AgenticRAGTracer: A Hop-Aware Benchmark for Diagnosing Multi-Step Retrieval Reasoning in Agentic RAG

Qijie You[1*], **Wenkai Yu**[2*], **Hao Liang**[2,3*†], **Zhen Hao Wong**[2], **Wentao Zhang**[2,3,4‡]
[1]University of Science and Technology Beijing
[2]Peking University
[3]Zhongguancun Academy
[4]Beijing Key Laboratory of Data Intelligence and Security (Peking University)
u202342615@xs.ustb.edu.cn, ywk2022@stu.pku.edu.cn, hao.liang@stu.pku.edu.cn
zhenhao1141@stu.pku.edu.cn, wentao.zhang@pku.edu.cn

Abstract

With the rapid advancement of agent-based methods in recent years, Agentic RAG has undoubtedly become an important research direction. Multi-hop reasoning, which requires models to engage in deliberate thinking and multi-step interaction, serves as a critical testbed for assessing such capabilities. However, existing benchmarks typically provide only final questions and answers, while lacking the intermediate hop-level questions that gradually connect atomic questions to the final multi-hop query. This limitation prevents researchers from analyzing at which step an agent fails and restricts more fine-grained evaluation of model capabilities. Moreover, most current benchmarks are manually constructed, which is both time-consuming and labor-intensive, while also limiting scalability and generalization. To address these challenges, we introduce AgenticRAG-Tracer , the first Agentic RAG benchmark that is primarily constructed automatically by large language models and designed to support step-by-step validation. Our benchmark spans multiple domains, contains 1,305 data points, and has no overlap with existing mainstream benchmarks. Extensive experiments demonstrate that even the best large language models perform poorly on our dataset. For instance, GPT-5 attains merely 22.6% EM accuracy on the hardest portion of our dataset. Hop-aware diagnosis reveals that failures are primarily driven by distorted reasoning chains—either collapsing prematurely or wandering into over-extension. This highlights a critical inability to allocate steps consistent with the task's logical structure, providing a diagnostic dimension missing in traditional evaluations. Our code and data are available at `https://github.com/YqjMartin/AgenticRAGTracer`.

*Equal contribution.
†Project leader.
‡Corresponding author.

1 Introduction

With the rapid evolution of the agentic paradigm, Large Language Models (LLMs) have expanded beyond simple text generation to master complex workflows and broad tool usage. Among these capabilities, Retrieval-Augmented Generation (RAG) is pivotal, enabling LLMs to access information beyond their parametric memory, thereby enhancing generalization and mitigating hallucinations. Consequently, advancing the capabilities of Agentic RAG—systems that autonomously plan, retrieve, and reason—has become a central research focus.

Compared to single-hop queries (e.g., NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017)), multi-hop questions serve as a rigorous testbed for Agentic RAG by demanding iterative reasoning and multi-document retrieval. However, current evaluations predominantly rely on benchmarks designed for traditional RAG, such as HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022). These datasets generally provide only final QA pairs with short supporting passages, assuming a retrieve-then-read paradigm where all content is processed at once. For Agentic RAG, which inherently relies on dynamic, multi-step interactions, evaluating performance solely on the final outcome fails to capture the intricate decision-making process. Through a rigorous analysis of existing benchmarks (see Appendix B), we identify four critical limitations that hinder the effective diagnosis of Agentic RAG systems:

Authenticity of Multi-hop Reasoning. The complexity labels in existing datasets are often inflated. We find that questions labeled as multi-hop can frequently be answered directly by the model's internal knowledge or require fewer reasoning steps than claimed. This misalignment compromises the validity of reasoning evaluations.

Faithfulness to RAG Settings. While the logical

Findings of the Association for Computational Linguistics: ACL 2026, pages 1301–1335
July 2-7, 2026 ©2026 Association for Computational Linguistics

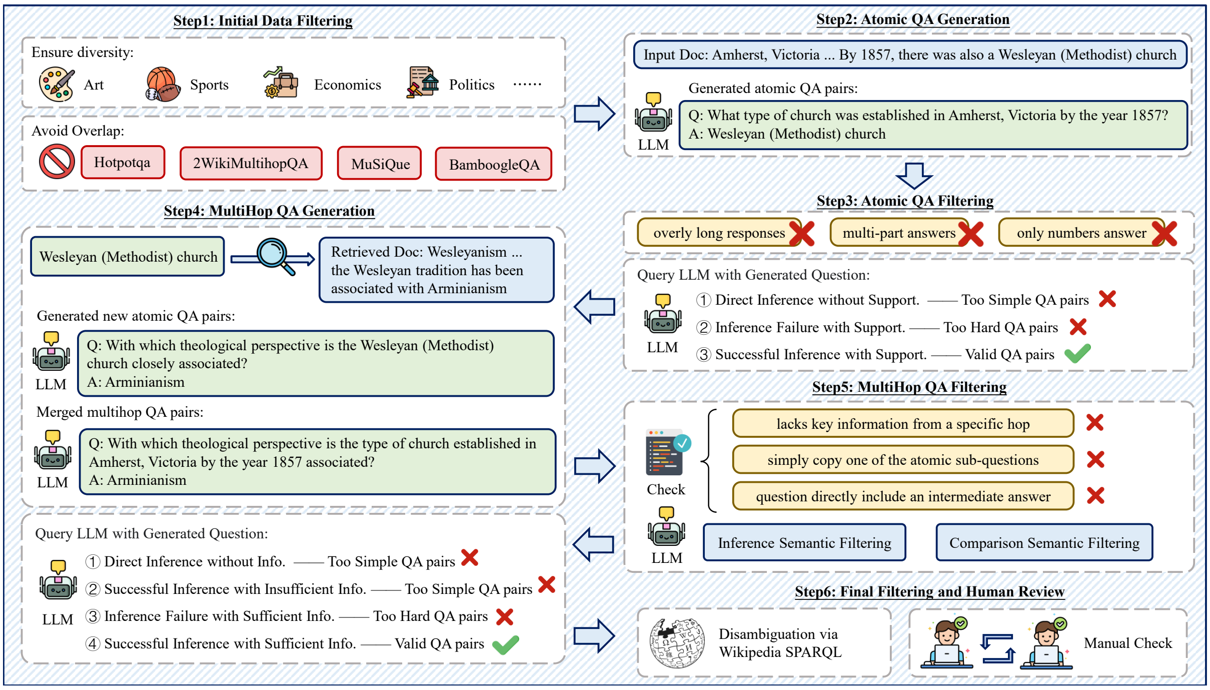

Figure 1: Data construction workflow.

connections between hops may appear plausible to humans, the provided supporting documents often fail to supply explicit links. in such cases, the model is forced to implicitly rely on human prior knowledge rather than actual retrievable evidence, violating the fundamental premise of RAG systems. **Neglect of Intermediate Reasoning.** Most benchmarks provide only the final supporting documents (or sentences) without the complete reasoning trajectory. For modern agents equipped with retrieval tools, errors often occur at specific intermediate steps (e.g., query formulation or document selection). Current benchmarks fail to capture which step fails, treating the agent as a black box. **Lack of Knowledge Bases and Indices.** Existing benchmarks typically do not release the specific corpus and retrieval index used during construction. This omission not only makes evaluation cumbersome but also severely hinders reproducibility, as retrieval performance is highly sensitive to the underlying index construction.

To address these limitations, we propose AgenticRAGTracer , a Hop-Aware benchmark specifically designed for diagnosing multi-step retrieval reasoning in Agentic RAG. We design an automated pipeline to generate high-quality questions with clear topological structures: Inference (Sequential) and Comparison (Parallel). By simulating the interactive workflow of agents, we iteratively combine atomic questions and apply rigorous filtering with final human verification. Evaluation results on mainstream LLMs reveal significant challenges in multi-hop reasoning, highlighting the need for this diagnostic tool.

Our main contributions are summarized as follows:

- We propose AgenticRAGTracer , a benchmark consisting of 1,305 examples, all of which are rigorously human-verified to ensure logical integrity. To the best of our knowledge, it is the first benchmark to enable hop-aware analysis, providing the granular evidence necessary to diagnose model performance at each specific reasoning and retrieval step.

- We design an automatic pipeline for constructing high-quality multi-hop Agentic RAG data, featuring a multi-stage verification protocol that blends automated logic filtering with human-in-the-loop adjudication. This framework enables the scalable and convenient generation of diverse, high-fidelity datasets spanning multiple domains.

- Through a comprehensive evaluation of 13 mainstream models, we demonstrate that existing LLMs—even state-of-the-art systems—exhibit severe degradation on complex multi-hop agentic reasoning. Leveraging hop-aware analysis, our benchmark uncovers how errors accumulate across reasoning steps, specifically how the mismatch between reasoning steps and task structure leads to collapsed or over-extended trajectories that diverge from the required logical progression, highlighting limitations that are not captured by conventional end-task evaluations.

2 Related Work

2.1 Agentic RAG

Retrieval-Augmented Generation (RAG) was originally introduced to combine sequence generation with retrieval, allowing models to leverage external knowledge (Lewis et al., 2020). Subsequent research developed Advanced (Gao et al., 2024a), Modular (Gao et al., 2024b), and Graph RAG (Peng et al., 2024) systems, which improved performance but largely operated as static pipelines. Recently, the field has shifted towards **Agentic RAG**, which integrates autonomous agent capabilities—such as planning, tool use, and reflection—into the retrieval framework. Recent studies formalize these systems, highlighting how agents can dynamically decide what to retrieve and orchestrate multi-step interactions (Singh et al., 2025; Liang et al., 2025b). Parallel efforts explore reinforcement learning to optimize these reasoning-search trajectories (Jin et al., 2025; Chen et al., 2025), transforming RAG from a static process into an adaptive, goal-oriented workflow. For more advanced data-centric agenticrag approaches, please refer to Deng et al. (2025).

2.2 Benchmarks for Multi-hop Retrieval

With the rapid development of Agentic RAG in recent years, there is a growing need for benchmarks that can effectively evaluate its performance. Multi-hop questions are particularly suitable for this purpose, as they require logical decomposition and the use of multiple documents, making them a natural testbed for assessing agent capabilities. Currently, most evaluations still rely on earlier benchmarks created through manual synthesis, such as HotpotQA, 2WikiMultihopQA, MuSiQue, and BamboogleQA. However, the construction of these datasets is labor-intensive, as it requires manually consulting large numbers of documents, and they also suffer from several inherent limitations.

More recently, benchmarks supported by LLM-assisted question generation have been proposed, including MoreHopQA (Schnitzler et al., 2024), MultiHop-RAG (Tang and Yang, 2024), and MINTQA (He et al., 2025). Yet, these approaches also face shortcomings: MoreHopQA simply extends the hop length of existing benchmarks, while MultiHop-RAG and MINTQA rely on directly prompting LLMs to generate questions from raw Wikipedia documents, without ensuring fine-grained quality control.

2.3 Data Synthesis

Recently, data synthesis has emerged as an important technique for improving the performance of large language models (LLMs) (Liang et al., 2026; Bai et al., 2024; An et al., 2025; Guo et al., 2025; Lin et al., 2025; An et al., 2024; Luo et al., 2024). Prior work has extensively explored data synthesis for both textual and multimodal domains. In the text domain, LLM-driven data synthesis pipelines are typically constructed using complex, workflow-based systems such as DataFlow (Liang et al., 2025a; Cai et al., 2025; Shen et al., 2025; Zheng et al., 2024; Liang et al., 2024), enabling high-quality synthetic data generation and achieving strong performance across a wide range of downstream tasks. In the multimodal domain, data synthesis has also proven effective. For example, prior studies synthesize large-scale image caption datasets (Liu et al., 2024) or multimodal verification trajectories (Sun et al., 2025) to enhance the training and reasoning capabilities of vision-language models.

3 Data Construction

3.1 Initial Data Filtering

As shown in Figure 1, we first randomly sampled a collection of documents from the Wikipedia dump[1]. To avoid overlap with existing benchmarks (including HotpotQA, 2Wiki, and MuSiQue), we compared their titles against those provided in the existing datasets and removed duplicates, we also conducted an analysis of similarity in F. We then employed an LLM [2] to annotate their types. This step prevents an overrepresentation of any single question type and ensures diversity within our benchmark.

[1]Given that numerous recent Agentic RAG training efforts have adopted the index and corpus released by FlashRAG, we follow the same collection to construct our benchmark, `https://www.modelscope.cn/datasets/hhjinjiajie/FlashRAG_Dataset` thereby ensuring a fair and comparable evaluation of model performance.

[2]by default GPT-4o-mini, unless otherwise specified in the paper

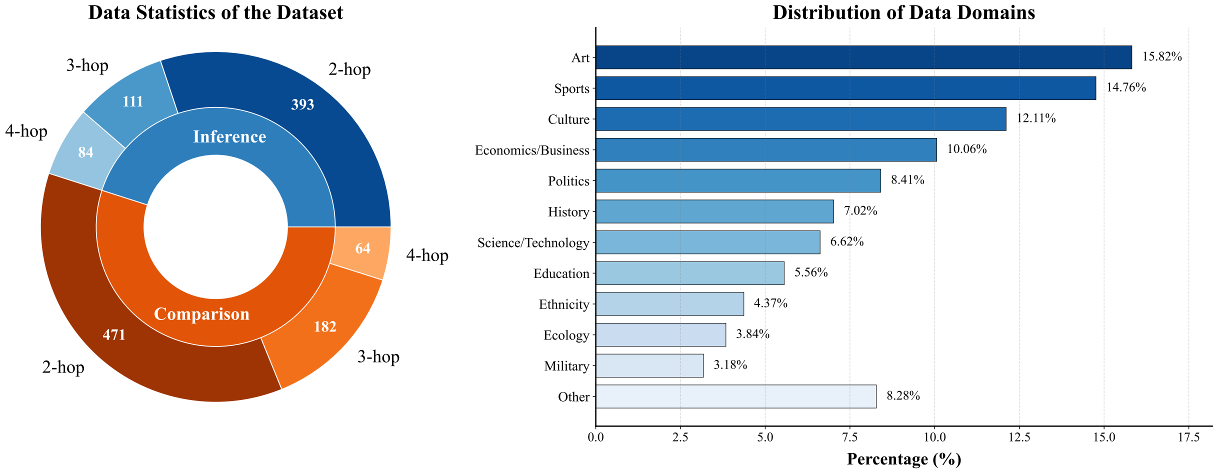

Figure 2: Data Statistics

3.2 Atomic Question Generation

We synthesized atomic QA pairs from the filtered Wikipedia documents using an LLM, followed by a rigorous quality assurance pipeline. The prompts used for data construction are provided in Appendix C. First, we applied heuristic filtering to remove malformed responses; crucially, we excluded answers consisting solely of numbers to preserve semantic context, thereby preventing relevance drift in subsequent retrieval hops. Next, to ensure the necessity of retrieval, we discarded questions that the LLM could correctly answer using only its parametric knowledge. Finally, we verified grounding solvability by re-prompting the LLM with the source document, retaining only those questions where the provided context was sufficient to derive the correct answer. We provide additional details and reliability analysis of the LLM-based judge in Appendix G.

3.3 Design of Multi-hop Questions

To comprehensively evaluate agentic reasoning, we design two distinct question topologies: **Sequential (Inference)**, which requires a stepwise logical chain where intermediate answers bridge to the next step; and **Parallel (Comparison)**, which necessitates gathering independent information about multiple entities before synthesizing a final answer.

Construction of 2-hop Questions. We synthesize 2-hop queries by retrieving additional documents relevant to our atomic QA pairs, explicitly excluding documents present in existing benchmarks to prevent leakage. An LLM is then prompted to combine original and retrieved atomic pairs into complex queries, dynamically selecting between inference or comparison templates. Each candidate question undergoes a rigorous three-stage verification protocol:

1) Structural Integrity Filtering. We discard candidates with obvious syntactic or logical flaws. Crucially, we filter out questions exhibiting *information leakage* (where intermediate answers are explicitly revealed in the question text) and those resulting from *trivial concatenation* (where sub-questions are simply joined without semantic integration).

2) Semantic Logic Verification. An LLM auditor is employed to reject incoherent reasoning chains. For *inference questions*, we eliminate spurious links, such as forcefully connecting unrelated entities or conflating distinct entities with similar names. For *comparison questions*, we discard invalid comparisons between dissimilar concepts or those lacking a clear comparative basis.

3) Multi-hop Necessity & Solvability Check. To guarantee the validity of the reasoning chain, we enforce three criteria: (i) *Retrieval Necessity*: The question must not be answerable by the model's parametric memory alone; (ii) *Dependency Irreducibility*: Given the full document set, removing *any single* supporting document must render the question unanswerable; (iii) *Grounded Solvability*: The model must successfully derive the correct answer when provided with the complete evidence set.

Extension to 3–4 Hop Questions. To challenge the limits of sequential and parallel planning, we iteratively extend verified 2-hop questions by appending additional inference steps. These higher-order questions undergo the same stringent filtering pipeline described above to ensure they remain

Table 1: Results of Closed-Sourced and Open-Sourced LLMs on our AgenticRAGTracer.

Models	Comparison									Inference								
	2hop			3hop			4hop			2hop			3hop			4hop		
	EM	F1	LLM	EM	F1	LLM	EM	F1	LLM	EM	F1	LLM	EM	F1	LLM	EM	F1	LLM
Closed-Sourced LLMs																		
o4-mini	77.7	85.1	86.0	67.6	74.0	79.1	**35.9**	**47.7**	**54.7**	47.8	57.2	66.2	**41.4**	**45.4**	**48.7**	21.4	31.8	34.5
GPT-5	76.2	85.4	88.1	**69.8**	**76.0**	81.3	26.6	39.3	45.3	**48.4**	**59.0**	**70.0**	33.3	40.3	42.3	**22.6**	**34.0**	**39.3**
GPT-4o	69.2	76.3	75.2	51.7	58.5	59.9	15.6	23.8	20.3	24.9	33.0	35.1	13.5	18.4	17.1	4.8	10.0	7.1
Grok-4	**80.7**	**87.0**	**89.0**	66.5	73.8	**81.9**	32.8	46.9	53.1	41.7	51.9	61.6	34.2	40.4	40.5	20.2	28.9	31.0
Grok-3	75.6	84.3	83.4	63.7	70.8	75.8	23.4	37.7	39.1	34.6	43.4	50.4	19.8	24.4	24.3	4.8	10.8	10.7
Open-Sourced LLMs																		
DeepSeek-R1	74.1	82.0	81.7	43.4	52.7	59.9	15.6	24.3	32.8	27.7	37.7	43.3	16.2	21.8	20.7	4.8	10.7	7.1
DeepSeek-V3	66.0	72.1	72.2	37.4	41.7	44.0	15.6	23.7	23.4	22.9	30.1	33.3	10.8	15.3	15.3	2.4	7.5	6.0
Qwen2.5-72B	78.8	85.0	84.3	62.1	73.2	**81.9**	21.9	34.1	35.9	37.7	46.8	53.4	18.0	25.1	26.1	8.3	14.3	11.9
Qwen2.5-32B	70.7	77.5	76.7	50.6	60.3	64.8	18.8	26.8	26.6	34.6	42.4	48.4	11.7	14.5	16.2	6.0	9.1	9.5
Qwen2.5-7B	52.4	65.1	70.5	36.8	44.0	47.3	7.8	16.1	9.4	24.7	32.8	35.1	7.2	13.1	13.5	2.4	8.2	4.8
R-Search	72.2	80.9	78.3	39.0	47.7	51.7	14.1	22.3	15.6	26.0	36.0	41.5	13.5	23.8	20.7	7.1	13.0	9.5
Search-R1	64.5	74.5	65.4	44.5	53.9	48.9	15.6	25.7	12.5	25.5	34.8	37.4	11.7	16.4	14.4	6.0	9.5	7.1
ReSearch	60.3	69.9	67.7	47.3	56.6	55.0	12.5	25.7	21.9	28.2	37.5	41.2	17.1	20.3	20.7	4.8	9.2	7.1

natural and logically solvable, yielding the final benchmark.

3.4 Human Review and Quality Control

In the final stage, we apply Wikipedia SPARQL queries to remove potentially ambiguous cases where multiple entities share the same name in Wikipedia. Following this automated disambiguation, to strictly guarantee data quality, we conducted a full-scale manual verification of the entire dataset.

Specifically, each generated data were independently evaluated by three trained human annotators. The evaluation was formulated as a binary decision (Retain or Discard) driven by two primary dimensions: (1) general quality checks to ensure strict factuality, faithfulness to the source text, and grammatical fluency; and (2) type-specific checks to validate the validity of intermediate reasoning steps and the dimensional consistency for comparison questions.

To assess the reliability of the human evaluation, we calculated Fleiss' Kappa (κ), achieving a score of 0.65, which indicates substantial inter-annotator agreement. For non-unanimous cases, the final labels were rigorously determined through a consensus discussion involving both the annotators and the authors. Further details are provided in Appendix E.

3.5 Data statistics

As shown in Figure 2, our benchmark exhibits both semantic diversity and structural complexity. The domain distribution (right) spans over 11 categories ranging from Art to Military, with no single domain dominating more than 16% of the dataset, ensuring a balanced evaluation. The question topology (left) features a balanced composition of Inference and Comparison types, facilitating a comprehensive assessment of both sequential and parallel reasoning capabilities. Moreover, the dataset spans a difficulty spectrum from fundamental 2-hop queries to complex 3- and 4-hop challenges. This progressive design allows for a fine-grained analysis, stepwise unveiling the capability boundaries of agentic models. Due to space constraints, representative examples of our benchmark are provided in Appendix A to illustrate its structure and complexity.

4 Experiments

4.1 Setup

We evaluate a diverse set of LLMs on our benchmark, including closed-source models (GPT and Grok series), open-source models (DeepSeek (DeepSeek-AI et al., 2025b,a) and Qwen (Qwen et al., 2025) series), as well as checkpoints of recent Agentic RAG models trained from Qwen2.5-7B-Instruct (Search-R1 (Jin et al., 2025), ReSearch (Chen et al., 2025), and R-Search (Zhao et al., 2025)). To ensure a fair comparison, all models are evaluated under a unified experimental setting using the Qwen-Agent framework, which follows the ReAct (Yao et al., 2023) paradigm to interleave reasoning and actions. We report Exact Match (EM), F1 score, and LLM-as-a-Judge scores, com-

Table 2: Reasoning depth and average steps on AgenticRAGTracer. MaxD denotes the maximum reasoning depth reached by each model. C-Steps and I-Steps denote the average number of steps on correctly and incorrectly answered samples, respectively.

Models	Comparison						Inference					
	3hop			4hop			3hop			4hop		
	MaxD	Steps-C	Steps-I	MaxD	Steps-C	Steps-I	MaxD	Steps-C	Steps-I	MaxD	Steps-C	Steps-I
Closed-Sourced LLMs												
o4-mini	2.62	3.02	5.87	3.23	4.34	6.03	1.77	3.02	5.87	2.11	4.72	7.78
GPT-5	2.70	3.13	5.85	3.20	4.10	7.03	1.79	3.14	5.85	2.40	4.48	8.25
GPT-4o	2.43	3.15	3.38	2.81	4.08	4.18	0.96	3.16	3.38	1.02	3.83	4.92
Grok-4	2.70	3.19	3.76	3.16	4.65	5.57	1.65	3.19	3.76	2.00	4.69	6.95
Grok-3	2.53	3.01	3.00	3.00	4.44	4.62	1.09	3.01	3.00	0.94	4.44	5.47
Open-Sourced LLMs												
DeepSeek-R1	2.38	3.04	3.26	2.94	4.10	3.28	1.04	3.05	3.26	0.93	4.17	2.76
DeepSeek-V3	2.07	3.08	2.81	2.53	4.07	3.49	0.86	3.09	2.81	0.83	4.20	3.61
Qwen2.5-72B	2.76	3.28	3.54	3.03	4.30	4.41	1.16	3.28	3.55	0.94	4.50	7.03
Qwen2.5-32B	2.49	3.08	2.83	2.46	4.18	3.51	1.03	3.08	2.83	0.83	4.13	2.96
Qwen2.5-7B	2.20	2.95	3.04	2.52	4.50	3.41	0.93	3.01	3.04	0.89	5.00	3.66
R-Search	2.32	3.08	2.97	2.78	4.20	3.65	0.96	3.09	2.97	0.93	4.13	3.57
Search-R1	2.28	3.03	2.83	2.27	4.38	3.88	1.03	3.03	2.83	0.90	4.17	4.50
ReSearch	2.26	3.26	2.99	2.58	4.21	3.88	1.09	3.26	2.99	0.85	4.00	4.06

bining objective metrics with model-based evaluation. Regarding the LLM-as-a-Judge scores, we employ GPT-4o-mini with the temperature set to 0 to ensure deterministic and reproducible results. To verify the rationality and stability of this protocol, we cross-checked the judge's verdicts on a random sample (20 instances per subset) against stronger models—specifically GPT-5 and Grok-4, as well as human annotations. The results demonstrate near-perfect agreement across these evaluators, confirming that the LLM judge provides a reliable and unbiased measure of model performance. The detailed prompts used for the evaluation are provided in the Appendix C.

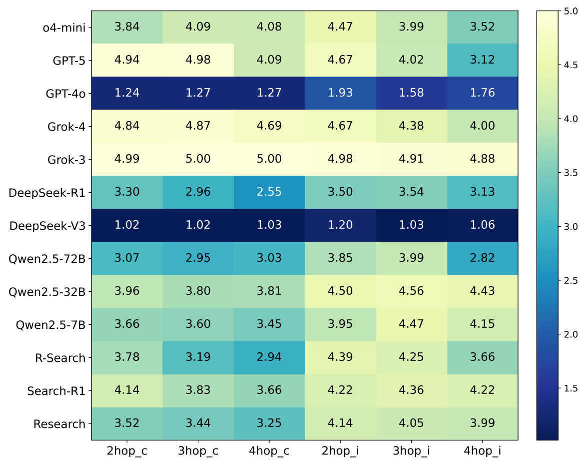

Figure 3: Average top-k values.

4.2 Main Results

The primary results of our evaluation on the proposed benchmark are detailed in Table 1. We summarize the main findings below:

Overall, the benchmark poses a challenge to LLM's multi-hop reasoning capability. Even GPT-5 only has a 22.6% EM score on the 4-hop inference benchmark.

Across all models, performance declines as the number of reasoning hops increases from two to four across both Comparison and Inference subsets; however, the magnitude of this degradation varies substantially across models, revealing different levels of robustness to long-chain reasoning. Similarly, while all models perform worse on the Inference subset, the resulting performance gap differs markedly, indicating that the benchmark effectively differentiates models in their ability to handle serial reasoning and dependency tracking.

Closed-source models such as GPT-5, o4-mini, and Grok-4 generally outperform most open-source counterparts, demonstrating their stronger integrated reasoning and retrieval coordination abilities. However, an exception is GPT-4o, which performs unexpectedly poorly on our benchmark. To investigate the underlying causes of this performance discrepancy, we conducted a quantitative analysis of the retrieval behavior across different models, specifically examining the retrieval volume per step.

As illustrated in Figure 3, there is a distinct correlation between the retrieval strategy and overall performance. Top-tier models like Grok-4 and GPT-5 consistently employ larger top-k values, reflecting a strategy of broad evidence integration. In the context of agentic RAG, this "information-seeking" behavior is crucial, as missing a single piece of evidence in the early hops of a multi-step chain can irreversibly derail the entire reasoning process.

In stark contrast, GPT-4o adopts a highly conservative retrieval strategy, consistently retrieving far fewer documents per step—a behavior pattern notably similar to DeepSeek-V3. This restrictive approach creates an information bottleneck, resulting in limited information accumulation that fails to satisfy the grounded solvability requirements of complex multi-hop queries. Consequently, even with strong internal reasoning capabilities, the model fails because it operates on an incomplete context. This finding not only explains GPT-4o's degradation but also highlights a key diagnostic capability of our benchmark: it effectively exposes suboptimal retrieval strategies (e.g., over-confidence or premature stopping) that limit multi-hop reasoning effectiveness, which traditional single-hop benchmarks may overlook.

Within open-source model families, performance scales positively with model size: Qwen2.5-72B-Instruct consistently outperforms its 32B and 7B variants, suggesting that larger capacity improves multistep reasoning stability. And Qwen2.5-72B-Instruct outperforms the DeepSeek series, indicating a stronger tool-calling and information retrieval capability. Moreover, checkpoints derived from Qwen2.5-7B-Instruct substantially outperform the base model; in some cases, they approach larger open-source models, though a gap to closed-source models remains.

4.3 Multi-Hop Analysis

To better understand model performance on multi-step reasoning, we examine the maximal reasoning depth that models can handle when failing on higher-hop questions. For each 3- and 4-hop question, we record the average of the highest corresponding lower-hop questions that the model answers correctly (Table 2). This metric captures the depth of reasoning that a model can achieve, even when it fails on longer chains. This analysis serves two purposes. First, it highlights the limitations of the current models in maintaining coherent, iterative reasoning: frequent failures on high-hop questions, despite solving the initial hops, indicate weaknesses in multi-step integration. Second, it validates our benchmark design: by quantifying performance degradation across hop levels, AgenticRAGTracer demonstrates its ability to stress agentic reasoning beyond end-task accuracy. Instead of reporting only overall accuracy, this cross-hop evaluation provides a layered, interpretable view of reasoning behavior, revealing that a model may succeed in 2-hop reasoning but struggle with 3- or 4-hop chains, which is unattainable with traditional benchmarks.

We also analyze the relationship between the reasoning step count and answer correctness on the 3- and 4-hop subsets (Table 2), which represents the challenging and representative setting of our benchmark. Across all models, correct responses exhibit step counts closely aligned with the target hop number, indicating structurally consistent reasoning trajectories. In contrast, incorrect responses show significantly more diverse behavior, with both over-extended and prematurely terminated reasoning chains. This gap reveals that success in multihop RAG depends not on longer reasoning but on allocating an appropriate number of steps consistent with task structure, which is a distinction enabled by our benchmark but obscured in prior evaluations.

4.4 Error Case Study

To gain a deeper understanding of model failures, we conducted a systematic analysis of the agentic reasoning trajectories. Our findings reveal that the vast majority of failures are not isolated errors in tool usage but rather cascading failures fundamentally rooted in a breakdown of initial task decomposition (Figure 4). As illustrated in the figure, the LLM incorrectly placed the specific token *'sixth career game-winner'* into the first reasoning step; subsequently, it introduced a spurious subgoal—*'the professional football team he first faced'*—which lacked a logical basis in the original query. This initial analytical misstep propagated a chain of errors throughout the reasoning trajectory, ultimately rendering the final inference completely incorrect.

The core bottleneck lies in the model's inability to correctly partition a complex, multi-hop query into a valid logical sequence. A single analytical misstep at this stage immediately diverts the agent onto an erroneous reasoning trajectory. Once the agent enters such a misaligned path, its entire subsequent interactive process becomes fundamentally unproductive; the agent effectively "loses its way,"

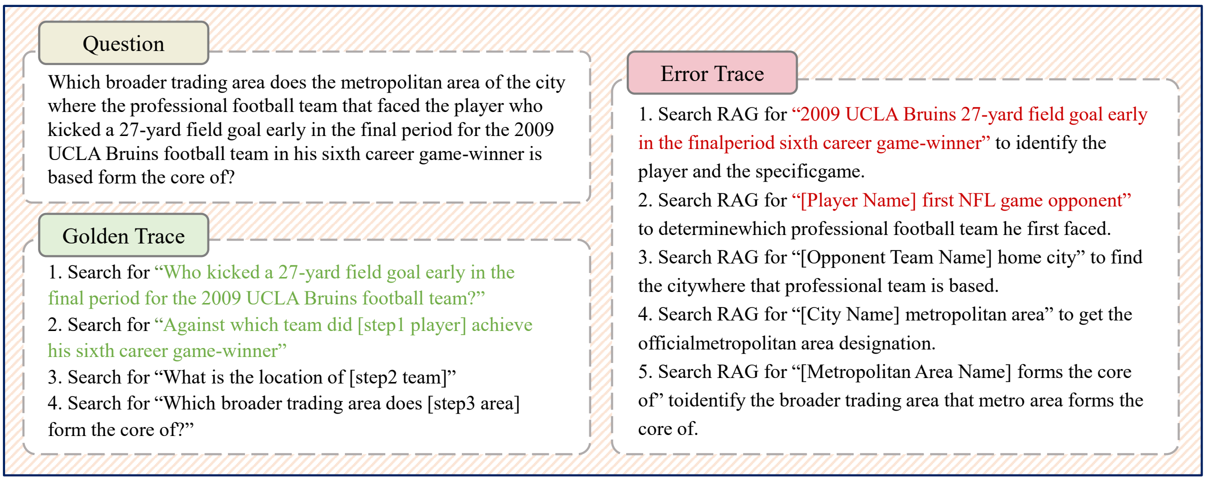

Figure 4: Error case study. We present the key information in the figure; additional details can be found in Appendix D.

pursuing a flawed objective where even successful retrieval steps fail to contribute to the correct answer. In this state of trajectory deviation, the model's efforts—such as issuing redundant queries or misinterpreting retrieved context—are merely inevitable byproducts of its commitment to a faulty plan. Ultimately, these observations underscore that the reliability of Agentic RAG is primarily governed by the robustness of its high-level logical planning, where the initial hop determines the viability of the entire reasoning chain.

4.5 Discussion

The empirical findings from AgenticRAGTracer reveal a fundamental gap between "following a planning prompt" and "possessing true reasoning agency." Although our evaluation pipeline mandates an initial planning stage, the high failure rates on complex tasks—coupled with the cascading errors illustrated in Figure 3—highlight a pervasive rigidity in plan execution. Most models treat their initial decomposition not as a flexible working hypothesis, but as an immutable script. When a model encounters a dead end or retrieves information that contradicts its initial assumptions, it lacks the autonomous meta-cognition to pause and re-evaluate its trajectory. This reasoning inertia is quantitatively evidenced in Table 2, where the reasoning steps for failed tasks (Steps-I) significantly exceed those for successful ones (Steps-C). This "flailing" behavior—characterized by redundant queries and over-extension—suggests that current agents cannot autonomously distinguish between a path that requires more evidence and one that has entered a logical abyss. True autonomous agency in RAG requires more than just better planning; it demands a dynamic self-audit mechanism capable of proactive course correction and informed termination. Our benchmark serves as a crucial diagnostic tool in this regard, proving that the bottleneck in multi-hop reasoning is not merely the lack of information, but the inability to strategically and autonomously manage the reasoning process itself.

5 Conclusion

In this paper, we introduce AgenticRAGTracer, the first hop-aware benchmark for evaluating Agentic RAG systems, together with an automated pipeline for constructing high-quality multi-hop questions. By iteratively synthesizing atomic facts into complex Inference and Comparison topologies, our pipeline ensures that each question possesses a clear, traceable reasoning structure backed by rigorous logical verification. Extensive experiments reveal that model failures are predominantly characterized by distorted reasoning chains, either collapsing prematurely or drifting into over-extension, reflecting an inability to allocate reasoning steps in alignment with the logical structure of the task. AgenticRAGTracer contributes a novel diagnostic perspective by shifting the focus toward step-level transparency. A core strength of our work lies in the high logical fidelity of our benchmark; by combining a controlled automated pipeline with rigorous human verification, we ensure a reliable and traceable dataset that goes beyond simple end-to-end metrics. We hope this resource provides a useful foundation for the community to identify subtle failure modes and encourages the development of more resilient, self-correcting Agentic RAG systems.

Limitations

We discuss the following limitations:

Dataset Scale Owing to the intrinsic complexity of multi-hop reasoning (not every document lends itself to multi-hop question generation) and consideration of resources, we deliberately refrain from significantly expanding the benchmark size.

Limited Types of Evaluated LLMs Due to regional and computational resource constraints, we were unable to evaluate some model series. Nonetheless, the current set of models is sufficient to highlight and reflect the key issues addressed in this study.

Ethical Considerations

Our work focuses on developing an automatically constructed benchmark for evaluating multi-hop reasoning in Agentic RAG systems. All data used in this study were derived from publicly available Wikipedia content and processed automatically, ensuring that no private or sensitive information was included. Manual verification was limited to quality assurance and did not involve annotators' personal data or subjective labeling beyond factual correctness checks. While our benchmark aims to advance transparent evaluation of reasoning processes, we acknowledge that automated data generation could potentially introduce biases from the underlying language models. We encourage future work to examine such biases and ensure fairness in subsequent model evaluations.

Acknowledgements

This work is supported by National Natural Science Foundation of China (92470121, 62402016), Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM113), National Key R&D Program of China (2024YFA1014003), Zhongguancun Academy (C20250204, C20250602), Beijing Major Science and Technology Project (Z251100008125043, Z251100008425023), and High-performance Computing Platform of Peking University.

References

Ruichuan An, Sihan Yang, Ming Lu, Renrui Zhang, Kai Zeng, Yulin Luo, Jiajun Cao, Hao Liang, Ying Chen, Qi She, et al. 2024. Mc-llava: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*.

Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, et al. 2025. Unictokens: Boosting personalized understanding and generation via unified concept tokens. *arXiv preprint arXiv:2505.14671*.

Tianyi Bai, Hao Liang, Binwang Wan, Yanran Xu, Xi Li, Shiyu Li, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, et al. 2024. A survey of multimodal large language model from a data-centric perspective. *arXiv preprint arXiv:2405.16640*.

Qifeng Cai, Hao Liang, Chang Xu, Tao Xie, Wentao Zhang, and Bin Cui. 2025. Text2sql-flow: A robust sql-aware data augmentation framework for text-to-sql. *arXiv preprint arXiv:2511.10192*.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025. Research: Learning to reason with search for llms via reinforcement learning. *Preprint*, arXiv:2503.19470.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jingwen Deng, Jihao Huang, Zhen Hao Wong, Hao Liang, Quanqing Xu, Bin Cui, and Wentao Zhang. 2025. Data-centric perspectives on agentic retrieval-augmented generation: A survey. *Authorea Preprints*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024a. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Yunfan Gao, Yun Xiong, Meng Wang, and Haofen Wang. 2024b. Modular rag: Transforming rag systems into lego-like reconfigurable frameworks. *Preprint*, arXiv:2407.21059.

Ziyu Guo, Xinyan Chen, Renrui Zhang, Ruichuan An, Yu Qi, Dongzhi Jiang, Xiangtai Li, Manyuan Zhang, Hongsheng Li, and Pheng-Ann Heng. 2025. Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. *arXiv preprint arXiv:2510.26802*.

Jie He, Nan Hu, Wanqiu Long, Jiaoyan Chen, and Jeff Z. Pan. 2025. Mintqa: A multi-hop question answering benchmark for evaluating llms on new and tail knowledge. *Preprint*, arXiv:2412.17032.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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 Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Hao Liang, Xiaochen Ma, Zhou Liu, Zhen Hao Wong, Zhengyang Zhao, Zimo Meng, Runming He, Chengyu Shen, Qifeng Cai, Zhaoyang Han, et al. 2025a. Dataflow: An llm-driven framework for unified data preparation and workflow automation in the era of data-centric ai. *arXiv preprint arXiv:2512.16676*.

Hao Liang, Linzhuang Sun, Jingxuan Wei, Xijie Huang, Linkun Sun, Bihui Yu, Conghui He, and Wentao Zhang. 2024. Synth-empathy: Towards high-quality synthetic empathy data. *arXiv preprint arXiv:2407.21669*.

Hao Liang, Zhen Hao Wong, Ruitong Liu, Yuhan Wang, Meiyi Qiang, Zhengyang Zhao, Chengyu Shen, Conghui He, Wentao Zhang, and Bin Cui. 2026. Data preparation for large language models. *Journal of Computer Science and Technology*.

Jintao Liang, Gang Su, Huifeng Lin, You Wu, Rui Zhao, and Ziyue Li. 2025b. Reasoning rag via system 1 or system 2: A survey on reasoning agentic retrieval-augmented generation for industry challenges. *Preprint*, arXiv:2506.10408.

Weifeng Lin, Xinyu Wei, Ruichuan An, Tianhe Ren, Tingwei Chen, Renrui Zhang, Ziyu Guo, Wentao Zhang, Lei Zhang, and Hongsheng Li. 2025. Perceive anything: Recognize, explain, caption, and segment anything in images and videos. *Preprint*, arXiv:2506.05302.

Zheng Liu, Hao Liang, Xijie Huang, Wentao Xiong, Qinhan Yu, Linzhuang Sun, Chong Chen, Conghui He, Bin Cui, and Wentao Zhang. 2024. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *arXiv preprint arXiv:2407.20756*.

Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. 2024. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *European Conference on Computer Vision*, pages 235–252. Springer.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. *Preprint*, arXiv:2408.08921.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Julian Schnitzler, Xanh Ho, Jiahao Huang, Florian Boudin, Saku Sugawara, and Akiko Aizawa. 2024. Morehopqa: More than multi-hop reasoning. *Preprint*, arXiv:2406.13397.

Chengyu Shen, Zhen Hao Wong, Runming He, Hao Liang, Meiyi Qiang, Zimo Meng, Zhengyang Zhao, Bohan Zeng, Zhengzhou Zhu, Bin Cui, et al. 2025. Let's verify math questions step by step. *arXiv preprint arXiv:2505.13903*.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *Preprint*, arXiv:2501.09136.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. 2025. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. *arXiv preprint arXiv:2502.13383*.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Qingfei Zhao, Ruobing Wang, Dingling Xu, Daren Zha, and Limin Liu. 2025. R-search: Empowering llm reasoning with search via multi-reward reinforcement learning. *Preprint*, arXiv:2506.04185.

Miao Zheng, Hao Liang, Fan Yang, Haoze Sun, Tianpeng Li, Lingchu Xiong, Yan Zhang, Youzhen Wu, Kun Li, Yanjun Shen, et al. 2024. Pas: Data-efficient plug-and-play prompt augmentation system. *arXiv preprint arXiv:2407.06027*.

A Data Example

Table 3: Data Example. To make it look clearer, We only present the key information from the supporting documents that is directly used in question generation.

Inference	2-hop	1-hop Question: What type of church was established in Amherst, Victoria by the year 1857? 1-hop Answer: Wesleyan (Methodist) church 1-hop Doc: Amherst, Victoria ... By 1857, there was also a Wesleyan (Methodist) church ... New Atomic Question: With which theological perspective is the Wesleyan (Methodist) church closely associated? New Atomic Answer: Arminianism New Doc: Wesleyanism ... the Wesleyan tradition has been associated with Arminianism ... 2-hop Question: With which theological perspective is the type of church established in Amherst, Victoria by the year 1857 associated? 2-hop Answer: Arminianism
	3-hop	New Atomic Question: Who is the Dutch Reformed theologian associated with the concept of Arminianism? New Atomic Answer: Jacobus Arminius New Doc: Arminianism ... is based on theological ideas of the Dutch Reformed theologian Jacobus Arminius ... 3-hop Question: Who is the Dutch Reformed theologian associated with the theological perspective of the church established in Amherst, Victoria by the year 1857? 3-hop Answer: Jacobus Arminius
	4-hop	New Atomic Question: Where did Jacobus Arminius teach his sermons on the Epistle of the Romans? New Atomic Answer: Amsterdam New Doc: Jacobus Arminius ... At Amsterdam, Arminius taught through "a number of sermons on the Epistle of the Romans." ... 4-hop Question: Where did the Dutch Reformed theologian associated with the theological perspective of the church established in Amherst, Victoria by the year 1857 teach his sermons on the Epistle of the Romans? 4-hop Answer: Amsterdam
Comparison	2-hop	1-hop Question: What is the effective temperature of Omega Persei? 1-hop Answer: 4,586 K 1-hop Doc: Omega Persei ... at an effective temperature of 4,586 K ... New Atomic Question: What is the effective temperature of the stellar atmosphere of HD 195564? New Atomic Answer: 5,421 K New Doc: HD 195564 ... The effective temperature of the stellar atmosphere is 5,421 K ... 2-hop Question: Which star has a higher effective temperature, Omega Persei or HD 195564? 2-hop Answer: HD 195564
	3-hop	New Atomic Question: What spacecraft provided parallax measurements for HD 195564? New Atomic Answer: Hipparcos New Doc: HD 195564 ... Parallax measurements from the Hipparcos spacecraft give us an estimate of its distance ... 3-hop Question: What spacecraft provided parallax measurements for the star that has a higher effective temperature, Omega Persei or HD 195564? 3-hop Answer: Hipparcos
	4-hop	New Atomic Question: How many entries does the Hipparcos Catalogue contain? New Atomic Answer: 118,218 entries New Doc: Hipparcos ... contains 118,218 entries (stars or multiple stars) ... 4-hop Question: How many entries does the catalogue produced by the spacecraft that provided parallax measurements for the star with a higher effective temperature between Omega Persei and HD 195564 contain? 4-hop Answer: 118,218 entries

B Case Studies of Limitations in Prior Work

The following examples illustrate the issues discussed above regarding prior work.

Multi-hop Questions with Incorrect Step Counts:

User:
You are a world expert at making efficient plans to solve any task using an RAG Search tool. ...
Here is your task: What is the name of this White House Cheif of Staff, born in 1950, who assumed the role at the same time as Elaine Duke became Secretary of Homeland Security?

Assistant:
1. Search "Elaine Duke Secretary of Homeland Security start date
2. Search "White House Chief of Staff assumed office on [date from step 1]
3. Search "[Name of Chief of Staff from step 2] born 1950

User:
Based on the plan and the search results before (if there is), first analyse what information you have gained and what other information you still need, then EXECUTE ONLY ONE MORE step using the RAG search tool. ...

Assistant:
```
"function_call": {"name": "RAG_search",
"arguments": "{query: Elaine Duke Secretary of
Homeland Security start date, topk: 3}"}
```

[RAGTool Results]
Query: "Elaine Duke Secretary of Homeland Security start date"
Topk = 3
Document:
"Elaine Duke" Elaine Duke Elaine Costanzo Duke (born June 26, 1958) is an American civil servant and former United States Deputy Secretary of Homeland Security, serving from April 10, 2017 until April 15, 2018. She became acting Secretary of Homeland Security on July 31, 2017, when John F. Kelly assumed the office of White House Chief of Staff. She left the acting position on December 6, 2017, upon the confirmation of Kirstjen Nielsen. The daughter of Francesco (Frank) Costanzo Duke and Concetta Scherma, Duke is a native of Ohio. Her maternal grandparents (Scherma) emigrated from Aidone, Sicily and her Costanzo family roots

User:
Based on the plan and the search results before (if there is), first analyse what information you have gained and what other information you still need, then EXECUTE ONLY ONE MORE step using the RAG search tool. ...

Assistant:
Final_Answer: John F. Kelly

Table 4: Authenticity Of multi-hop reasoning case

The following presents the QA content from prior work (Hotpotqa):

Question: What is the name of this White House Cheif of Staff, born in 1950, who assumed the role at the same time as Elaine Duke became Secretary of Homeland Security?

Answer: John Francis Kelly

Supporting_Facts:
title: Elaine Duke
Elaine Costanzo Duke (born 1958) is an American civil servant who is the current Acting United States Secretary of Homeland Security.
On January 30, 2017, she was nominated by President Donald Trump to become the United States Deputy Secretary of Homeland Security under John F. Kelly.
She became acting Secretary of Homeland Security on July 31, 2017, when John F. Kelly assumed the office of White House Chief of Staff.

Unretrievable Questions for RAG:

Question: Who is the spouse of the Green performer?

Answer: Miquette Giraudy

supporting_facts:
Question: Green » performer
Answer: Steve Hillage
Doc: Green is the fourth studio album by British progressive rock musician Steve Hillage. Written in spring 1977 at the same time as his previous album, the funk-inflected "Motivation Radio" (1977), "Green" was originally going to be released as "The Green Album" as a companion to "The Red Album" (the originally intended name for "Motivation Radio"). However, this plan was dropped and after a US tour in late 1977, "Green" was recorded alone, primarily in Dorking, Surrey, and in London.
Question: #1 » spouse
Answer: Miquette Giraudy
Doc: Miquette Giraudy (born 9 February 1953, Nice, France) is a keyboard player and vocalist, best known for her work in Gong and with her partner Steve Hillage. She and Hillage currently form the core of the ambient band System 7. In addition to her performances in music, she has also worked as an actress, film editor and writer. In each role, she has used different stage names.

In the above (from prior work Musique), it can be observed that the document does not contain any information about *Miquette Giraudy spouse Green (Steve Hillage album)*, but only provides a brief description of Miquette Giraudy's biography. As a result, the RAG tool fails to retrieve this information and consequently cannot obtain the correct answer.

C Prompt Used

Prompts Used in Data Construction Pipeline 5–23:
Prompts Used in Evaluate 24 and 25:

gen_atomic_qa_prompt

```
You are an information extraction and question generation system.
# Task:
Given a document, extract a set of **atomic, verifiable facts** and convert each into a **QA pair**, where:
- The **question** focuses on a specific, retrievable detail from the document.
- The **answer** is concise, factual, and directly grounded in the document.
- For each document, generate **at most {gen_qa_num} QA pairs**. Prioritize the most concrete, unique, and verifiable facts.
- Only generate questions that require consulting the document to answer — avoid trivial facts or common-sense knowledge.
# Rules for QA Generation
1. Atomicity
- Each QA must be based on a single indivisible fact (no conjunctions).
✖ "A increased and B decreased" → must split into two questions.
2. Verifiability
- The answer must include at least one of:
- ✓ Numeric value (e.g., 59.0%)
- ✓ Time or date (e.g., 2025/04/28)
- ✓ Unique name/entity (e.g., Humpback65B)
- ✖ Reject vague expressions: "Performance has improved"
3. Time specificity
- Explicitly mark time ranges when containing time-sensitive information
- Examples:
✓ "Global GDP grew by 3.0% in 2023"
✖ "Recent GDP growth of 3.0%"
4. Relevance and Precision
- Avoid abstract questions. Focus on measurable and database-friendly details.
5. Answer Uniqueness
- The question must be **specific enough** to yield a **unique answer** from the document.
- ✖ Avoid under-specified questions that allow **multiple correct answers**.
- ✖ "What awards did Author X receive?" (if multiple awards are listed in the document)
- ✓ "What award did Author X receive in 2022?" (if only one is given for that year)
- ✖ Avoid vague superlatives like "notable," "important," "significant" without clear criteria.
# Output Format:
A JSON array of QA pairs. Each item contains:
- `question`: A specific, answerable question.
- `answer`: The factual value from the document.
# Examples
```json
[
{{
"question": "What is the number 1 sport in the usa?",
"answer": "American football"
}},
{{
"question": "Where did the Ottoman slave trade flourish?",
"answer": "In the Balkans"
}},
{{
"question": "Who was president when the white house was built?",
"answer": "John Adams"
}}
]
```
The document content to be processed is as follows:
{input_doc}
```

Figure 5: Gen atomic qa prompt

merge_qa_prompt(part1)

You are an expert in constructing multi-hop questions grounded in document-based facts.

Task
You are given multiple question-answer-document triples to generate **0 to {max_num}** multi-hop questions that require reasoning over the **latest previous hop** (i.e., the final element in Previous_Hops) together with New_pair. Use Previous_Hops strictly as supporting context: they may be consulted to verify entities/constraints and must be preserved (not removed, weakened, or contradicted) by the final question.
Only produce a multi-hop question if it is logically valid, unambiguous, and well-supported by both documents. If there is any uncertainty or weak connection, return empty JSON list instead of forcing a question.

Input Schema
Previous_Hops:
type: "list of objects (ordered oldest → latest)"
each_item:
- Hop_number
- Question: string
- Answer: string
- Doc: string
New_pair:
type: object (This is the candidate QA+Doc to attach to the chain.)
fields:
- Question: string
- Answer: string
- Doc: string

Output Format
Return a JSON list of up to {max_num} objects. Each object must be in one of the following formats:
Valid multi-hop case:
{{
"type": "inference" | "comparison",
"final_question": "...",
"final_answer": "..."
}}
If no high quality multi-hop question can be created:
Return an empty JSON list []

Types
- inference:
The final question chains information so that answering it requires (1) the last previous hop's facts and (2) the New_pair's facts. For inference cases, final_answer MUST equal New_pair Answer.
- comparison:
The final question compares a shared measurable dimension (e.g., date, numeric quantity, size). Both compared values must be supported explicitly in the documents. The answer should be one of the compared entities.

Rules
- 1. Use Previous_Hops as supporting context. Do NOT remove, weaken, or contradict any fact, constraint, or entity presented in any Previous_Hops's Question/Answer/Doc.
- 2. The final question MUST be generated from (latest Previous_Hop) + (New_pair). DO NOT remove or weaken any Previous_Hops's Question important information in final_question
- 3. DO NOT leak intermediate answers (no explicit exposure of any Previous_Hop answer in the final_question).
- 4. Do not simply restate the New_pair question, and do not return an unexpanded Hop question that is missing or only partially uses the information from the Previous Hop. The final question MUST depend on all QA-doc pairs to be answerable.
- 5. Ensure that there is **sufficient and accurate** evidence to get the answer.
- 6. When linking entities across docs, explicitly verify from the document contents (not just question text) that they refer to the exactly same real-world entity or fact
- 7. For comparison, only compare facts on the same axis (e.g., date vs date, size vs size), not unrelated attributes.
- 8. Do not create final questions that just state multiple facts independently, without a reasoning link between them.
- 9. Favor precision and factual grounding over producing more questions. If the logical connection is weak, ambiguous, or speculative, reject (return []).
- 10. Output only the final JSON list. Do NOT include chain-of-thought, explanations, or any extra text.

Examples
final_question is inference
Case 1 (Extend 1-hop question to 2-hop):
Input:
Hop_1:
Question: What is the name of the performer of "Qui de nous deux"?
Answer: Matthieu Chedid
Doc: "Qui de nous deux" is performed by Matthieu Chedid.

Figure 6: Merge qa prompt(part1)

merge_qa_prompt(part2)

New_pair:
Question: Who is the father of Matthieu Chedid?
Answer: Louis Chedid
Doc: Matthieu Chedid is the son of Louis Chedid.
Good Output:
{{
"type": "inference",
"final_question": "Who is the father of the performer of 'Qui de nous deux'?",
"final_answer": "Louis Chedid"
}}
✖ Error Case (leaks intermediate answer):
-"final_question": "Who is the father of Matthieu Chedid?"
Case 2 (Extend 2-hop inference question to 3-hop):
Input:
Hop_1:
Question: Who is the composer of "Al gran sole carico d'amore"?
Answer: Luigi Nono
Doc: "Al gran sole carico d'amore" is an opera with music by Luigi Nono.
Hop_2:
Question: Where did the composer of "Al gran sole carico d'amore" work?
Answer: Venice
Doc: Luigi Nono was active as a painter in Venice.
type: inference
New_pair:
Question: What is the name of the oldest bridge in Venice?
Answer: Rialto Bridge
Doc: The Rialto Bridge is the oldest bridge spanning the Grand Canal in Venice.
Good Output:
{{
"type": "inference",
"final_question": "What is the name of the oldest bridge in the city where the composer of 'Al gran sole carico d'amore' worked?",
"final_answer": "Rialto Bridge"
}}
✖ Error Case (leaks intermediate answer):
-"final_question": "What is the name of the oldest bridge in Luigi Nono worked?"
Case 3 (Extend 3-hop inference question to 4-hop):
Input:
Hop_1:
Question: Where was Francisco Vázquez born?
Answer: Guadalajara
Doc: Francisco H. Vázquez (born June 11, 1949 in Guadalajara, Jalisco, Mexico)
Hop_2:
Question: On which continent is Guadalajara located?
Answer: North America
Doc: Guadalajara is located in North America.
type: inference
Hop_3:
Question: Who was the Italian navigator who sailed for England and explored the east coast of the continent where Francisco Vázquez was born?
Answer: John Cabot
Doc: John Cabot (Italian: Giovanni Caboto; c. 1450 -- c. 1500) was a Venetian navigator and explorer whose 1497 discovery of the coast of North America under the commission of Henry VII of England
type: inference
New_pair:
Question: What is the name of the child of John Cabot?
Answer: Sebastian Cabot
Doc: Sebastian Cabot was the son of Italian explorer John Cabot (Giovanni Caboto)

Figure 7: Merge qa prompt(part2)

merge_qa_prompt(part3)

Good Output:
{{
"type": "inference",
"final_question": "What is the name of the child of the Italian navigator who sailed for England and explored the east coast of the continent where Francisco Vázquez was born?",
"final_answer": "Sebastian Cabot"
}}
✖ Error Case (leaks intermediate answer):
-"final_question": "What is the name of the child of the Italian navigator who sailed for England and explored the east coast of North America?"
✖ Error Case (remove or weaken important information in Previous_Hops):
-"final_question": "What is the name of the child of the navigator who explored the east coast of Francisco Vázquez was born?"
final_question is comparison
Input:
Question: When was John Beach born?
Answer: January 1, 1812
Doc: Major John Beach(January 1, 1812 - August 31, 1874) was a United States Army officer during the Black Hawk and American Civil War.
New_pair:
Question: When was Seth Gordon Persons born?
Answer: February 5, 1902
Doc: Seth Gordon Persons(February 5, 1902 - May 29, 1965) was an American Democratic politician who was the 43rd Governor of Alabama from 1951 to 1955.
Good Output:
{{
"type": "comparison",
"final_question": "Who was born first, John Beach or Seth Gordon Persons?",
"final_answer": "John Beach"
}}
Another Good Output:
{{
"type": "comparison",
"final_question": "Was John Beach born before Seth Gordon Persons?",
"final_answer": "Yes"
}}
✖ Error Case (leaks intermediate answer):
-"final_question": "Who was born first, John Beach (January 1, 1812 - August 31, 1874) or Seth Gordon Persons?"
✖ Invalid (spurious linkage: the QA-doc pairs contain unrelated facts that are superficially similar but logically disconnected.)
Input:
Hop_1:
Question: How many cardinals entered the papal conclave on March 31?
Answer: 27
Doc: Only twenty-seven cardinals entered the conclave on March 31, 1721.
New_pair:
Question: Which band did 27 open for in the Czech Republic?
Answer: Robert Plant
Doc: 27 is a rock band that opened for Robert Plant in Prague.
Correct Output:
[]
✖ Error Output (In Example: Cardinals and the rock band '27' are unrelated entities):
final_question: Which band did the cardinals who entered the papal conclave on March 31 open for in the Czech Republic?
final_answer: Robert Plant
✖ Invalid (spurious linkage: there is no necessary logical connection between them.)
Input:
Hop_1:
Question: What was the deployment order date for the 16th Army to the Ukraine?
Answer: 25 May 1941
Doc: The 16th Army was ordered to deploy to Ukraine on 25 May 1941.

Figure 8: Merge qa prompt(part3)

merge_qa_prompt(part4)

New_pair:
Question: Which two spheres of influence were involved in the division of Europe in the 1940s?
Answer: The Western world and the Soviet Union
Doc: Postwar Europe was divided into the Western and Soviet spheres of influence.

Correct Output:
[]
✖ Error Output (In Example: No causal or thematic link between army deployment date and geopolitical division):
final_question: What were the two major spheres of influence following the deployment of the 16th Army to the Ukraine in 1941?
final_answer: The Western world and the Soviet Union

✖ Invalid: (entity-based false link: a false connection between facts or documents that arises solely because different entities share identical or highly similar names, without any actual semantic or factual relationship.)
Input:
Hop_1:
Question: Who presents the Statewide Drive program at 107.9 ABC Ballarat?
Answer: Nicole Chvastek
Doc: "107.9 ABC Ballarat" has a total of 16 full time employees. A breakfast program is presented by Steve Martin from 6.15 am to 10.00 am weekdays. A mornings program is presented by Gavin McGrath from 10.00 am to 11.00 am weekdays. The regional "Statewide Drive" program (3.00 pm to 6.00 pm weekdays) is also broadcast from the Ballarat studios. It is presented by Nicole Chvastek and covers Victoria, southern New South Wales and a small part of eastern South Australia. It does not broadcast into the Melbourne metro area. 107.9 ABC Ballarat, callsign 3CRR, is an ABC Local Radio station.
New_pair:
Question: What toolkit has Nicole Joseph designed for breast care?
Answer: Breast-CareSolutions toolkit
Doc: Nicole Joseph introduced the global and multi-lingual breast care awareness campaign "The Gesture That Saves" in San Francisco in 2016 to 100 global peers from 40 countries during the VV100 retreat. She has designed a comprehensive Breast-CareSolutions toolkit and is currently designing a reproductive-health advocacy program. Nicole Joseph-Chin is the Chief Innovator, Founder and CEO of Ms. Brafit Limited.

Correct Output:
[]
✖ Error Output (In Example: Nicole Chvastek and Nicole Joseph are different individuals):
final_question: What toolkit has the presenter of the Statewide Drive program at 107.9 ABC Ballarat designed for breast care?
final_answer: Breast-CareSolutions toolkit

✖ Invalid: (trivial concatenation: the final question simply combines facts without needing reasoning or integration.)
Input:
Hop_1:
Question: Where was the State Normal School at Cheney located by the end of the term's first week?
Answer: Pomeroy building
Doc: By the end of the term's first week, the State Normal School at Cheney was located in the Pomeroy building.
New_pair:
Question: Who designed the Cheney Building?
Answer: H. H. Richardson
Doc: The Cheney Building was designed by H. H. Richardson.

Correct Output:
[]
✖ Error Output (In Example: Final_question simply concatenates the two questions together using 'and' without needing reasoning.):
final_question: Which building was used for the State Normal School at Cheney by the end of the term's first week, and who designed the Cheney Building?
final_answer: Pomeroy building, H. H. Richardson

Invalid (lacking_evidence: one or both compared values are not explicitly supported in the provided documents with sufficient precision to support the asserted comparison.)
Hop_1:
Question: What year was the Cambridge Battery completed for the 100-ton gun?
Answer: 1886
Doc: Cambridge Battery was ready in 1886.

Figure 9: Merge qa prompt(part4)

merge_qa_prompt(part5)

New_pair:
Question: When did the 1886 United Kingdom general election take place?
Answer: July 1-27, 1886
Doc: The 1886 United Kingdom general election took place from 1 July to 27 July 1886.

Correct Output:
[]
✖ Error Output (In Example: The Cambridge Battery is only dated to the year (1886), while the election records include exact dates in July. This lack of precise dating in the document makes it impossible to determine which event occurred first.):
final_question: Which event occurred first, the completion of the Cambridge Battery in 1886 or the United Kingdom general election led by Charles Stewart Parnell in 1886?
final_answer: The completion of the Cambridge Battery in 1886.

Checklist
- [ ] Do all docs describe logically connectable facts?
- [ ] For comparison: are both facts from the same measurable dimension?
- [ ] For inference: does the reasoning chain correctly lead to New_pair Answer?
- [ ] Is the final question truly unanswerable without all QA-doc pairs?
- [ ] Are all intermediate answers hidden?
- [ ] Are the linked entities explicitly confirmed as the same in both documents?
- [ ] Is there sufficient and accurate evidence to get the answer?

The data need to be processed is as follows:
{Data}

New_pair:
Question: {New_question}
Answer: {New_answer}
Doc: {New_document}
Only output the final JSON object. Do not explain your reasoning.

Figure 10: Merge qa prompt(part5)

inferenec_check_prompt(part1)

```
You are a multi-hop QA verification system.
## Task
You are given a multi-hop QA construction based on two question-answer-document triples:
(Question1, Answer1, Doc1) and (Question2, Answer2, Doc2), and a final multi-hop QA:
- Final_question
- Final_answer
- type: "inference"
Your job is to **verify whether the final QA is logically valid** according to the reasoning paths and documents.
## Input Fields
- Question1, Answer1, Doc1
- Question2, Answer2, Doc2
- Final_question, Final_answer
- type: "inference"
## Output Format
Return a JSON object:
{{
"valid": "true" | "false",
"error_type": "bad_linkage" | "entity_false_link" | "trivial_concatenation" | "other",
"justification": "Short explanation of the issue"
}}
## Definitions & Rules
- "inference": Final question requires combining QA1 and QA2 in a reasoning chain. The final_answer must exactly
match Answer2. No intermediate answers should appear in final_question.
- "bad_linkage": If the two QA-doc pairs contain unrelated facts that are superficially similar but logically
disconnected.
- "entity_false_link": a false connection between two facts or documents that arises solely because different entities
share identical or highly similar names, without any actual semantic or factual relationship.
- "trivial_concatenation": If the final question is formed by simply joining two or more independent facts from the
given QA-document pairs into a single sentence (often using "and" or similar conjunctions), without any logical
reasoning beyond listing the facts.
- "other": Other errors that you think are not included above.
## Examples
### ✓ Valid
Input:
Question1: What is the name of the performer of 'Qui de nous deux'?
Answer1: Matthieu Chedid
Doc1: "Qui de nous deux" is performed by Matthieu Chedid.
Question2: Who is the father of Matthieu Chedid?
Answer2: Louis Chedid
Doc2: Matthieu Chedid is the son of Louis Chedid.
Final_question: Who is the father of the performer of 'Qui de nous deux'?
Final_answer: Louis Chedid
type: "inference"
Expected Output:
{{
"valid": "true",
"error_type": "No Error",
"justification": "Reasoning chain is valid."
}}
### ✖ Invalid (bad linkage)
Input:
Question1: How many cardinals entered the papal conclave on March 31?
Answer1: 27
Doc1: 27 cardinals entered the 1721 Papal Conclave.
Question2: Which band did 27 open for in the Czech Republic?
Answer2: Robert Plant
Doc2: 27 is a rock band that opened for Robert Plant.
Final_question: Which band did the cardinals who entered the papal conclave on March 31 open for in the Czech
Republic?
Final_answer: Robert Plant
type: "inference"
```

Figure 11: Inference check prompt(part1)

inferenec_check_prompt(part2)

Expected Output:
{{
"valid": "false",
"error_type": "bad_linkage",
"justification": "Cardinals and the rock band '27' are unrelated entities"
}}
✖ Invalid (bad linkage)
Input:
Question1: What was the deployment order date for the 16th Army to the Ukraine?
Answer1: 25 May 1941
Doc1: The 16th Army was ordered to deploy to Ukraine on 25 May 1941.
Question2: Which two spheres of influence were involved in the division of Europe in the 1940s?
Answer2: The Western world and the Soviet Union
Doc2: Postwar Europe was divided into the Western and Soviet spheres of influence.
Final_question: What were the two major spheres of influence following the deployment of the 16th Army to the Ukraine in 1941?
Final_answer: The Western world and the Soviet Union
type: "inference"
Expected Output:
{{
"valid": "false",
"error_type": "bad_linkage",
"justification": "No causal or thematic link between army deployment date and geopolitical division"
}}
✖ Invalid (entity_false_link)
Input:
Question1: Who presents the Statewide Drive program at 107.9 ABC Ballarat?
Answer1: Nicole Chvastek
Doc1: "107.9 ABC Ballarat" has a total of 16 full time employees. A breakfast program is presented by Steve Martin from 6.15 am to 10.00 am weekdays. A mornings program is presented by Gavin McGrath from 10.00 am to 11.00 am weekdays. The regional "Statewide Drive" program (3.00 pm to 6.00 pm weekdays) is also broadcast from the Ballarat studios. It is presented by Nicole Chvastek and covers Victoria, southern New South Wales and a small part of eastern South Australia. It does not broadcast into the Melbourne metro area. 107.9 ABC Ballarat, callsign 3CRR, is an ABC Local Radio station.
Question2: What toolkit has Nicole Joseph designed for breast care?
Answer2: Breast-CareSolutions toolkit
Doc2: Nicole Joseph introduced the global and multi-lingual breast care awareness campaign "The Gesture That Saves" in San Francisco in 2016 to 100 global peers from 40 countries during the VV100 retreat. She has designed a comprehensive Breast-CareSolutions toolkit and is currently designing a reproductive-health advocacy program. Nicole Joseph-Chin is the Chief Innovator, Founder and CEO of Ms. Brafit Limited.
Final_question: What toolkit has the presenter of the Statewide Drive program at 107.9 ABC Ballarat designed for breast care?
Final_answer: Breast-CareSolutions toolkit
type: "inference"
Expected Output:
{{
"valid": "false",
"error_type": "entity_false_link",
"justification": "Nicole Chvastek and Nicole Joseph are different individuals"
}}
✖ Invalid: (trivial_concatenation)
Input:
Question1: Where was the State Normal School at Cheney located by the end of the term's first week?
Answer1: Pomeroy building
Doc1: By the end of the term's first week, the State Normal School at Cheney was located in the Pomeroy building.
Question2: Who designed the Cheney Building?
Answer2: H. H. Richardson
Doc2: The Cheney Building was designed by H. H. Richardson.
Final_question: Which building was used for the State Normal School at Cheney by the end of the term's first week, and who designed the Cheney Building?
Final_answer: Pomeroy building, H. H. Richardson
type: "inference"
Expected Output:
{{
"valid": "false",
"error_type": "trivial_concatenation",
"justification": "Final_question simply concatenates the two questions together using 'and' without needing reasoning."
}}
The data need to be processed is as follows:
Question1: {Question1}
Answer1: {Answer1}
Doc1:{Document1}
Question2: {Question2}
Answer2: {Answer2}
Doc2:{Document2}
Final_question:{Final_question}
Final_answer:{Final_answer}
type:{qa_type}
Only return the JSON object as described. Do not include explanations unless requested.

Figure 12: Inference check prompt(part2)

inference_check_prompt(part3)

✖ Invalid: (trivial_concatenation)
Input:
Question1: Where was the State Normal School at Cheney located by the end of the term's first week?
Answer1: Pomeroy building
Doc1: By the end of the term's first week, the State Normal School at Cheney was located in the Pomeroy building.
Question2: Who designed the Cheney Building?
Answer2: H. H. Richardson
Doc2: The Cheney Building was designed by H. H. Richardson.
Final_question: Which building was used for the State Normal School at Cheney by the end of the term's first week, and who designed the Cheney Building?
Final_answer: Pomeroy building, H. H. Richardson
type: "inference"
Expected Output:
{{
"valid": "false",
"error_type": "trivial_concatenation",
"justification": "Final_question simply concatenates the two questions together using 'and' without needing reasoning."
}}
The data need to be processed is as follows:
Question1: {Question1}
Answer1: {Answer1}
Doc1:{Document1}
Question2: {Question2}
Answer2: {Answer2}
Doc2:{Document2}
Final_question:{Final_question}
Final_answer:{Final_answer}
type:{qa_type}
Only return the JSON object as described. Do not include explanations unless requested.

Figure 13: Inference check prompt(part3)

comparison_check_prompt(part1)

```
You are a multi-hop QA verification system.
## Task
You are given two question-answer-document triples:
(Question1, Answer1, Doc1) and (Question2, Answer2, Doc2), plus a final multi-hop QA:
- Final_question
- Final_answer
- type: "comparison"
Your job is to **verify whether the final QA is logically valid** according to the reasoning paths and documents.
## Input Fields
- Question1, Answer1, Doc1
- Question2, Answer2, Doc2
- Final_question, Final_answer
- type: "inference"
## Output Format
Return a JSON object:
{{
"valid": "true" | "false",
"error_type": "forced_pairing" | "lacking_evidence" | "trivial_concatenation" | "other",
"justification": "Short explanation of the issue"
}}
## Definitions & Rules
- "comparison": Final question compares a **shared attribute/dimension** (e.g., date, numeric quantity, size)
between two entities derived from QA1 and QA2.
- "forced_pairing": The two QA-doc pairs do not share a meaningful, comparable dimension — the comparison is
forced or domain-incoherent.
- "lacking_evidence": One or both compared values are not explicitly supported in the provided documents with
sufficient precision to support the asserted comparison.
- "trivial_concatenation": If the final question is formed by simply joining two or more independent facts from the
given QA-document pairs into a single sentence (often using "and" or similar conjunctions), without any logical
comparing beyond listing the facts.
- "other": Other errors that you think are not included above.
## Examples
### ✓ Valid
Input:
Question1: When was John Beach born?
Answer1: January 1, 1812
Doc1: Major John Beach (January 1, 1812 - August 31, 1874) was a United States Army officer during the Black Hawk
and American Civil War.
Question2: When was Seth Gordon Persons born?
Answer2: February 5, 1902
Doc2: Seth Gordon Persons (February 5, 1902 - May 29, 1965) was an American Democratic politician and the 43rd
Governor of Alabama.
Final_question: Who was born first, John Beach or Seth Gordon Persons?
Final_answer: John Beach
type: "comparison"
Expected Output:
{{
"valid": "true",
"error_type": "No Error",
"justification": "Both docs provide explicit birth dates (1812 vs 1902) and John Beach is earlier."
}}
### ✖ Invalid (forced_pairing)
Input:
Question1: What was the renumbering date of the 17th Lancers?
Answer1: April 1763
Doc1: The regiment was renumbered the 17th Regiment of (Light) Dragoons in April 1763.
Question2: What year was the 67th English cricket season?
Answer2: 1763
Doc2: The 1763 English cricket season was the 67th English cricket season.
Final_question: Was the renumbering of the 17th Lancers in April 1763 before or during the 67th English cricket
season?
Final_answer: During
type: "comparison"
```

Figure 14: Comparison check prompt(part1)

comparison_check_prompt(part2)

Expected Output:
{{
"valid": "false",
"error_type": "forced_pairing",
"justification": "This forces a link between a military renumbering (point event) and a sports season (period) without a meaningful shared comparison dimension."
}}
✖ Invalid (lacking_evidence)
Input:
Question1: What year was the Cambridge Battery completed for the 100-ton gun?
Answer1: 1886
Doc1: Cambridge Battery was ready in 1886.
Question2: When did the 1886 United Kingdom general election take place?
Answer2: July 1-27, 1886
Doc2: The 1886 United Kingdom general election took place from 1 July to 27 July 1886.
Final_question: Which event occurred first, the completion of the Cambridge Battery in 1886 or the United Kingdom general election led by Charles Stewart Parnell in 1886?
Final_answer: The completion of the Cambridge Battery in 1886
type: "comparison"
Expected Output:
{{
"valid": "false",
"error_type": "lacking_evidence",
"justification": "The Cambridge Battery is only dated to the year (1886), while the election records include exact dates in July. This lack of precise dating in the document makes it impossible to determine which event occurred first."
}}
✖ Invalid: (trivial_concatenation)
Input:
Question1: Where was the State Normal School at Cheney located by the end of the term's first week?
Answer1: Pomeroy building
Doc1: By the end of the term's first week, the State Normal School at Cheney was located in the Pomeroy building.
Question2: Who designed the Cheney Building?
Answer2: H. H. Richardson
Doc2: The Cheney Building was designed by H. H. Richardson.
Final_question: Which building was used for the State Normal School at Cheney by the end of the term's first week, and who designed the Cheney Building?
Final_answer: Pomeroy building, H. H. Richardson
type: "comparison"
Expected Output:
{{
"valid": "false",
"error_type": "trivial_concatenation",
"justification": "Final_question simply concatenates the two questions together using 'and' without needing reasoning."
}}
The data need to be processed is as follows:
Question1: {Question1}
Answer1: {Answer1}
Doc1: {Document1}
Question2: {Question2}
Answer2: {Answer2}
Doc2: {Document2}
Final_question: {Final_question}
Final_answer: {Final_answer}
type: {qa_type}
Only return the JSON object as described. Do not include explanations unless requested.

Figure 15: Comparison check prompt(part2)

refine_prompt

```
You are an AI agent tasked with cleaning and extracting concise answers from original QA pairs.
## Input:
You are given a **question** and its corresponding **original answer**. Your task is to extract the most precise and concise information that directly answers the question.
## Processing Rules:
1. Extract **only** the exact information requested in the question.
2. Keep the original index numbering or order if present.
3. **Do not** omit essential information.
4. **Never add or infer** information not explicitly stated in the original answer.
5. Follow strict formatting conventions:
- Percentages: use format like `8%` (not "eight percent" or "8 percent")
- Currency: use `$1,000` format
- Dates: use `YYYY-MM-DD`
- Units: include units (e.g., `5kg`, `10cm`)
6. For answers that consist of multiple parts or are comparative in nature, multiple core components and comparative statements should be included.
## Output Format (JSON):
For each input QA pair, output the following JSON object:
{{
"question": "<original question>",
"original_answer": "<original answer>",
"refined_answer": "<clean, concise, and direct answer>"
}}
## Example:
Input:
question: What edition of the Wightman Cup was held in 1931?
original_answer: The 1931 Wightman Cup was its 9th edition.
Output:
{{
"question": "What edition of the Wightman Cup was held in 1931?",
"original_answer": "The 1931 Wightman Cup was its 9th edition.",
"refined_answer": "The 9th edition."
}}
Input:
question: How does the percentage of individuals under age 18 living below the poverty line in Farina, Illinois compare to the statewide percentage in Illinois?
original_answer: In Farina, Illinois, 10.7% of individuals under age 18 were living below the poverty line, which is lower than the statewide percentage in Illinois of 16.1%.
Output:
{{
"question": "How does the percentage of individuals under age 18 living below the poverty line in Farina, Illinois compare to the statewide percentage in Illinois?",
"original_answer": "In Farina, Illinois, 10.7% of individuals under age 18 were living below the poverty line, which is lower than the statewide percentage in Illinois of 16.1%.",
"refined_answer": "10.7%, lower than the statewide percentage in Illinois of 16.1%."
}}
The data need to be processed is as follows:
question: {question}
original_answer: {original_answer}
```

Figure 16: Refine prompt

more_optional_answer_prompt

You are an expert in **linguistic variation** and **data augmentation**. Your task is to generate a comprehensive list of all plausible and commonly recognized alternative expressions, formats, and aliases for a given input entity or piece of information. The goal is to create high-quality training data that captures diverse ways of referring to the same concept.

Key Guidelines:

1. **Equivalence:** Each alternative expression must refer to *exactly the same entity or information* as the original input. Do not include broader categories, narrower sub-types, or related but distinct concepts.
2. **Scope of Variation:** Focus on:
* Different **formatting conventions** (e.g., dates, numbers, units).
* Common **abbreviations, acronyms, or initialisms**.
* Well-known **aliases, nicknames, or shorter forms** in common usage.
* Synonyms or rephrasing should *only* be included if they are direct, commonly accepted equivalents.
3. **Context-Agnosticism:** Unless the input itself implies a specific context, generate general-purpose variations. Avoid creating variations that are only valid in very niche or obscure contexts.
4. **Inclusion of Original:** Always include the original input as the first item in the generated list.
5. **Format:** Output the variations as a JSON list of strings.

Examples:

Input: 1977-01-26
Output: ["1977-01-26", "1977 01 26", "1977.01.26", "January 26, 1977", "26 Jan 1977", "Jan 26, 1977"]

Input: United Nations
Output: ["United Nations", "U.N.", "UN"]

Input: 3.14159
Output: ["3.14159", "π", "pi", "PI"]

Input: Doctor of Philosophy
Output: ["Doctor of Philosophy", "Ph.D.", "PhD", "Doctorate"]

Input: New York City
Output: ["New York City", "NYC", "The Big Apple"]

Input: kilogram
Output: ["kilogram", "kg", "kilograms"]

Input: {refined_answer}
Please list all possible textual expressions that have the same meaning or refer to the same entity, especially in different formats (e.g., dates, names, abbreviations).
Respond with a JSON list of strings. Do not explain.

Figure 17: More optional answer prompt

reasoning_prompt

Please solve the following problem and return result. Ensure responses are as concise as possible, focusing only on key information while omitting redundant details.
The problem is:
{problem}

Figure 18: Reasoning prompt

reasoning_prompt_comparison

Please solve the following problem and return result.
For comparison question, if you are unsure of the answer, please do not guess or choose randomly. Instead, return "I cannot answer this question."
The problem is:
{problem}
Ensure responses are as concise as possible, focusing only on key information while omitting redundant details.

Figure 19: Reasoning prompt comparison

singlehop_prompt

You are given the following document that contains relevant information to help answer a question.
Document:
{Document}
Question:
{Question}
Please answer the question using the information in the provided document. Ensure responses are as concise as possible, focusing only on key information while omitting redundant details.
If you cannot answer the question, return "I cannot answer this question. <the reason why cannot answer this question>".
For Example:
✓ Can answer:
Document:
"Olive Winchester"
1851 in Corinna, Maine; died October 2, 1892 in Yankton, South Dakota), and Sarah A. ""Sadie"" Blackstone Winchester (born May 1, 1853 in Pownal, Maine; died February 6, 1949 in Los Angeles, California). Winchester's parents were married in Portland, Maine on February 22, 1879 in the Methodist Episcopal Church. Winchester was a relative of Oliver Fisher Winchester (born November 30, 1810 in Brookline, Massachusetts; died December 11, 1880 in New Haven, Connecticut), the manufacturer and marketer of the Winchester repeating rifle. After June 25, 1880, the Winchester family left Monson, Maine and by 1881 had relocated to Forestburg, in Sanborn
Question:
When was the first woman to graduate with a Bachelor of Divinity degree from Glasgow, Olive Winchester, born?
Answer:
1851
Document:
"1996 North Indian Ocean cyclone season"
due to flash flooding. Elsewhere in India, the storm killed 111 people, including 44 in Tamil Nadu where 18 boats were damaged or missing. In some areas, the rains helped end a drought. After the storm passed, the Andhra Pradesh government provided each family RS$1,000 (US$30) if their house was destroyed, and RS$100,000 (US$3,000) if they lost a family member. While the previous storm was paralleling the east Indian coastline, another disturbance formed off the west coast on June 15, also associated with the monsoon. The new area of convection persisted, developing a distinct circulation by the next day.
Question:
What is the prime factorization of the number of people killed by the storm in India during the 1996 North Indian Ocean cyclone season?
Answer:
3 x 37
Document:
Seth Gordon Persons (February 5, 1902 - May 29, 1965) was an American Democratic politician who was the 43rd Governor of Alabama from 1951 to 1955.
Question:
Who was born first, John Beach (January 1, 1812 - August 31, 1874) or Seth Gordon Persons?
Answer:
John Beach
✖ Cannot answer:
Document:
The Rialto Bridge (Italian: Ponte di Rialto; Venetian: Ponte de Rialto) is the oldest of the four bridges spanning the Grand Canal in Venice, Italy. Connecting the sestieri (districts) of San Marco and San Polo, it has been rebuilt several times since its first construction as a pontoon bridge in the 12th century, and is now a significant tourist attractiojn the city.
Question:
What is the name of the famous bridge in the place where Al gran sole carico d'amore's composer worked?
Answer:
I cannot answer this question. I don’t know where the composer of Al gran sole carico d’amore has worked.
Document:
Worst (manga) Worst is a Japanese delinquent manga series written and illustrated by Hiroshi Takahashi. It has the same setting as Takahashi's previous manga ""Crows"" and ""QP"" and revolves around a group of teenage boys who fight their way through the notorious high school, ""Suzuran"". The manga was first published by Shōnen Champion in 2002. The series is currently being serialized in Japan and has been collected into twenty-five tankōbon volumes. In North America, Digital Manga Publishing has released only three volumes, with the last graphic novel released in November 2004. The series is currently on hiatus but Digital Manga
Question:
Which of the following was critiqued as one of the worst manga, .hack//Legend of the Twilight or the manga Worst, published in 2002?
Answer:
I cannot answer this question. I don't have sufficient information about "hack/Legend of the Twilight".

Figure 20: Singlehop prompt

```
multihop_inference_prompt

You are an expert at solving problems. Now you need to solve a multi-hop inference problem.
Multi-hop inference promblem: a question that requires combining information from multiple sources in a logical chain to reach an answer.
## For Example:
Input:
Question1: "What is the name of the performer of Qui de nous deux?"
Answer1: "Matthieu Chedid"
Supporting Document1: "'Qui de nous deux' is performed by Matthieu Chedid."
Question2: "Who is the father of Matthieu Chedid?"
Supporting Document2: "Matthieu Chedid is the son of Louis Chedid."
FinalQuestion: "Who is the father of the performer of Qui de nous deux?"
Output:
"Louis Chedid"
## The Example's Logic Chain (Just to help you better understand the multihop problem. Don't output something like this):
FinalQuestion:"Who is the father of the performer of Qui de nous deux?" -> Q:"What is the name of the performer of Qui de nous deux?" A:"Matthieu Chedid" -> Q:"Who is the father of Matthieu Chedid?" A:"Louis Chedid" -> FinalAnswer: "Louis Chedid"
## Now you are given some supporting fact to help answer a question.
{Data}
FinalQuestion: {FinalQuestion}
Now you need sole the promblem and return result. Ensure responses are as concise as possible, focusing only on key information while omitting redundant details.
```

Figure 21: Multi-hop inference prompt

```
multihop_comparison_prompt

You are an expert at solving problems. Now you need to solve a multi-hop comparison problem.
Multi-hop comparison promblem: a question that requires retrieving and comparing information from multiple sources to determine a relative fact.
## For Example:
Input:
Question1: "When was John Beach born?"
Answer1: "January 1, 1812"
Supporting Document1: "Major John Beach( January 1, 1812 - August 31, 1874) was a United States Army officer during the Black Hawk and American Civil War."
Question2: "When was Seth Gordon Persons born?"
Answer2: "February 5, 1902"
Supporting Document2: "Seth Gordon Persons( February 5, 1902 - May 29, 1965) was an American Democratic politician who was the 43rd Governor of Alabama from 1951 to 1955."
FinalQuestion: "Who was born first, John Beach or Seth Gordon Persons?"
Output:
"John Beach"
## The Example's Logic Chain (Just to help you better understand the multihop problem. Don't output something like this):
FinalQuestion: "Who was born first, John Beach or Seth Gordon Persons?" -> Q:"When was John Beach born?" A:"January 1, 1812" + Q:"When was Seth Gordon Persons born?" A:"February 5, 1902" -> FinalAnswer: "John Beach"
## Now you are given some supporting fact to help answer a question.
{Data}
FinalQuestion: {FinalQuestion}
Now you need sole the promblem and return result. Ensure responses are as concise as possible, focusing only on key information while omitting redundant details.
```

Figure 22: Multi-hop comparison prompt

llm_judge_prompt

You are an expert evaluator. Evaluate whether the OTHER ANSWER preserves **all essential information** in the GOLDEN ANSWER, **with respect to the QUESTION**.

Scoring Criteria
- **2 points** → OTHER ANSWER is fully equivalent to the GOLDEN ANSWER. Same meaning, even if reworded or paraphrased. No missing or incorrect information.
- **1 point** → OTHER ANSWER includes ALL key information from the GOLDEN ANSWER but adds **extra non-contradictory information** that may not be strictly necessary but is still valid in context.
- **0 points** → OTHER ANSWER is **missing** critical information from the GOLDEN ANSWER or introduces **incorrect/contradictory** information, based on the QUESTION.

Always consider what the QUESTION is asking when judging whether information is essential.

✓ Positive Examples

✓ 2 points

- Question: What year did the war end?
Golden: 1848
Other: The year was 1848.

- Question: Who became the first African American U.S. president?
Golden: Barack Obama
Other: Obama
- Question: When did the battle begin?
Golden: The battle began in 1775.
Other: The conflict started in the year 1775.

- Question: Which field of Nobel Prize did Marie Curie receive?
Golden: the Nobel Prize in Physics.
Other: Physics.

✓ 1 point

- Question: What is the cause of death of Mercedesz Henger's father?
Golden: diabetes mellitus.
Other: Diabetes mellitus type 2.

- Question: When did the war end?
Golden: 1848
Other: The war ended in 1848.
- Question: Who became the first African American U.S. president?
Golden: Barack Obama
Other: Barack Obama, the 44th president of the United States.

- Question: Where is the Eiffel Tower located?
Golden: The Eiffel Tower is in Paris.
Other: The Eiffel Tower is in Paris, the capital of France.

✖ Negative Examples (0 points)

- Question: How much is the price?
Golden: 50
Other: 25 dollars
- Question: When did the war end?
Golden: 1848
Other: In 1846
- Question: Where is the Eiffel Tower located?
Golden: The Eiffel Tower is in Paris.
Other: The Eiffel Tower is in Berlin.
- Question: Who became the first African American U.S. president?
Golden: Barack Obama
Other: Barack Obama and Abraham Lincoln were presidents during the same era.

Output Format:
Return ONLY JSON, no extra text.
```json
{
"answer_reason": "reason for the score",
"answer_score": 0/1/2
}
```

Figure 23: LLM judge prompt

thinking_prompt

You are a world expert at making efficient plans to solve any task using an RAG Search tool.
Now for the given task, develop a step-by-step high-level plan.
Each step should be a concise action that uses the RAG search tool, specifying the keyword you want to search for.
Do not skip steps, do not add any superfluous steps. Only write the high-level plan, DO NOT Execute TOOL CALLS in this planning step.
Each step should be clearly numbered.

Here is your task:
{task['final_question']}

Now begin! Write your plan below.

Figure 24: Thinking prompt

action_prompt

Based on the plan and the search results before (if there is), first analyse what information you have gained and what other information you still need, then Execute ONLY ONE MORE step using the RAG search tool.
You should never assume or invent any search results that are not explicitly provided in the context.
If there is no search results before, just write 'Currently, no results are available.' directly instead of guessing and directly Execute the first step in your plan using the RAG search tool.
If you find there is nothing more to search, do not search for the same or similar thing you have searched again.
Instead, directly provide the final concise answer to the original question using the Final_Answer tool.
When providing the final answer, DO NOT include any other content such as search results or any thoughts, just the concise final answer itself.

Figure 25: Action prompt

The LLM judge prompt is the same as the prompt used in the data construction process.

D Error Case Analysis

To ensure conciseness and highlight the key points, we present below specific model output examples for the **Error Case Study**, with some non-essential parts omitted.

For Table 5 6, it can be observed that the model incorrectly transformed *"sixth career game-winner"* in the second step into *"the professional football team he first faced"*, thereby causing the entire reasoning trajectory to deviate.

User: You are a world expert at making efficient plans to solve any task using an RAG Search tool. ... Here is your task: Which broader trading area does the metropolitan area of the city where the professional football team that faced the player who kicked a 27-yard field goal early in the final period for the 2009 UCLA Bruins football team in his sixth career game-winner is based form the core of?
Assistant: 1. Search RAG for "2009 UCLA Bruins 27-yard field goal early in the final period sixth career game-winner" to identify the player and the specific game. 2. Search RAG for "[Player Name] first NFL game opponent" to determine which professional football team he first faced. 3. Search RAG for "[Opponent Team Name] home city" to find the city where that professional team is based. 4. Search RAG for "[City Name] metropolitan area" to get the official metropolitan area designation. 5. Search RAG for "[Metropolitan Area Name] forms the core of" to identify the broader trading area that metro area forms the core of.
User: Based on the plan and the search results before (if there is), first analyse what information you have gained and what other information you still need, then EXECUTE ONLY ONE MORE step using the RAG search tool. ...
Assistant: `"function_call": {"name": "RAG_search", "arguments": "{query: 2009 UCLA Bruins 27-yard field goal early in the final period sixth career game-winner, topk: 5}"}`
[RAGTool Results] Query: "2009 UCLA Bruins 27-yard field goal early in the final period sixth career game-winner" Topk = 5 Document: "2009 UCLA Bruins football team" ... Kai Forbath kicked a 27-yard field goal early in the final period
User: Based on the plan and the search results before (if there is), first analyse what information you have gained and what other information you still need, then EXECUTE ONLY ONE MORE step using the RAG search tool. ...
Assistant: `"function_call": {"name": "RAG_search", "arguments": "{query: Kai Forbath first NFL game opponent, topk: 5}"}`
[RAGTool Results] Query: "Kai Forbath first NFL game opponent" Topk = 5 Document: Kai Forbath" ... Forbath made his NFL and Redskins debut knocking through a 50-yard attempt as his first career NFL field goal in the home game against the Minnesota Vikings in Week 6 on October 14, 2012.
... As a consequence, the entire subsequent reasoning trajectory deviates accordingly.
Assistant: Final_Answer: Minneapolis–St. Paul MN–WI Combined Statistical Area

Table 5: Error Case

Hop1:
Atomic Question: Who kicked a 27-yard field goal early in the final period for the 2009 UCLA Bruins football team?
Atomic Answer: Kai Forbath
Doc: "2009 UCLA Bruins football team" ... Kai Forbath kicked a 27-yard field goal early in the final period

Hop2:
Atomic Question: Against which team did Kai Forbath achieve his sixth career game-winner?
Atomic Answer: Atlanta Falcons
Doc: "Kai Forbath" ... season finale in Week 17 against the Atlanta Falcons as time expired, his sixth career game-winner.
Final Question: Against which team did the player who kicked a 27-yard field goal early in the final period for the 2009 UCLA Bruins football team achieve his sixth career game-winner?
Final Answer: Atlanta Falcons

Hop3:
Atomic Question: What is the location of the Atlanta Falcons' professional football team?
Atomic Answer: Atlanta, Georgia
Doc: "Atlanta Falcons" ... Atlanta Falcons The Atlanta Falcons are a professional American football team based in Atlanta, Georgia.
Final Question: What is the location of the professional football team that faced the player who kicked a 27-yard field goal early in the final period for the 2009 UCLA Bruins football team in his sixth career game-winner?
Final Answer: Atlanta, Georgia

Hop4:
Atomic Question: Which broader trading area does the Atlanta, Georgia metropolitan area form the core of?
Atomic Answer: Atlanta–Athens-Clarke–Sandy Springs Combined Statistical Area
Doc: "Atlanta metropolitan area" ... The metro area forms the core of a broader trading area, the Atlanta–Athens-Clarke–Sandy Springs Combined Statistical Area. The Combined Statistical Area spans up to 39 counties in north Georgia and has an
Final Question: Which broader trading area does the metropolitan area of the city where the professional football team that faced the player who kicked a 27-yard field goal early in the final period for the 2009 UCLA Bruins football team in his sixth career game-winner is based form the core of?
Final Answer: Atlanta–Athens-Clarke–Sandy Springs Combined Statistical Area

Table 6: Golden Trace

E Human Evaluation Details and Ethical Considerations

Annotator Recruitment. Human evaluators were recruited from adult participants fluent in English. All annotators were informed of the study's purpose and participated voluntarily. To ensure the reliability of judgments and high-quality annotations, we conducted a comprehensive training session for all participants to ensure they fully understood the guidelines and could correctly perform the evaluation tasks.

Annotation Procedure and Criteria. Annotators were presented with data samples requiring human verification. The task was designed as a binary classification problem: annotators were asked to either "Retain" or "Discard" the sample based on strict quality standards. The evaluation criteria consisted of two main categories: *General Quality Checks:* Annotators verified Factuality & Faithfulness by ensuring the answer was strictly grounded in the source text and free from any information fabrication or hallucination. They also checked Logic & Fluency to ensure questions were grammatically natural and logically sound. *Type-Specific Checks:* For Reasoning Questions, annotators reviewed the reasoning chain layer by layer to ensure that every intermediate step constituted a valid deduction. For Comparison Questions, they verified the consistency of the comparison dimensions and the factual accuracy of the final conclusion.

Inter-Annotator Agreement and Adjudication. All data samples requiring human verification were independently evaluated by three annotators. To assess the reliability of the annotation process, we calculated the Inter-Annotator Agreement (IAA) using Fleiss' Kappa (κ). The resulting κ score was 0.65, indicating substantial agreement among annotators. For cases where annotators disagreed (i.e., non-unanimous decisions), the final labels were determined through a consensus discussion involving both the annotators and the authors to ensure the highest data quality.

Compensation. Annotators were compensated at a fixed rate of $15 per hour, which is consistent with or above standard minimum wage guidelines. The compensation was independent of the label distribution to avoid incentive bias.

Ethical Considerations. The annotation process did not involve the collection of any personally identifiable information (PII). All content was anonymized. Given the non-invasive nature of the task, this study aligns with standard ethical guidelines for NLP data annotation.

F Overlap Analysis with Existing Multi-hop Benchmarks

We analyze the overlap between our dataset and existing multi-hop QA benchmarks, including HotpotQA, 2Wiki, and MuSiQue. **TF-IDF (1–3 n-grams) cosine similarity:** mean = 0.1344, median = 0.1231. **SBERT semantic similarity:** mean = 0.5213, median = 0.5178. The low TF-IDF scores suggest limited lexical overlap, while the moderate SBERT similarity reflects general semantic relatedness without substantial content duplication. Overall, these results indicate minimal overlap with existing benchmarks.

G LLM-as-a-Judge Specification and Reliability

We provide detailed specifications of the LLM-based judge and evaluate its reliability. **Judge model.** The main experiments use GPT-4o-mini as the evaluation model, with temperature set to 0.0 to ensure deterministic outputs. **Stability and cross-model consistency.** To assess robustness, we randomly sampled 60 instances (10 from each subset) and re-evaluated all model predictions using Grok-4 as an independent judge. The judgments from Grok-4 were identical to those from GPT-4o-mini on all 60 cases, indicating strong cross-model consistency. **Human agreement.** We further conducted manual inspection on the sampled instances and observed no disagreements between model judgments and human evaluation. Overall, these results demonstrate that our LLM-as-a-Judge setup is stable under deterministic decoding, robust across different model families, and well-aligned with human judgment.