	92.9	84.3
Qwen2.5-7B	93.4	63.2	45.5	61.0
Qwen2.5-14B	91.1	75.2	72.6	81.9
Qwen2.5-32B	93.7	94.1	97.1	96.9
Qwen2.5-72B	92.7	79.8	78.9	86.3
Gemma				
Gemma-2-2B	100.0	51.6	0.0	0.0
Gemma-2-9B	100.0	53.7	8.1	14.9
Gemma-2-27B	97.1	66.4	49.1	65.2
Phi				
Phi-3-mini	100.0	51.7	0.1	0.2
Phi-3-small	99.7	56.8	19.1	32.1
Phi-3-medium	100.0	51.8	0.9	1.7
Mistral				
Mistral-7B-v0.3	100.0	51.7	0.2	0.3
Mixtral-8x7B-v0.1	98.9	51.8	2.3	4.6
Ministral-8B-2410	100.0	51.6	0.0	0.0

Table 10: The per-label accuracy and F1 scores for the task where the model is required to determine whether the answer to the main question has been found, given the sub-questions and their answers.

sub-answers is less than the main question's hop number, the model should continue decomposition.

As shown in Table 8, performance generally correlates with model size. Qwen2.5-32B achieves 95% accuracy on MINTQA-POP but drops to 62.53% on MINTQA-TI, reflecting new knowledge poses challenges for synthesizing. Some models like Mistral-7B, show extreme biases, predicting the main answer always within sub-answers for 99.90% cases of MINTQA-POP.

G.4 Accuracy and F1 Across Categories

Table 9 and 10 reports the accuracy and F1 scores for each category under the evaluation setup described in Section G.2 and G.3. From the table, we can observe that most models demonstrate high accuracy, often exceeding 90% or even reaching 100% in identifying sub-questions that can directly generate answers. However, the F1 scores are significantly lower. This discrepancy indicates that models tend to predict that all examples are solvable, revealing an overconfidence in their ability to answer our constructed benchmarks.

The table also highlights similar phenomena across models, particularly for LLaMA-3.1-8B, LLaMA-3.2-1B, LLaMA-3.2-3B, Qwen2.5-1.5B, Gemma-2-2B, and Ministral-8B-2410. These models consistently predict that the main question can be derived from existing sub-question answers. On the other hand, models in the same series, such as Qwen2.5 variants, exhibit more balanced accuracy and F1 scores across categories. This reflects significant inconsistencies among large models in determining whether sub-question answers suffice to answer the main question.

Such findings indicate the challenges of relying on large models for complex reasoning tasks and highlight the need for more robust evaluation metrics and methodologies.

G.5 Effectiveness of Only Decomposition

In this section, we investigate whether only generating and answering sub-questions or providing sub-questions for answering improves the accuracy on our benchmark. Results can be seen in Table 12.

Error Type	MINTQA-POP Example	TI Example
1. Poor question decomposition	*Q*: Where was Juan R. Correa-Pérez born? *Subq*: What is the population of the city where the 2024 Summer Olympics will be held? *Issue*: Subquestion diverges from the main question.	*Q*: Who is the composer of the work where the place of birth of Elian Morenzo's sibling is present? *Subq*: Who is the place of the composer of the birthplace that features the sibling of Elian Morenzo in the work? *Issue*: Complex composition not decomposed meaningfully.
2. Poor question answering	*Q*: Which country hosted the 2022 South American Games? *Pred*: Chile *Gold*: Paraguay *Issue*: Model skips decomposition and answer the question incorrectly.	*Q*: Where did Felix Braverman die? *Subq*: Where did Felix Braverman die? *Pred*: Warsaw, Poland *Gold*: Vilnius, Lithuania *Issue*: Subquestion is relevant, but answer is wrong due to retrieval failure.
3. Final answer incorrect despite good reasoning	*Q*: Which country is Vaucluse located in? *Subq*: Which Vaucluse (France or Australia)? *Sub-pred*: France *Final-Pred*: Australia *Gold*: France *Issue*: Subquestions and answers are correct, but final prediction is incorrect.	*Q*: What academic degree does Lara Kimura hold? *Subq*: What degree does Lara Kimura hold? *Sub-pred*: Ph.D. in Astrophysics *Final-Pred*: Bachelor's degree *Gold*: Ph.D. in Astrophysics *Issue*: Final summarization failed despite correct subanswer.
4. Output truncation or formatting issue	– Not observed –	*Q*: What is the native language of the spouse of the sibling of Alina Voren's mother? *Subq*: What is the native language of the spouse? *Sub-pred*: The native language is ... *Gold*: Estonian *Issue*: Model output is truncated, incomplete final answer.

Table 11: Representative error examples for each error type in POP and TI datasets (Q = main question, Subq = subquestion, Sub-pred = model prediction for subquestion, Final-pred = model prediction for the main question, Gold = gold answer).

Self-Generated Sub-Questions: On MINTQA-POP, self-generated sub-questions improve performance slightly (e.g., LLaMA-3.1-8B: 34.83% to 37.28%), but they degrade accuracy on MINTQA-TI (12.83% to 9.28%). This contrast reflects the reliance on models' knowledge bases: for known but unpopular facts in MINTQA-POP, decomposition organizes existing knowledge, while on MINTQA-TI, knowledge gaps might lead to flawed decomposition and additional errors.

Providing Sub-Questions: Gold sub-questions significantly boost performance on MINTQA-POP (e.g., LLaMA-3.1-8B sees a 33.41% increase) by clarifying reasoning paths and allowing models to focus on synthesis. On MINTQA-TI, improvements are modest, with the best accuracy (23.75%) still from LLaMA-3.1-8B. This differences can be expected. While decomposition can help models better utilize their existing knowledge, it cannot compensate for the fundamental lack of information when handling questions about new facts.

Model	MINTQA-POP		MINTQA-TI	
	(1)	(2)	(1)	(2)
LLaMA-3.2-1B	19.62	41.89	6.88	14.23
LLaMA-3.2-3B	27.02	59.93	7.42	18.26
LLaMA-3.1-8B	37.28	70.69	9.82	**23.75**
LLaMA-3.1-70B	**54.08**	**72.97**	**16.73**	23.15
Mistral-7B-v0.3	31.41	58.80	9.52	16.00
Ministral-8B-2410	35.76	59.61	10.54	17.44
Mixtral-8x7B-v0.1	45.66	68.90	12.08	20.93
Phi-3-mini	21.41	37.61	5.84	11.46
Phi-3-small	26.98	38.53	8.09	13.52
Phi-3-medium	38.91	64.29	10.18	19.45
Qwen2.5-1.5B	25.74	59.29	9.07	18.33
Qwen2.5-3B	23.86	55.90	6.99	15.87
Qwen2.5-7B	26.32	55.83	8.93	17.85
Qwen2.5-14B	41.34	65.67	12.96	20.01
Qwen2.5-32B	39.16	63.44	12.51	19.79
Qwen2.5-72B	44.99	54.62	14.10	20.51
Gemma-2-2B	25.67	49.97	8.04	14.32
Gemma-2-9B	38.18	59.65	11.17	18.09
Gemma-2-27B	44.54	66.29	12.36	18.90

Table 12: The accuracy (%) of LLMs evaluated under query decomposition settings: (1) the model generates and answers sub-questions itself, and (2) the model answers given gold sub-questions.

H Error Analysis

To better understand the current limitations of the model, we manually analyze 100 incorrectly predicted samples from the Qwen2.5-72B model under the decomposition-then-retrieval (DTR) setting, using examples drawn from the MINTQA-POP and MINTQA-TI datasets. The errors can be cate-

gorized into four major types, with representative examples shown in Table 11:

1. **Poor Question Decomposition:** The model generates subquestions that are syntactically valid but semantically irrelevant or misaligned with the original question. This is the most common error type in both datasets. It accounts for 40.13% of errors in the MINTQA-POP dataset (301 out of 750) and 42% in the MINTQA-TI dataset (21 out of 50).

2. **Poor Subquestion Answering:** This category includes two failure modes: (a) the model generates meaningful subquestions but fails to retrieve or synthesize correct answers, and (b) the model does not generate any subquestions at all. This is the dominant error type in the MINTQA-POP dataset, making up 59.33% of errors (445 out of 750), and is also frequent in the MINTQA-TI dataset at 44% (22 out of 50).

3. **Final Answer Incorrect Despite Good Reasoning:** Although the subquestions and their answers are correct, the model produces a final answer that contradicts the available evidence. This is a rare but important category, responsible for 0.53% of errors in MINTQA-POP (4 out of 750) and 4% in MINTQA-TI (2 out of 50). These cases highlight deficiencies in final synthesis or judgment.

4. **Output Truncation or Formatting Issue:** These are rare cases where the final output is incomplete due to decoding failures or formatting errors. This error type does not appear in the MINTQA-POP dataset but is observed in 2% of the MINTQA-TI dataset (1 out of 50).

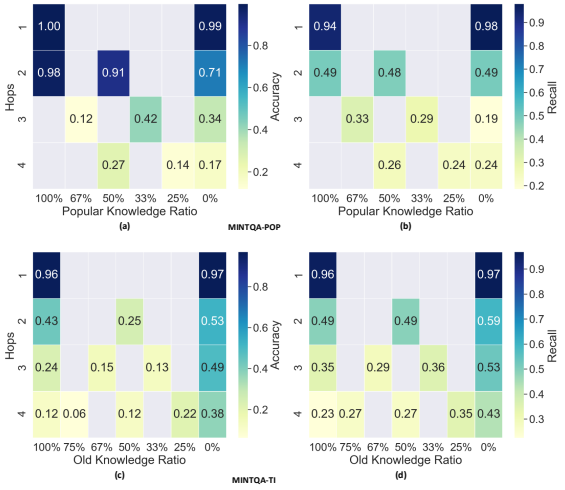


Figure 12: Heatmaps (a) and (c) show Qwen2.5-72B with BM25 performance on two datasets, while heatmaps (b) and (d) shows BM25 recall. The X-axis represents the proportion of popular knowledge required in the question, and the Y-axis indicates question hops.

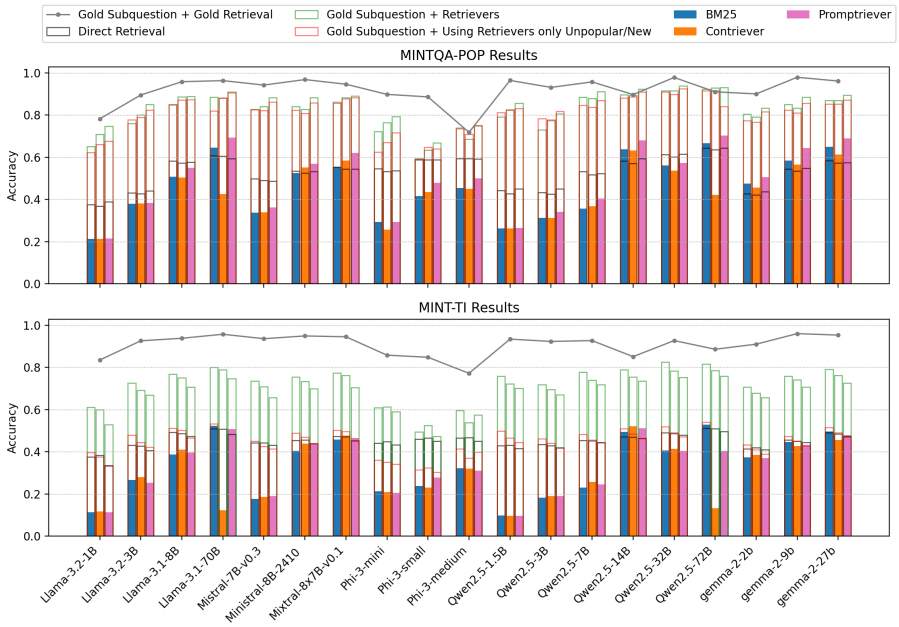


Figure 13: Full datasets evaluation results of performance of all models with three retrievers (i.e. BM25, Contriever and Promptriever) using decomposition-retrieval approach on two datasets. Gold Subquestion + Gold Retriever means that the model uses gold subquestion and gold retrieval results. Direct Retrieval means that the model uses different retrievers for direct retrieval instead of the decomposition-retrieval approach. Gold Subquestion + Retrievers indicates that the model uses gold subquestion and uses different retrievers to retrieve. Gold Subquestion + Using Retrievers only Unpopular/New indicates that the model uses gold subquestion and uses different retrievers to only retrieve the question involving unpopular knowledge or new knowledge, while models rely on their own to direct answer the popular knowledge or old knowledge.

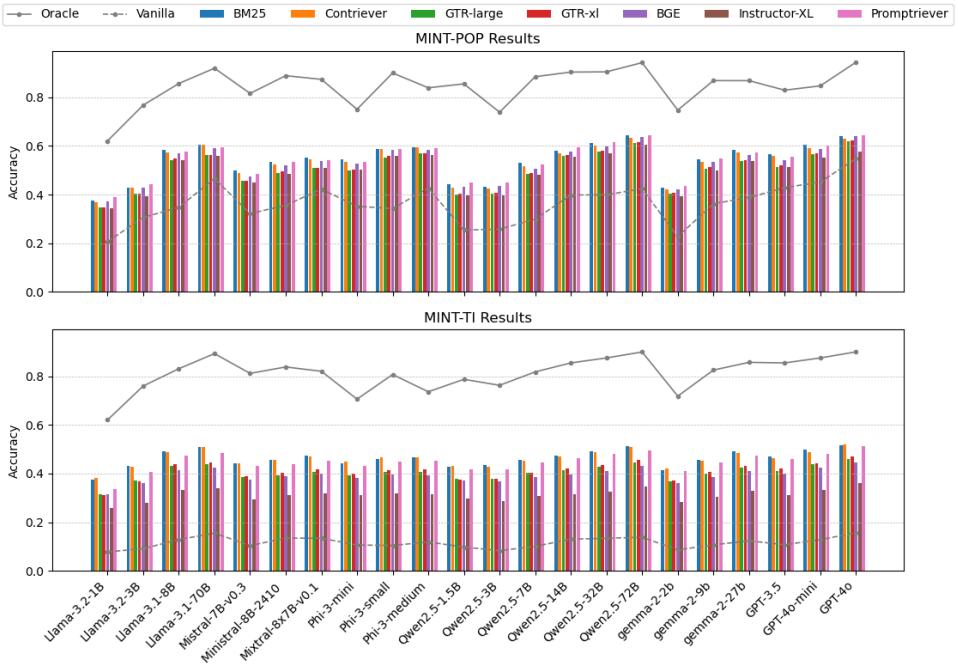


Figure 14: Performance comparison of LLMs on MINTQA-POP and MINTQA-TI using different retrieval methods. "Oracle" uses gold-standard retrieval passages, while "Vanilla"involves models answering without retrieval content.

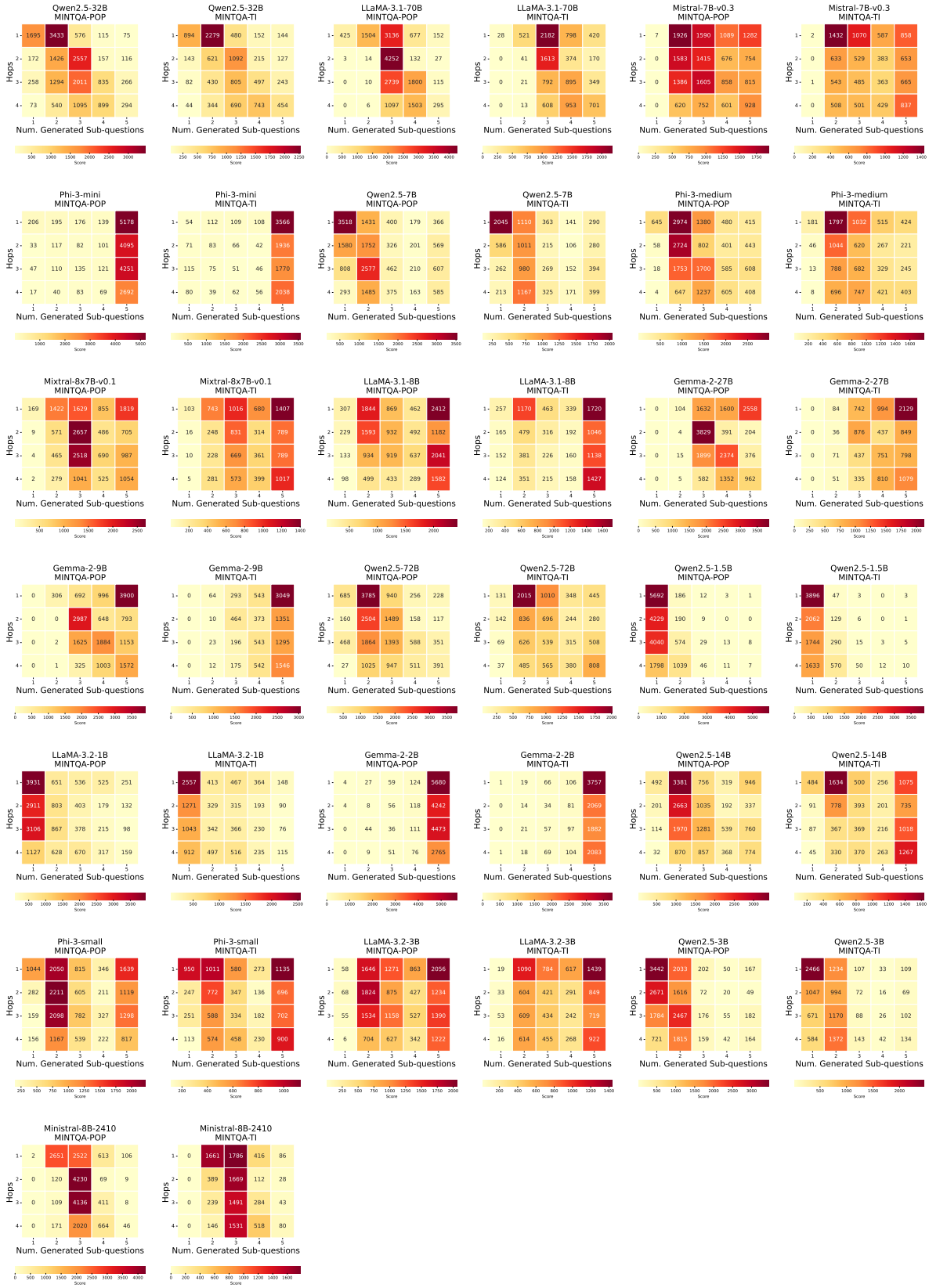


Figure 15: The confusion matrix of the number of sub-questions generated by the LLMs for main questions categorized by hops in the setting of purely generating sub-questions.

Model	BM25			Contriever			PromptRetrieval		
	Acc (%)	Avg. Sub	Avg. Ret	Acc (%)	Avg. Sub	Avg. Ret	Acc (%)	Avg. Sub	Avg. Ret
MINTQA-POP									
Qwen Models									
Qwen2.5-1.5B	25.86 (-0.50)	1.13 (+1.00)	0.32 (+0.19)	26.02 (-0.36)	1.15 (+1.02)	0.31 (+0.18)	26.13 (-0.35)	1.15 (+1.01)	0.32 (+0.18)
Qwen2.5-3B	31.16 (-0.13)	1.78 (+0.95)	1.54 (+0.71)	29.56 (-1.64)	1.70 (+0.89)	1.48 (+0.67)	31.55 (-2.50)	1.69 (+0.89)	1.47 (+0.67)
Qwen2.5-7B	32.98 (-2.68)	2.18 (+0.97)	1.14 (-0.07)	34.89 (-1.90)	2.11 (+0.93)	1.09 (-0.09)	36.58 (-3.93)	2.09 (+0.92)	1.10 (-0.07)
Qwen2.5-14B	53.77 (-10.02)	3.44 (+1.10)	1.22 (-1.12)	53.35 (-9.80)	3.45 (+0.99)	1.23 (-1.23)	55.53 (-12.39)	3.41 (+1.10)	1.21 (-1.10)
Qwen2.5-32B	50.33 (-5.86)	2.79 (+1.02)	1.18 (-0.59)	48.56 (-5.03)	2.77 (+0.92)	1.19 (-0.66)	50.84 (-6.40)	2.81 (+0.91)	1.18 (-0.72)
Qwen2.5-72B	58.63 (-7.93)	3.01 (+1.10)	1.35 (-0.56)	57.42 (+15.28)	3.05 (+3.05)	1.36 (+1.36)	60.89 (-9.33)	3.02 (+1.09)	1.32 (-0.61)
Llama Models									
LLaMA-3.2-1B	20.53 (-0.75)	1.79 (+1.00)	1.36 (+0.57)	20.86 (-0.47)	1.79 (+1.00)	1.32 (+0.53)	21.13 (-0.46)	1.80 (+1.00)	1.32 (+0.52)
LLaMA-3.2-3B	37.23 (-0.64)	3.48 (+0.72)	3.26 (+0.50)	37.70 (-0.53)	3.49 (+0.68)	3.28 (+0.47)	38.04 (-0.27)	3.50 (+0.70)	3.29 (+0.49)
LLaMA-3.1-8B	50.01 (-0.72)	3.88 (+0.56)	3.79 (+0.47)	50.19 (-0.13)	4.05 (+0.49)	3.96 (+0.40)	54.38 (-0.54)	4.01 (+0.56)	3.91 (+0.46)
LLaMA-3.1-70B	64.80 (+0.17)	3.41 (+0.89)	3.40 (+0.88)	62.55 (+19.98)	3.38 (+3.38)	3.37 (+3.37)	69.21 (-0.07)	3.31 (+0.96)	3.30 (+0.95)
Mistral Models									
Mistral-7B-v0.3	29.47 (-4.23)	3.13 (+0.76)	1.84 (-0.53)	28.80 (-5.18)	3.24 (+0.67)	1.77 (-0.80)	30.73 (-5.60)	3.20 (+0.72)	1.84 (-0.64)
Ministral-8B-2410	35.91 (-16.49)	2.95 (+0.96)	0.02 (-1.97)	36.04 (-19.19)	2.94 (+0.95)	0.02 (-1.97)	36.04 (-20.94)	2.94 (+0.97)	0.01 (-1.96)
Mixtral-8x7B-v0.1	48.28 (-7.08)	3.61 (+0.77)	1.29 (-1.55)	48.05 (-10.47)	3.60 (+0.86)	1.30 (-1.44)	49.00 (-13.10)	3.59 (+0.83)	1.23 (-1.53)
Phi Models									
Phi-3-mini	26.81 (-2.54)	4.75 (+0.04)	2.23 (-2.48)	25.64 (-0.12)	4.71 (+0.03)	2.38 (-2.30)	27.63 (-1.79)	4.71 (+0.03)	2.28 (-2.40)
Phi-3-small	37.09 (-4.58)	2.92 (+0.44)	0.78 (-1.70)	37.19 (-6.28)	2.88 (+0.65)	0.79 (-1.44)	39.14 (-8.85)	2.87 (+0.62)	0.77 (-1.48)
Phi-3-medium	40.16 (-5.24)	2.98 (+0.87)	1.08 (-1.03)	39.65 (-5.31)	2.94 (+0.80)	1.07 (-1.07)	41.90 (-8.05)	2.96 (+0.81)	1.07 (-1.08)
Gemma Models									
Gemma-2-2B	34.20 (-13.31)	4.96 (+0.13)	0.99 (-3.84)	34.62 (-11.07)	4.96 (+0.07)	1.00 (-3.89)	35.62 (-14.95)	4.96 (+0.07)	0.99 (-3.90)
Gemma-2-9B	39.99 (-18.52)	3.93 (+0.18)	0.32 (-3.43)	40.20 (-16.35)	3.92 (+0.11)	0.34 (-3.47)	40.56 (-23.73)	3.92 (-0.07)	0.33 (-3.66)
Gemma-2-27B	64.64 (-0.37)	4.05 (+0.74)	4.01 (+0.70)	60.82 (-0.50)	4.09 (+0.71)	4.04 (+0.66)	68.64 (-0.40)	4.32 (+0.57)	4.28 (+0.53)
MINTQA-TI									
Qwen Models									
Qwen2.5-1.5B	9.49 (-0.29)	1.13 (+1.00)	0.35 (+0.22)	9.39 (-0.20)	1.14 (+1.01)	0.36 (+0.23)	9.47 (-0.11)	1.14 (+1.01)	0.36 (+0.23)
Qwen2.5-3B	18.18 (+0.00)	1.83 (+0.95)	1.65 (+0.77)	17.55 (-1.41)	1.78 (+0.91)	1.62 (+0.75)	17.76 (-1.27)	1.75 (+0.89)	1.59 (+0.73)
Qwen2.5-7B	20.36 (-2.60)	2.35 (+0.92)	1.58 (+0.15)	21.83 (-3.93)	2.21 (+0.92)	1.54 (+0.25)	21.34 (-3.26)	2.21 (+0.91)	1.56 (+0.26)
Qwen2.5-14B	39.40 (-10.01)	3.65 (+0.84)	2.01 (-0.80)	41.53 (-10.61)	3.63 (+0.89)	2.00 (-0.74)	40.91 (-10.25)	3.62 (+0.85)	1.98 (-0.79)
Qwen2.5-32B	33.87 (-6.81)	2.86 (+1.02)	1.87 (+0.03)	34.91 (-6.53)	2.87 (+0.98)	1.89 (+0.00)	33.75 (-6.75)	2.89 (+0.97)	1.89 (-0.03)
Qwen2.5-72B	44.79 (-8.13)	3.42 (+0.95)	2.04 (-0.43)	45.99 (+32.78)	3.45 (+3.45)	2.06 (+2.06)	44.44 (-9.64)	3.47 (+0.92)	2.07 (-0.48)
Llama Models									
LLaMA-3.2-1B	9.25 (-2.17)	1.94 (+0.99)	1.41 (+0.46)	9.47 (-2.36)	1.98 (+1.00)	1.34 (+0.36)	8.99 (-2.29)	1.99 (+0.99)	1.36 (+0.36)
LLaMA-3.2-3B	25.89 (-0.69)	3.73 (+0.65)	3.47 (+0.39)	27.39 (-0.68)	3.75 (+0.64)	3.49 (+0.38)	24.82 (-0.57)	3.74 (+0.62)	3.48 (+0.36)
LLaMA-3.1-8B	38.24 (-0.51)	3.99 (+0.49)	3.90 (+0.40)	40.40 (-0.64)	4.10 (+0.46)	4.01 (+0.37)	39.02 (-0.61)	4.11 (+0.48)	4.02 (+0.39)
LLaMA-3.1-70B	51.99 (-0.31)	3.70 (+0.91)	3.68 (+0.89)	52.16 (+39.78)	3.69 (+3.69)	3.67 (+3.67)	50.59 (-0.27)	3.71 (+0.90)	3.69 (+0.88)
Mistral Models									
Mistral-7B-v0.3	9.86 (-7.75)	3.19 (+0.76)	1.72 (-0.71)	9.77 (-8.87)	3.41 (+0.65)	1.76 (-1.00)	10.16 (-8.92)	3.36 (+0.70)	1.81 (-0.85)
Ministral-8B-2410	10.69 (-29.81)	2.99 (+0.89)	0.10 (-2.00)	10.75 (-33.21)	2.99 (+0.88)	0.10 (-2.01)	10.71 (-33.46)	2.98 (+0.90)	0.10 (-1.98)
Mixtral-8x7B-v0.1	24.91 (-20.91)	3.97 (+0.78)	1.33 (-1.86)	26.19 (-21.47)	3.95 (+0.77)	1.32 (-1.86)	25.72 (-20.78)	3.96 (+0.80)	1.29 (-1.87)
Phi Models									
Phi-3-mini	17.38 (-3.95)	4.68 (+0.05)	2.44 (-2.19)	16.94 (-3.91)	4.65 (+0.06)	2.64 (-1.95)	17.21 (-3.43)	4.65 (+0.05)	2.51 (-2.09)
Phi-3-small	15.38 (-8.37)	3.15 (+0.48)	1.23 (-1.44)	15.71 (+15.71)	3.11 (+3.11)	1.22 (+1.22)	16.20 (-11.60)	3.11 (+0.61)	1.21 (-1.29)
Phi-3-medium	19.58 (-12.64)	3.19 (+0.80)	1.35 (-1.04)	19.12 (-12.92)	3.21 (+0.77)	1.34 (-1.10)	19.18 (-11.87)	3.19 (+0.77)	1.33 (-1.09)
Gemma Models									
Gemma-2-2B	24.67 (-12.69)	4.93 (+0.06)	1.78 (-3.09)	25.34 (-13.12)	4.94 (+0.05)	1.82 (-3.07)	24.48 (-12.59)	4.93 (+0.07)	1.80 (-3.06)
Gemma-2-9B	15.29 (-29.38)	4.59 (+0.29)	0.91 (-3.39)	16.08 (-26.57)	4.58 (+0.27)	0.91 (-3.40)	16.50 (-26.89)	4.58 (+0.19)	0.92 (-3.47)
Gemma-2-27B	48.39 (-1.21)	4.49 (+0.51)	4.43 (+0.45)	45.35 (-0.30)	4.53 (+0.47)	4.47 (+0.41)	47.45 (-0.47)	4.53 (+0.48)	4.47 (+0.42)

Table 13: The full results for Decomposition-Dynamic Retrieve. **Acc** represents the model's accuracy (%), **Avg. Sub** is the average number of sub-questions generated, **Avg. Ret** is the average number of sub-questions actually used for retrieval. The difference between **DDR** and **DTR** is shown in brackets (>=0 is green, <0 is red).

You are a powerful multi-hop question generator. Users will provide a chain of Wikidata triplets, and you will help write questions to ask the tail entity from the head entity. The format of a wikidata triple is (subject, relation, object). You shouldn't include bridge entities in generated questions. The questions should only include the head entity. **All involved relations must be reflected in the question.**

#Example 1
Wikidata triplets:(Four Peaks, mountain range, x1), (x1, located in the administrative territorial entity, x2), (x2, located in the administrative territorial entity, x3), (x3, office held by head of government, x4)
Generated question: Who holds the office of the head of government for the administrative entity where the mountain range Four Peaks is located??

#Example 2
Wikidata triplets: (Alena Vostrá, place of birth, x1)
Generated question: Where was Alena Vostrá born?

#Example 3
Wikidata triplets: (Anguilla, country, x1), (x1, capital, x2)
Generated question: what is the capital of the country of the Anguilla?

#Example 4
Wikidata triplets: (Nazko River, mouth of the watercourse, x1), (x1, mouth of the watercourse, x2), ⟨x2, country, x3)
Generated question: In which country does the Nazko River ultimately discharge its waters?

#Example 5
Wikidate triplets: {Sampled facts}
Generated question:

Table 14: The prompt used to generate questions is based on sampled facts. Additionally, we include 4 demonstrations showcasing examples ranging from 1-hop to 4-hop reasoning.

You are a powerful question answering system. Users will provide a question and useful context. The provided context are some wikidata triplets which format is (subject, relation, object). You should answer the question based on the context. The answer should be a single entity or a list of entities. If the answer is a list of entities, you should return the most relevant one.
Context: {related documents}
Question: {question}

Table 15: The prompt used for question quality inspection provides a given question and its corresponding facts. We aim for the GPT-4o to correctly answer the question based on this information.

Below is a question, please answer it directly and keep your answer as short as possible. **Question**: {question } **Answer**:

Table 16: The prompt designed to guide the model in providing a concise answer directly to the question.

Given some related documents: {retrieved_documents}. This is a question: {question}. Please answer the question directly. Please keep your answer as short as possible. **Answer**:

Table 17: The prompt instructs the model to provide a concise answer to the question based on the retrieved documents.

Here is a question: {question}
To answer this question. You have to three choices now:

⟨**choice A**⟩ Generate a sub-question.
⟨**choice B**⟩ Answer the question directly if you are confident to answer it.
⟨**choice C**⟩ retrieve some document to help you answer the question.

If you choose ⟨**choice A**⟩, please output:
{{"choice A": {{"sub-question": "your_sub_question_here"}}}}

If you choose ⟨**choice B**⟩, please output:
{{"choice B": {{"answer": "your_answer_here"}}}}

If you choose ⟨**choice C**⟩, please output:
{{"choice C": retrieval}}

The final output should be in the form of a JSON string, without any additional content. Please keep your answer as short as possible.
Output:

Table 18: The prompt is used for retrieval tasks, directly generating answers or creating sub-questions for judgment purposes.

Given a question: {question}
The subsequent sub-questions: {sub_questions}

You have two choices now:

⟨**choice A**⟩ answer the final sub-question directly.
⟨**choice B**⟩ retrieve some document to help you answer the question. Just output retrieval as a placeholder.
If you choose ⟨**choice A**⟩, please output:
{{"choice A": {{"answer": "your_answer_here"}}}}

If you choose ⟨**choice B**⟩, please output:
{{"choice B": retrieval}}

The final output should be in the form of a JSON string, without any additional content. Please keep your answer as short as possible.
Output:

Table 19: The prompt is used for evaluating sub-questions, performing retrieval, or directly generating answers.

Given a main question: {question}
And sub-question-answer pairs: {sub_question_answer_pairs}

Please judge if the main question has been finished. You have two choices now:

⟨**choice A**⟩ The answer can be found in the sub-question-answer pairs. If you choose this choice, please output the final answer.
⟨**choice B**⟩ The answer cannot be found and a new sub-question needs to be generated.

If you choose ⟨**choice A**⟩, please output:
{{"choice A": {{"answer": "final_answer_here"}}}}

If you choose ⟨**choice B**⟩, please output:
{{"choice B": {{"sub-question": "new_sub-question_here"}}}}

The final output should be in the form of a JSON string, without any additional content. Please keep your answer as short as possible.
Output:

Table 20: The prompt provides sub-questions and their answers, requiring the model to determine whether the answer to the main question has been found.

To answer this question, you may need to generate subquestions following these guidelines:
Given a main question and optional previous subquestion-answer pairs, you may need to generate subquestions to help answer this main question. Please ensure to only generate subquestions that are relevant to answering the main question. When there are no more subquestions needed, output "finish".
Input Format
liRequired:
- Main Question: [question]
Optional:
- Previous Subquestion: [subquestion]
- Previous Answer: [subanswer]
Output Format
One of:
- Next Subquestion: [new subquestion]
- "finish" (when no further subquestions are needed)
Generation Guidelines
1. Subquestions should:
- Break down complex aspects of the main question
- Follow a logical progression
- Be specific and focused
- Build upon previous answers when available
2. Output "finish" when:
- All relevant aspects have been covered
- Further breakdown would not add value
- The question has been fully addressed
Examples
Example 1:
Input:
- Main Question: "What is the location of the headquarters of the institution where Percival Lowell was educated?"
- Previous Subquestion: "Where did Percival Lowell receive his education?"
- Previous Answer: "Harvard University."
Output:
- Next Subquestion: "Where is the headquarters of Harvard University?"
Example 2:
Input:
- Main Question: "What is the capital of France?"
Output:
- "finish"
Main Question: {question}
{previous_subquestion_answer_pairs}
Output:

Table 21: This prompt indicates that the model determines whether to generate sub-questions based on the previous answer history, to further break down the main question, or to finish the continued generation of sub-questions.

Based on the main question and all subquestion-answer pairs, please provide a comprehensive final answer. Please keep your answer as short as possible.
Main Question: {main_question}
Previous Subquestions and Answers:
{history_str}
Final Answer:

Table 22: The prompt instructs the model to summarize and generate the answer to the main question based on the sub-questions and their answers.

Answer the following question based on your internal knowledge with one or few words.
Add a confidence indicator after your answer: - "certain" if you are completely confident in the accuracy - "uncertain" if you have any doubts
Input Format
Input:
- Question: [question]
Output Format
Output:
- Answer: [brief answer]
- Confidence: [certain/uncertain]
Question: {question}
Output:

Table 23: The prompt requires the model to output a confidence score for the generated sub-questions, which will be used to determine whether retrieval is necessary.

Triplets: [[Pigeon Bay Domain, country, New Zealand]] **Main Question:** In which country is Pigeon Bay Domain located? **Main Answer:** New Zealand **Type:** New
Triplets: [[Eveline Hoffmann, place of detention, Theresienstadt Ghetto]] **Main Question:** Where was Eveline Hoffmann detained? **Main Answer:** Theresienstadt Ghetto **Type:** Old

Table 24: One-hop question-answer pairs and their corresponding types in MINTQA-TI.

Triplets: [[Scram Kitty and his Buddy on Rails, publisher, Dakko Dakko], [Dakko Dakko, industry, video game industry]] **Main Question:** In which industry does the publisher of Scram Kitty and his Buddy on Rails operate? **Main Answer:** video game industry **Subquestion pairs:** **Sub-question** 0: Who is the publisher of Scram Kitty and his Buddy on Rails? **Sub-answer** 0: Dakko Dakko. **Type**: New **Sub-question** 1: In which industry does Dakko Dakko operate? **Sub-answer** 1: video game industry. **Type**: New
Triplets: [[CineKink NYC, location, New York City], [New York City, capital of, United States of America]] **Main Question:** CineKink NYC is located in the city that is the capital of which entity? **Main Answer:** United States of America **Subquestion pairs:** **Sub-question** 0: Where is CineKink NYC located? **Sub-answer** 0: New York City. **Type**: New **Sub-question** 1: What entity has New York City as its capital? **Sub-answer** 1: United States of America. **Type**: Old
Triplets: [[Sanna Aunesluoma, residence, Espoo], [Espoo, member of, Union of the Baltic Cities]] **Main Question:** Which organization or group is the residence of Sanna Aunesluoma a member of? **Main Answer:** Union of the Baltic Cities **Subquestion pairs:** **Sub-question** 0: Where does Sanna Aunesluoma reside? **Sub-answer** 0: Espoo. **Type**: Old **Sub-question** 1: Of which entity is Espoo a member? **Sub-answer** 1: Union of the Baltic Cities. **Type**: New
Triplets: [[Horst Hoffmann, country of citizenship, German Democratic Republic], [German Democratic Republic, legislative body, Volkskammer]] **Main Question:** What is the legislative body of the country where Horst Hoffmann holds citizenship? **Main Answer:** Volkskammer **Subquestion pairs:** **Sub-question** 0: What is the country of citizenship of Horst Hoffmann? **Sub-answer** 0: German Democratic Republic. **Type**: Old **Sub-question** 1: What is the legislative body of the German Democratic Republic? **Sub-answer** 1: Volkskammer. **Type**: Old

Table 25: Two-hop question-answer pairs and their corresponding types in MINTQA-TI.

Triplets: [[Systems and methods for mesh augmentation and prevention of incisional hernia, owned by, The Trustees of the University of Pennsylvania], [The Trustees of the University of Pennsylvania, headquarters location, Philadelphia], [Philadelphia, member of, Organization of World Heritage Cities]]
Main Question: Of which entity is the headquarters location of the owner of the "Systems and methods for mesh augmentation and prevention of incisional hernia" a member?
Main Answer: Organization of World Heritage Cities
Subquestion pairs:
Sub-question 0: Who owns the patent for Systems and methods for mesh augmentation and prevention of incisional hernia? **Sub-answer** 0: The Trustees of the University of Pennsylvania. **Type**: New
Sub-question 1: Where is the headquarters of The Trustees of the University of Pennsylvania located? **Sub-answer** 1: Philadelphia. **Type**: New
Sub-question 2: What is Philadelphia a member of? **Sub-answer** 2: Organization of World Heritage Cities. **Type**: New

Triplets: [[De grote Gwen en Geraldine show, nominated for, Dutch Podcast Award for Chatcast Vermaak], [Dutch Podcast Award for Chatcast Vermaak, country, Netherlands], [Netherlands, language used, Dutch]]
Main Question: What is the language used in the country for which "De grote Gwen en Geraldine show" was nominated?
Main Answer: Dutch
Subquestion pairs:
Sub-question 0: For what award was "De grote Gwen en Geraldine show" nominated? **Sub-answer** 0: Dutch Podcast Award for Chatcast Vermaak. **Type**: New
Sub-question 1: In which country is the Dutch Podcast Award for Chatcast Vermaak given? **Sub-answer** 1: Netherlands. **Type**: New
Sub-question 2: What language is used in the Netherlands? **Sub-answer** 2: Dutch. **Type**: Old

Triplets: [[Gathering to Celebrate Old Age, creator, Tomioka Tessai], [Tomioka Tessai, location, Tokyo National Museum], [Tokyo National Museum, member of, Japan Consortium for Open Access Repository]]
Main Question: Which organization or group is the location associated with the creator of "Gathering to Celebrate Old Age" a member of?
Main Answer: Japan Consortium for Open Access Repository
Subquestion pairs:
Sub-question 0: Who is the creator of Gathering to Celebrate Old Age? **Sub-answer** 0: Tomioka Tessai. **Type**: New
Sub-question 1: Where is Tomioka Tessai located? **Sub-answer** 1: Tokyo National Museum. **Type**: Old
Sub-question 2: What organization or association is the Tokyo National Museum a member of? **Sub-answer** 2: Japan Consortium for Open Access Repository. **Type**: New

Triplets: [[The Woman Who Cooked Her Husband, author, Debbie Isitt], [Debbie Isitt, country of citizenship, United Kingdom], [United Kingdom, continent, Europe]]
Main Question: On which continent does the author of "The Woman Who Cooked Her Husband" hold citizenship?
Main Answer: Europe
Subquestion pairs:
Sub-question 0: Who is the author of "The Woman Who Cooked Her Husband"? **Sub-answer** 0: Debbie Isitt. **Type**: New
Sub-question 1: What is the country of citizenship of Debbie Isitt? **Sub-answer** 1: United Kingdom. **Type**: Old
Sub-question 2: On which continent is the United Kingdom located? **Sub-answer** 2: Europe. **Type**: Old

Triplets: [[Mubarak Shah, religion or worldview, Islam], [Islam, item operated, Qalab], [Qalab, cause of death, Ajal]]
Main Question: What was the cause of death for the operator of the religion or worldview followed by Mubarak Shah?
Main Answer: Ajal
Subquestion pairs:
Sub-question 0: What is the religion or worldview of Mubarak Shah? **Sub-answer** 0: Islam. **Type**: Old
Sub-question 1: What item is operated by Islam? **Sub-answer** 1: Qalab. **Type**: New
Sub-question 2: What was the cause of death for Qalab? **Sub-answer** 2: Ajal. **Type**: New

Triplets: [[Felipe Borrego Estrada, place of birth, Zacatecas], [Zacatecas, member of, Organization of World Heritage Cities], [Organization of World Heritage Cities, headquarters location, Quebec City]]
Main Question: Where is the headquarters of the entity that the birthplace of Felipe Borrego Estrada is a member of?
Main Answer: Quebec City
Subquestion pairs:
Sub-question 0: Where was Felipe Borrego Estrada born? **Sub-answer** 0: Zacatecas. **Type**: Old
Sub-question 1: Of which organization is Zacatecas a member? **Sub-answer** 1: Organization of World Heritage Cities. **Type**: New
Sub-question 2: Where is the headquarters of the Organization of World Heritage Cities located? **Sub-answer** 2: Quebec City. **Type**: Old

Triplets: [[Hykjeberget, operator, Dalarna County Administrative Board], [Dalarna County Administrative Board, headquarters location, Falun], [Falun, twinned administrative body, Hamina]]
Main Question: What administrative body is twinned with the location of the headquarters of the operator of Hykjeberget?
Main Answer: Hamina
Subquestion pairs:
Sub-question 0: Who operates Hykjeberget? **Sub-answer** 0: Dalarna County Administrative Board. **Type**: Old
Sub-question 1: Where is the headquarters of the Dalarna County Administrative Board located? **Sub-answer** 1: Falun. **Type**: Old
Sub-question 2: Which administrative body is twinned with Falun? **Sub-answer** 2: Hamina. **Type**: New

Triplets: [[University of California Italian Studies Multicampus Research Group, country, United States of America], [United States of America, highest point, Denali], [Denali, mountain range, Alaska Range]]
Main Question: What is the mountain range that contains the highest point in the country where the University of California Italian Studies Multicampus Research Group is located?
Main Answer: Alaska Range
Subquestion pairs:
Sub-question 0: In which country is the University of California Italian Studies Multicampus Research Group located? **Sub-answer** 0: United States of America. **Type**: Old
Sub-question 1: What is the highest point in the United States of America? **Sub-answer** 1: Denali. **Type**: Old
Sub-question 2: In which mountain range is Denali located? **Sub-answer** 2: Alaska Range. **Type**: Old

Table 26: Three-hop question-answer pairs and their corresponding types in MINTQA-TI.

Triplets: [[Patricia Florence Suthers, sibling, Elaine Suthers], [Elaine Suthers, mother, Elsie Suthers], [Elsie Suthers, country of citizenship, United Kingdom], [United Kingdom, highest point, Ben Nevis]]
Main Question: What is the highest point in the country where the mother of Patricia Florence Suthers' sibling is a citizen?
Main Answer: Ben Nevis
Subquestion pairs:
Sub-question 0: Who is the sibling of Patricia Florence Suthers? **Sub-answer** 0: Elaine Suthers. **Type**: New
Sub-question 1: Who is the mother of Elaine Suthers? **Sub-answer** 1: Elsie Suthers. **Type**: New
Sub-question 2: Which country is Elsie Suthers a citizen of? **Sub-answer** 2: United Kingdom. **Type**: New
Sub-question 3: What is the highest point in the United Kingdom? **Sub-answer** 3: Ben Nevis. **Type**: Old

Triplets: [[Patricia Florence Suthers, mother, Elsie Suthers], [Elsie Suthers, spouse, Robert Suthers], [Robert Suthers, relative, Miriam Farid], [Miriam Farid, country of citizenship, United Kingdom]]
Main Question: What is the country of citizenship of the relative of Patricia Florence Suthers' mother's spouse?
Main Answer: United Kingdom
Subquestion pairs:
Sub-question 0: Who is the mother of Patricia Florence Suthers? **Sub-answer** 0: Elsie Suthers. **Type**: New
Sub-question 1: Who is the spouse of Elsie Suthers? **Sub-answer** 1: Robert Suthers. **Type**: New
Sub-question 2: Who is a relative of Robert Suthers? **Sub-answer** 2: Miriam Farid. **Type**: New
Sub-question 3: Which country is Miriam Farid a citizen of? **Sub-answer** 3: United Kingdom. **Type**: New

Triplets: [[May Hnin Aw Kanya, mother, May Hnin Htapi], [May Hnin Htapi, father, Loethai], [Loethai, child, Lithai], [Lithai, notable work, Traibhumikatha]]
Main Question: What is the notable work of the child of the father of the mother of May Hnin Aw Kanya?
Main Answer: Traibhumikatha
Subquestion pairs:
Sub-question 0: Who is the mother of May Hnin Aw Kanya? **Sub-answer** 0: May Hnin Htapi. **Type**: New
Sub-question 1: Who is May Hnin Htapi's father? **Sub-answer** 1: Loethai. **Type**: New
Sub-question 2: Who is the child of Loethai? **Sub-answer** 2: Lithai. **Type**: Old
Sub-question 3: What is a notable work created by Lithai? **Sub-answer** 3: Traibhumikatha. **Type**: New

Triplets: [[SEOlytics, parent organization, Sistrix], [Sistrix, country, Germany], [Germany, continent, Europe], [Europe, shares border with, Asia]]
Main Question: Which continent shares a border with the continent where the country of SEOlytics' parent organization is located?
Main Answer: Asia
Subquestion pairs:
Sub-question 0: What is the parent organization of SEOlytics? **Sub-answer** 0: Sistrix. **Type**: New
Sub-question 1: In which country is Sistrix located? **Sub-answer** 1: Germany. **Type**: New
Sub-question 2: On which continent is Germany located? **Sub-answer** 2: Europe. **Type**: Old
Sub-question 3: Which continent shares a border with Europe? **Sub-answer** 3: Asia. **Type**: Old

Triplets: [[Sri Dhamasokaraj, relative, Saileuthai], [Saileuthai, father, Lithai], [Lithai, sibling, May Hnin Htapi], [May Hnin Htapi, place of death, Mottama]]
Main Question: Where did the sibling of the father of Sri Dhamasokaraj pass away?
Main Answer: Mottama
Subquestion pairs:
Sub-question 0: Who is a relative of Sri Dhamasokaraj? **Sub-answer** 0: Saileuthai. **Type**: New
Sub-question 1: Who was the father of Saileuthai? **Sub-answer** 1: Lithai. **Type**: Old
Sub-question 2: Who is Lithai's sibling? **Sub-answer** 2: May Hnin Htapi. **Type**: New
Sub-question 3: Where did May Hnin Htapi die? **Sub-answer** 3: Mottama. **Type**: New

Triplets: [[Frank Gailor, educated at, New College], [New College, founded by, William of Wykeham], [William of Wykeham, country of citizenship, Kingdom of England], [Kingdom of England, replaced by, Kingdom of Great Britain]]
Main Question: Which entity replaced the country of citizenship of the founder of the institution where Frank Gailor was educated?
Main Answer: Kingdom of Great Britain
Subquestion pairs:
Sub-question 0: Where was Frank Gailor educated? **Sub-answer** 0: New College. **Type**: New
Sub-question 1: Who founded New College? **Sub-answer** 1: William of Wykeham. **Type**: Old
Sub-question 2: Which country was William of Wykeham a citizen of? **Sub-answer** 2: Kingdom of England. **Type**: New
Sub-question 3: What entity replaced the Kingdom of England? **Sub-answer** 3: Kingdom of Great Britain. **Type**: Old

Triplets: [[The Life You Can Save, author, Peter Singer], [Peter Singer, mother, Cora Singer], [Cora Singer, father, David Ernst Oppenheim], [David Ernst Oppenheim, academic degree, doctorate]]
Main Question: What academic degree does the father of the author of "The Life You Can Save" hold?
Main Answer: doctorate
Subquestion pairs:
Sub-question 0: Who is the author of "The Life You Can Save"? **Sub-answer** 0: Peter Singer. **Type**: Old
Sub-question 1: Who is Peter Singer's mother? **Sub-answer** 1: Cora Singer. **Type**: New
Sub-question 2: Who is the father of Cora Singer? **Sub-answer** 2: David Ernst Oppenheim. **Type**: New
Sub-question 3: What academic degree does David Ernst Oppenheim hold? **Sub-answer** 3: doctorate. **Type**: Old

Triplets: [[Geoffrey Howe, creator, June Mendoza], [June Mendoza, place of birth, Melbourne], [Melbourne, located in or next to body of water, Yarra River], [Yarra River, continent, Australian continent]]
Main Question: On which continent is the body of water located next to the place where the creator Geoffrey Howe was born?
Main Answer: Australian continent
Subquestion pairs:
Sub-question 0: What did Geoffrey Howe create? **Sub-answer** 0: June Mendoza. **Type**: Old
Sub-question 1: Where was June Mendoza born? **Sub-answer** 1: Melbourne. **Type**: New
Sub-question 2: Which body of water is Melbourne located near? **Sub-answer** 2: Yarra River. **Type**: Old
Sub-question 3: On which continent is the Yarra River located? **Sub-answer** 3: Australian continent. **Type**: New

Table 27: Four-hop question-answer pairs and their corresponding types in MINTQA-TI (part 1).

Triplets: [[Descenso a los fascismos, place of publication, Barcelona], [Barcelona, member of, Creative Cities Network], [Creative Cities Network, operator, UNESCO], [UNESCO, operating area, worldwide]]
Main Question: In what area does the operator of the organization that includes the place where "Descenso a los fascismos" was published operate?
Main Answer: worldwide
Subquestion pairs:
Sub-question 0: Where was "Descenso a los fascismos" published? **Sub-answer** 0: Barcelona. **Type**: New
Sub-question 1: What organization or group is Barcelona a member of? **Sub-answer** 1: Creative Cities Network. **Type**: Old
Sub-question 2: Who operates the Creative Cities Network? **Sub-answer** 2: UNESCO. **Type**: Old
Sub-question 3: What is the operating area of UNESCO? **Sub-answer** 3: worldwide. **Type**: New

Triplets: [[Monument to Terenzio Mamiani, commemorates, Terenzio, Count Mamiani della Rovere], [Terenzio, Count Mamiani della Rovere, award received, Order of the Redeemer], [Order of the Redeemer, founded by, Otto of Greece], [Otto of Greece, spouse, Amalia of Oldenburg]]
Main Question: Who is the spouse of the founder of the award received by the person commemorated by the Monument to Terenzio Mamiani?
Main Answer: Amalia of Oldenburg
Subquestion pairs:
Sub-question 0: Who is commemorated by the Monument to Terenzio Mamiani? **Sub-answer** 0: Terenzio, Count Mamiani della Rovere. **Type**: New
Sub-question 1: What award did Terenzio, Count Mamiani della Rovere receive? **Sub-answer** 1: Order of the Redeemer. **Type**: Old
Sub-question 2: Who founded the Order of the Redeemer? **Sub-answer** 2: Otto of Greece. **Type**: Old
Sub-question 3: Who was the spouse of Otto of Greece? **Sub-answer** 3: Amalia of Oldenburg. **Type**: Old

Triplets: [[Tansen, religion or worldview, Islam], [Islam, item operated, Qalab], [Qalab, cause of death, Ajal], [Ajal, location, treasures of God in Islam]]
Main Question: Where did the cause of death of the religious figure associated with Tansen occur?
Main Answer: treasures of God in Islam
Subquestion pairs:
Sub-question 0: What is the religion or worldview associated with Tansen? **Sub-answer** 0: Islam. **Type**: Old
Sub-question 1: What item is operated by Islam? **Sub-answer** 1: Qalab. **Type**: New
Sub-question 2: What was the cause of death for Qalab? **Sub-answer** 2: Ajal. **Type**: New
Sub-question 3: Where is Ajal located? **Sub-answer** 3: treasures of God in Islam. **Type**: New

Triplets: [[Irma Stern, place of birth, Bratislava], [Bratislava, member of, League of Historical Cities], [League of Historical Cities, headquarters location, Kyoto], [Kyoto, highest point, Mount Minako]]
Main Question: What is the highest point of the location where the headquarters of the entity that includes the birthplace of Irma Stern is situated?
Main Answer: Mount Minako
Subquestion pairs:
Sub-question 0: Where was Irma Stern born? **Sub-answer** 0: Bratislava. **Type**: Old
Sub-question 1: Of which organization is Bratislava a member? **Sub-answer** 1: League of Historical Cities. **Type**: New
Sub-question 2: Where is the headquarters of the League of Historical Cities located? **Sub-answer** 2: Kyoto. **Type**: Old
Sub-question 3: What is the highest point in Kyoto? **Sub-answer** 3: Mount Minako. **Type**: Old

Triplets: [[Andrew Cogglesby, present in work, Evan Harrington], [Evan Harrington, author, George Meredith], [George Meredith, spouse, Mary Meredith], [Mary Meredith, cause of death, kidney failure]]
Main Question: What was the cause of death of the spouse of the author who created the work featuring Andrew Cogglesby?
Main Answer: kidney failure
Subquestion pairs:
Sub-question 0: In which work does Andrew Cogglesby appear? **Sub-answer** 0: Evan Harrington. **Type**: Old
Sub-question 1: Who is the author of "Evan Harrington"? **Sub-answer** 1: George Meredith. **Type**: Old
Sub-question 2: Who is the spouse of George Meredith? **Sub-answer** 2: Mary Meredith. **Type**: New
Sub-question 3: What was the cause of death of Mary Meredith? **Sub-answer** 3: kidney failure. **Type**: New

Triplets: [[Federico Cocozza, employer, Curie Institute], [Curie Institute, founded by, Marie Curie], [Marie Curie, ethnic group, Poles], [Poles, language used, Church Slavonic]]
Main Question: What language is used by the ethnic group of the founder of Federico Cocozza's employer?
Main Answer: Church Slavonic
Subquestion pairs:
Sub-question 0: Who employs Federico Cocozza? **Sub-answer** 0: Curie Institute. **Type**: Old
Sub-question 1: Who founded the Curie Institute? **Sub-answer** 1: Marie Curie. **Type**: Old
Sub-question 2: What is the ethnic group of Marie Curie? **Sub-answer** 2: Poles. **Type**: New
Sub-question 3: Which language is used by Poles? **Sub-answer** 3: Church Slavonic. **Type**: Old

Triplets: [[Devespresso Games, headquarters location, Seoul], [Seoul, member of, Creative Cities Network], [Creative Cities Network, operator, UNESCO], [UNESCO, operating area, worldwide]]
Main Question: What is the operating area of the operator of the member organization where Devespresso Games' headquarters is located?
Main Answer: worldwide
Subquestion pairs:
Sub-question 0: Where is the headquarters of Devespresso Games located? **Sub-answer** 0: Seoul. **Type**: Old
Sub-question 1: Of which organization is Seoul a member? **Sub-answer** 1: Creative Cities Network. **Type**: Old
Sub-question 2: Who operates the Creative Cities Network? **Sub-answer** 2: UNESCO. **Type**: Old
Sub-question 3: What is the operating area of UNESCO? **Sub-answer** 3: worldwide. **Type**: New

Triplets: [[Sonetto I, author, Vittorio Alfieri], [Vittorio Alfieri, place of death, Florence], [Florence, present in work, Civilization V], [Civilization V, developer, Firaxis Games]]
Main Question: Who is the developer of the work where the place of death of the author of Sonetto I is present?
Main Answer: Firaxis Games
Subquestion pairs:
Sub-question 0: Who is the author of Sonetto I? **Sub-answer** 0: Vittorio Alfieri. **Type**: Old
Sub-question 1: Where did Vittorio Alfieri die? **Sub-answer** 1: Florence. **Type**: Old
Sub-question 2: In which work is Florence present? **Sub-answer** 2: Civilization V. **Type**: Old
Sub-question 3: Who developed Civilization V? **Sub-answer** 3: Firaxis Games. **Type**: Old

Table 28: Four-hop question-answer pairs and their corresponding types in MINTQA-TI (part 2)..

Triplets: [[Papanasam taluk, country, India]]
Main Question: In which country is Papanasam taluk located?
Main Answer: India
Type: Popular

Triplets: [[Jerod Swallow, sports discipline competed in, ice dance]]
Main Question: In which sports discipline does Jerod Swallow compete?
Main Answer: ice dance
Type: Unpopular

Table 29: One-hop question-answer pairs and their corresponding types in MINTQA-POP.

Triplets: [[Gmina Szypliszki, country, Poland], [Poland, capital, Warsaw]]
Main Question: What is the capital of the country where Gmina Szypliszki is located?
Main Answer: Warsaw
Subquestion pairs:
Sub-question 0: In which country is Gmina Szypliszki located? **Sub-answer** 0: Poland. **Type**: Popular
Sub-question 1: What is the capital of Poland? **Sub-answer** 1: Warsaw. **Type**: Popular

Triplets: [[Canary Islands, country, Spain], [Spain, legislative body, Cortes Generales]]
Main Question: What is the legislative body of the country to which the Canary Islands belong?
Main Answer: Cortes Generales
Subquestion pairs:
Sub-question 0: Which country are the Canary Islands part of? **Sub-answer** 0: Spain. **Type**: Popular
Sub-question 1: What is the legislative body of Spain? **Sub-answer** 1: Cortes Generales. **Type**: Unpopular

Triplets: [[Pabna Cadet College, country, Bangladesh], [Bangladesh, capital, Dhaka]]
Main Question: What is the capital of the country where Pabna Cadet College is located?
Main Answer: Dhaka
Subquestion pairs:
Sub-question 0: In which country is Pabna Cadet College located? **Sub-answer** 0: Bangladesh. **Type**: Unpopular
Sub-question 1: What is the capital of Bangladesh? **Sub-answer** 1: Dhaka. **Type**: Popular

Triplets: [[Brackendale Eagles Provincial Park, country, Canada], [Canada, highest point, Mount Logan]]
Main Question: What is the highest point in the country where Brackendale Eagles Provincial Park is located?
Main Answer: Mount Logan
Subquestion pairs:
Sub-question 0: In which country is Brackendale Eagles Provincial Park located? **Sub-answer** 0: Canada. **Type**: Unpopular
Sub-question 1: What is the highest point in Canada? **Sub-answer** 1: Mount Logan. **Type**: Unpopular

Table 30: Two-hop question-answer pairs and their corresponding types in MINTQA-POP.

Triplets: [[Cuzco Department, country, Peru], [Peru, capital, Lima], [Lima, located in or next to body of water, Rímac River]]
Main Question: Which body of water is located in or next to the capital of the country where the Cuzco Department is found?
Main Answer: Rímac River
Subquestion pairs:
Sub-question 0: In which country is the Cuzco Department located? **Sub-answer** 0: Peru. **Type**: Popular
Sub-question 1: What is the capital of Peru? **Sub-answer** 1: Lima. **Type**: Popular
Sub-question 2: Which body of water is Lima located next to? **Sub-answer** 2: Rímac River. **Type**: Unpopular

Triplets: [[Kirkovo Municipality, country, Bulgaria], [Bulgaria, highest point, Musala], [Musala, mountain range, Rila]]
Main Question: Which mountain range includes the highest point in the country of Kirkovo Municipality?
Main Answer: Rila
Subquestion pairs:
Sub-question 0: Which country is Kirkovo Municipality located in? **Sub-answer** 0: Bulgaria. **Type**: Popular
Sub-question 1: What is the highest point in Bulgaria? **Sub-answer** 1: Musala. **Type**: Unpopular
Sub-question 2: In which mountain range is Musala located? **Sub-answer** 2: Rila. **Type**: Unpopular

Triplets: [[Nicu Stroia, participant in, 1992 Summer Olympics], [1992 Summer Olympics, country, Spain], [Spain, capital, Madrid]]
Main Question: What is the capital of the country where Nicu Stroia participated in an event?
Main Answer: Madrid
Subquestion pairs:
Sub-question 0: In which events or activities did Nicu Stroia participate? **Sub-answer** 0: 1992 Summer Olympics. **Type**: Unpopular
Sub-question 1: In which country were the 1992 Summer Olympics held? **Sub-answer** 1: Spain. **Type**: Popular
Sub-question 2: What is the capital of Spain? **Sub-answer** 2: Madrid. **Type**: Popular

Triplets: [[Bunk Moreland, present in work, The Wire], [The Wire, original broadcaster, HBO], [HBO, parent organization, WarnerMedia]]
Main Question: What is the parent organization of the original broadcaster of the work featuring Bunk Moreland?
Main Answer: WarnerMedia
Subquestion pairs:
Sub-question 0: In which work does the character Bunk Moreland appear? **Sub-answer** 0: The Wire. **Type**: Unpopular
Sub-question 1: What is the original broadcaster of The Wire? **Sub-answer** 1: HBO. **Type**: Popular
Sub-question 2: What is the parent organization of HBO? **Sub-answer** 2: WarnerMedia. **Type**: Unpopular

Triplets: [[Ewout van Asbeck, sport, field hockey], [field hockey, country of origin, England], [England, capital, London]]
Main Question: What is the capital of the country of origin of the sport in which Ewout van Asbeck participates?
Main Answer: London
Subquestion pairs:
Sub-question 0: What sport does Ewout van Asbeck participate in? **Sub-answer** 0: field hockey. **Type**: Unpopular
Sub-question 1: Which country is the origin of field hockey? **Sub-answer** 1: England. **Type**: Unpopular
Sub-question 2: What is the capital of England? **Sub-answer** 2: London. **Type**: Popular

Triplets: [[College Hockey in the D, sport, ice hockey], [ice hockey, authority, International Ice Hockey Federation], [International Ice Hockey Federation, headquarters location, Zürich]]
Main Question: Where is the headquarters of the authority governing the sport of College Hockey in the D located?
Main Answer: Zürich
Subquestion pairs:
Sub-question 0: What sport is associated with College Hockey in the D? **Sub-answer** 0: ice hockey. **Type**: Unpopular
Sub-question 1: Which organization is the governing authority for ice hockey? **Sub-answer** 1: International Ice Hockey Federation. **Type**: Unpopular
Sub-question 2: Where are the headquarters of the International Ice Hockey Federation located? **Sub-answer** 2: Zürich. **Type**: Unpopular

Table 31: Three-hop question-answer pairs and their corresponding types in MINTQA-POP.

Triplets: [[National Hockey League, sport, ice hockey], [ice hockey, authority, International Ice Hockey Federation], [International Ice Hockey Federation, country, Switzerland], [Switzerland, continent, Europe]]
Main Question: On which continent is the country that has authority over the sport played in the National Hockey League located?
Main Answer: Europe
Subquestion pairs:
Sub-question 0: What sport is played in the National Hockey League? **Sub-answer** 0: ice hockey. **Type**: Popular
Sub-question 1: Which organization is the governing authority for ice hockey? **Sub-answer** 1: International Ice Hockey Federation. **Type**: Unpopular
Sub-question 2: Which country is the International Ice Hockey Federation based in? **Sub-answer** 2: Switzerland. **Type**: Unpopular
Sub-question 3: On which continent is Switzerland located? **Sub-answer** 3: Europe. **Type**: Unpopular

Triplets: [[Rafael Bejarano, place of birth, Arequipa], [Arequipa, country, Peru], [Peru, capital, Lima], [Lima, located in or next to body of water, Rímac River]]
Main Question: Which body of water is the capital of the country where Rafael Bejarano was born located next to?
Main Answer: Rímac River
Subquestion pairs:
Sub-question 0: Where was Rafael Bejarano born? **Sub-answer** 0: Arequipa. **Type**: Unpopular
Sub-question 1: In which country is Arequipa located? **Sub-answer** 1: Peru. **Type**: Popular
Sub-question 2: What is the capital of Peru? **Sub-answer** 2: Lima. **Type**: Popular
Sub-question 3: Which body of water is Lima located next to? **Sub-answer** 3: Rímac River. **Type**: Unpopular

Triplets: [[The Perfect Cocktail, part of the series, How I Met Your Mother], [How I Met Your Mother, original broadcaster, CBS], [CBS, owned by, Paramount Global], [Paramount Global, industry, mass media]]
Main Question: In which industry does the owner of the original broadcaster of the series that includes "The Perfect Cocktail" operate?
Main Answer: mass media
Subquestion pairs:
Sub-question 0: Of which series is "The Perfect Cocktail" a part? **Sub-answer** 0: How I Met Your Mother. **Type**: Unpopular
Sub-question 1: Which network originally broadcasted "How I Met Your Mother"? **Sub-answer** 1: CBS. **Type**: Popular
Sub-question 2: Who owns CBS? **Sub-answer** 2: Paramount Global. **Type**: Unpopular
Sub-question 3: In which industry does Paramount Global operate? **Sub-answer** 3: mass media. **Type**: Unpopular

Triplets: [[Saint George Killing the Dragon, creator, Bernat Martorell], [Bernat Martorell, place of death, Barcelona], [Barcelona, country, Spain], [Spain, capital, Madrid]]
Main Question: What is the capital of the country where the creator of Saint George Killing the Dragon died?
Main Answer: Madrid
Subquestion pairs:
Sub-question 0: Who is the creator of Saint George Killing the Dragon? **Sub-answer** 0: Bernat Martorell. **Type**: Unpopular
Sub-question 1: Where did Bernat Martorell die? **Sub-answer** 1: Barcelona. **Type**: Unpopular
Sub-question 2: In which country is Barcelona located? **Sub-answer** 2: Spain. **Type**: Popular
Sub-question 3: What is the capital of Spain? **Sub-answer** 3: Madrid. **Type**: Popular

Triplets: [[DWNX-FM, owned by, Radio Mindanao Network], [Radio Mindanao Network, headquarters location, Makati], [Makati, country, Philippines], [Philippines, continent, Asia]]
Main Question: On which continent is the country located where the headquarters of the owner of DWNX-FM is situated?
Main Answer: Asia
Subquestion pairs:
Sub-question 0: Who owns DWNX-FM? **Sub-answer** 0: Radio Mindanao Network. **Type**: Unpopular
Sub-question 1: Where is the headquarters of Radio Mindanao Network located? **Sub-answer** 1: Makati. **Type**: Unpopular
Sub-question 2: In which country is Makati located? **Sub-answer** 2: Philippines. **Type**: Popular
Sub-question 3: On which continent is the Philippines located? **Sub-answer** 3: Asia. **Type**: Unpopular

Triplets: [[2008 FA Trophy Final, location, Wembley Stadium], [Wembley Stadium, owned by, The Football Association], [The Football Association, applies to jurisdiction, England], [England, capital, London]]
Main Question: What is the capital of the jurisdiction that owns the location of the 2008 FA Trophy Final?
Main Answer: London
Subquestion pairs:
Sub-question 0: Where was the 2008 FA Trophy Final held? **Sub-answer** 0: Wembley Stadium. **Type**: Unpopular
Sub-question 1: Who owns Wembley Stadium? **Sub-answer** 1: The Football Association. **Type**: Unpopular
Sub-question 2: Which jurisdiction does The Football Association apply to? **Sub-answer** 2: England. **Type**: Unpopular
Sub-question 3: What is the capital of England? **Sub-answer** 3: London. **Type**: Popular

Triplets: [[Rothschild banking family of France, founded by, James Mayer de Rothschild], [James Mayer de Rothschild, place of birth, Frankfurt], [Frankfurt, located in or next to body of water, Main], [Main, mouth of the watercourse, Rhine]]
Main Question: Into which body of water does the river located next to the birthplace of the founder of the Rothschild banking family of France ultimately flow?
Main Answer: Rhine
Subquestion pairs:
Sub-question 0: Who founded the Rothschild banking family of France? **Sub-answer** 0: James Mayer de Rothschild. **Type**: Unpopular
Sub-question 1: Where was James Mayer de Rothschild born? **Sub-answer** 1: Frankfurt. **Type**: Unpopular
Sub-question 2: Which body of water is Frankfurt located next to? **Sub-answer** 2: Main. **Type**: Unpopular
Sub-question 3: Into which watercourse does the Main River flow? **Sub-answer** 3: Rhine. **Type**: Unpopular

Table 32: Four-hop question-answer pairs and their corresponding types in MINTQA-POP.