CRiT-QA: Evaluating Multi-hop Reasoning with Counterfactual Chains and Distractor Traps

JungMin Yun[1*], **JuneHyoung Kwon**[1*], **YoungBin Kim**[1, 2]
[1] Department of Artificial Intelligence, Chung-Ang University
[2] Graduate School of Advanced Imaging Sciences, Multimedia and Film, Chung-Ang University
{cocoro357, dirchdmltnv, ybkim85}@cau.ac.kr

Abstract

Evaluating the multi-hop reasoning capabilities of large language models remains a significant challenge. Although current models achieve strong results on existing multi-hop question answering datasets, such performance often masks two critical vulnerabilities: (1) reliance on internal parametric knowledge rather than adherence to the provided context, and (2) exploitation of dataset shortcuts, such as single-document cues or type-matching, that diminish the need for genuine evidence aggregation across multiple documents. We introduce CRiT-QA (Counterfactual Reasoning with Traps), a dataset explicitly designed to address both limitations. To neutralize reliance on memorized knowledge and enforce strict context dependency, CRiT-QA transforms factual reasoning chains with counterfactual entities. Furthermore, it injects multi-anchor distractor chains, plausible but incorrect reasoning paths that diverge at different hops. These traps require models to follow the entire reasoning process rather than exploiting shallow heuristics. Our experiments show that LLMs exhibit substantial performance degradation on CRiT-QA compared to standard datasets, exposing their vulnerability to counterfactual conditions and distractor traps. CRiT-QA thus serves as a rigorous diagnostic tool for evaluating genuine multi-hop reasoning and provides a foundation for developing more reliable, evidence-grounded LLMs.

Keywords: Question Answering, Multi-Hop Reasoning, Large Language Models

1. Introduction

Retrieval-Augmented Generation (RAG) has become a dominant approach for enhancing large language models (LLMs) by grounding outputs in external knowledge sources rather than relying solely on internal parametric memory (Lewis et al., 2020; Fan et al., 2024). This paradigm has demonstrated promising improvements in factuality and adaptability across diverse domains (Tang and Yang, 2024; Gao et al., 2024). The effectiveness of RAG systems, however, critically depends on the model's ability to identify, link, and aggregate multiple pieces of evidence distributed across different documents (Fang et al., 2024; Suryawanshi et al., 2025).

Multi-hop Question Answering (QA) has thus become a standard task for evaluating the higher-order reasoning capabilities of LLMs. In principle, such datasets are designed to assess whether models can identify and synthesize intermediate evidence to derive logically coherent final answers (Kim et al., 2024; Li et al., 2024; Liu et al., 2025). However, recent studies have revealed that the strong performance of LLMs on existing multi-hop QA datasets does not necessarily reflect genuine reasoning ability (Wu et al., 2024a; Jiang et al., 2024; Parmar et al., 2024). Instead, models often exploit artifacts in the dataset or rely on shallow surface-level patterns, thereby inflating their performance without demonstrating robust reasoning (Schlegel et al., 2020; Trivedi et al., 2022; Suryawanshi et al., 2025).

From the model perspective, LLMs often bypass the provided context and instead rely on internally memorized parametric knowledge acquired during pretraining (Wu et al., 2025; Bi et al., 2025; Cheng et al., 2024). This tendency undermines the validity of the evaluation, as models may generate correct

(A)	**Question:** Who founded the company that distributed the film UHF? —— *without any context* —— **LLM-generated Answer:** Mike Medavoy ✓
(B)	**Question:** In which **county** is Mark Dismore's birthplace located? **Paragraph 1:** Greenfield is a city in and the county seat of **Hancock County**, Indiana, United States, and a part of the Indianapolis metropolitan area. (...) **LLM-generated Answer:** Hancock County ✓ ***Sub-question:*** *What is the place of birth of Mark Dismore?* ***LLM-generated Answer:*** ***Unanswerable***

Table 1: Examples illustrating evaluation vulnerabilities in multi-hop QA: (A) an LLM answering correctly without context, and (B) a model exploiting single-paragraph cues to bypass the full multi-hop reasoning process.

*Equal contribution.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5246–5255
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

answers without engaging with the evidence. For example, as illustrated in Table 1(A), an LLM can answer "Who founded the company that distributed the film *UHF*?" correctly as "Mike Medavoy," even when no context is supplied, indicating reliance on internal memory rather than reasoning over the provided passages. Moreover, large-scale pretraining poses risks of data leakage, leading to contamination that further compromises the reliability of reasoning evaluation.

On the dataset side, existing benchmarks contain shortcuts that enable models to arrive at correct answers without completing the full reasoning process (Guo et al., 2023; Yang et al., 2024; Ho et al., 2023). A question phrased with a type cue such as "which person" can sometimes be solved by selecting the only person entity in a passage; in other cases, a single paragraph suffices, and cross-document aggregation is unnecessary. Table 1(B) demonstrates this failure case for a question about "Mark Dismore's birthplace," where the model produces "Hancock County" by matching the "county" type in the question to a surface string in the first paragraph, while failing to resolve the requisite intermediate sub-question about the actual place of birth. Such artifacts encourage heuristic pattern matching and obscure whether models truly perform multi-hop reasoning.

To address these limitations, we introduce CRiT-QA (Counterfactual Reasoning with Traps), a dataset designed to rigorously evaluate LLM reasoning under counterfactual and distractor conditions. CRiT-QA targets both weaknesses in a systematic manner. First, to reduce dependence on internal knowledge, factual reasoning chains are transformed using counterfactual entities so that correct answers cannot be recalled from pretraining memory and must be inferred strictly from the given context. Second, to overcome dataset shortcuts, CRiT-QA injects multi-anchor distractor traps, plausible but ultimately incorrect reasoning paths that diverge at different hops while maintaining type consistency via named entity recognition. These distractors prevent models from relying on shallow cues and require stepwise verification of the evidence path.

Our experiments reveal that existing LLMs experience substantial performance degradation on CRiT-QA, standing in sharp contrast to their strong results on traditional datasets. This drop is consistent across both proprietary and open-source models and becomes more severe as the required reasoning chain length increases. For example, even the strongest proprietary systems, such as `GPT-4o` and `Gemini-2.5-Pro`, show marked declines when moving from 2-hop to 4-hop questions. The performance collapse is even more dramatic for open-source models like `LLaMA-3-8B` and `Qwen2.5-7B`, whose accuracy is reduced by more than half. These findings indicate that CRiT-QA successfully stresses models beyond surface-level recall, exposing their vulnerability to counterfactual reasoning constraints and the increasing density of distractor traps as reasoning depth grows. In summary, our main contributions are as follows:

1. We propose CRiT-QA, a diagnostic multi-hop QA dataset that simultaneously addresses two major limitations of existing evaluations: over-reliance on internal parametric knowledge and the exploitation of dataset shortcuts.

2. We design an automated data construction pipeline leveraging LLMs to transform existing datasets into counterfactual versions and to generate multi-anchor distractor chains, thereby creating more challenging contexts for evaluating reasoning.

3. We empirically demonstrate that state-of-the-art LLMs, despite excelling on standard datasets, substantially underperform on CRiT-QA, underscoring the persistent gap between surface-level success and genuine multi-hop reasoning ability.

Ultimately, CRiT-QA serves as a rigorous diagnostic tool for evaluating genuine multi-hop reasoning in LLMs. By exposing critical vulnerabilities to counterfactuals and distractors, it provides a robust foundation for developing more reliable, evidence-grounded language models.

2. Related Work

Multi-hop QA Datasets. Multi-hop QA datasets have been widely adopted as a standard benchmark for evaluating a model's ability to perform complex reasoning by integrating evidence from multiple sources (Press et al., 2023; Geva et al., 2021; Tang and Yang, 2024). HotpotQA (Yang et al., 2018) requires multi-hop reasoning across linked Wikipedia articles, with annotated supporting facts that enable explainable evaluation. 2WikiMultiHopQA (Ho et al., 2020) systematically pairs two Wikipedia pages to focus on structured cross-document reasoning, while MuSiQue (Trivedi et al., 2022) constructs multi-hop questions from single-hop components, providing explicit decompositions into sub-questions with intermediate answers.

Reasoning under Distractors. Although these datasets have substantially advanced multi-hop QA evaluation, subsequent analyzes have revealed that models often exploit dataset-specific artifacts, such as type matching or single-paragraph cues, or rely on memorized parametric knowledge (Bhuiya

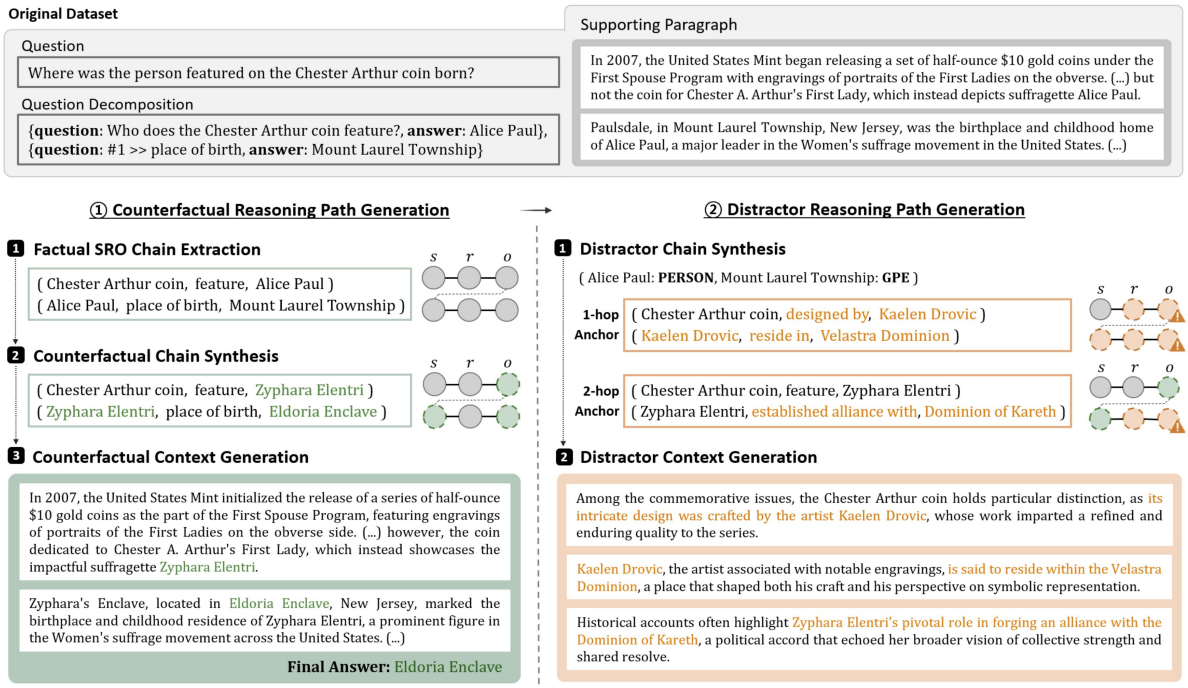

Figure 1: Overall pipeline of the proposed CRiT-QA dataset construction.

et al., 2024; Li et al., 2024; Schnitzler et al., 2024). As a result, models can achieve inflated performance without genuine multi-hop reasoning. For instance, many questions can be solved through simple pattern matching or even one-hop shortcuts, as question terms frequently overlap with tokens in the answer sentences (Schlegel et al., 2020; Bhuiya et al., 2024).

To mitigate these weaknesses, several datasets have been designed to deliberately introduce distractions into the reasoning process. One of the works (Jiang and Bansal, 2019) proposes adversarial data construction strategies that compel models to follow the correct reasoning chain, while other works (Ding et al., 2021) investigate reasoning robustness under fact-level conflicts. ClashEval (Wu et al., 2024b), for example, perturbs retrieved passages with subtle or overt errors to quantify the interplay between a model's internal knowledge and external evidence. Another study (Bhuiya et al., 2024) augments multi-hop datasets by injecting semantically similar distractor paragraphs, creating plausible yet misleading reasoning paths to evaluate robustness.

Reasoning under Counterfactual Knowledge. Beyond distractor-based evaluation, a complementary line of research investigates reasoning under counterfactual conditions. This paradigm probes whether models can prioritize contextual reasoning over potentially conflicting parametric knowledge (Frohberg and Binder, 2022). By introducing hypothetical alternatives to established facts, counterfactual scenarios challenge models to perform genuine reasoning rather than mere factual recall (Chen et al., 2025). DisentQA (Neeman et al., 2023) adopts counterfactual data augmentation to disentangle parametric and contextual knowledge, although it remains limited to single-hop settings. Other works manipulate the question itself: CREPE (Yu et al., 2023b) incorporates false presuppositions into questions to evaluate whether models can detect and correct them, and IFQA (Yu et al., 2023a) embeds counterfactual presuppositions directly in questions, requiring models to integrate these assumptions with retrieved factual evidence in an open-domain setting. While these approaches provide valuable diagnostic probes, they are confined to local, question-level edits rather than assessing reasoning consistency across an entire multi-hop reasoning chain. Closest to our work, CofCA (Wu et al., 2025) leverages an LLM to rewrite original documents with counterfactual content and subsequently generates new multi-hop QA pairs from the modified context. However, CofCA focuses on regenerating new questions aligned with rewritten documents, while we deliberately retain the original multi-hop questions and systematically replace the entire underlying reasoning chain along with its supporting evidence, ensuring that the original compositional reasoning demands remain intact. Moreover, we combine counterfactual chain replacement with multi-anchor distractors within a single unified framework, enabling a controlled, incremental escalation of reasoning difficulty that neither challenge alone can provide.

3. Dataset Construction Pipeline

We construct CRiT-QA through a multi-stage, LLM-driven pipeline designed to rigorously evaluate reasoning in multi-hop QA under both counterfactual and distracting conditions. Within this pipeline, factual reasoning chains are transformed into counterfactual and distractor variants, ensuring that models engage in authentic reasoning processes while preventing reliance on shallow inference driven by memorized associations or surface-level patterns.

3.1. Counterfactual Reasoning Path

Factual SRO Chain Extraction. We begin with the MusiQue dataset (Trivedi et al., 2022), which is suitable for our setting, as it explicitly provides step-by-step decompositions for multi-hop questions. Each sample consists of a multi-hop question Q, a final factual answer A_{fact}, a set of supporting paragraphs P, and a reasoning decomposition D. The decomposition consists of an ordered sequence of sub-questions q_i and their corresponding sub-answers a_i. We first filter the paragraphs to retain only the gold evidence paragraphs P_{fact} that support this decomposition. We then employ an LLM to transform the decomposition D into a structured Subject–Relation–Object (SRO) chain C_{fact}. This chain is a sequence of N hops, $C_{fact} = \{h_1, h_2, \ldots, h_N\}$, where each hop h_i is a triple (s_i, r_i, o_i). The relation r_i is derived from the sub-question q_i, and the object o_i corresponds to the sub-answer a_i. The subject s_i is either an entity from the original question Q or the object o_k $(k < i)$ of a previous hop, explicitly linking the reasoning steps. This LLM-guided mapping enables us to construct an explicit and coherent reasoning chain C_{fact} that faithfully captures the stepwise, and potentially non-linear, reasoning trajectory.

Counterfactual Chain Synthesis. To introduce systematic counterfactual variation, we employ an LLM to transform the factual chain C_{fact} into a corresponding counterfactual chain $C_{cf} = \{h_1^{(c)}, h_2^{(c)}, \ldots, h_N^{(c)}\}$, where each hop is $h_i^{(c)} = (s_i^{(c)}, r_i^{(c)}, o_i^{(c)})$. This synthesis is guided by the following principles to ensure coherence and logical consistency. First, each factual object o_i is replaced by a newly generated fictional counterfactual object $o_i^{(c)}$. Second, subjects originating from the original question Q and all relations are preserved to maintain structural alignment and logical flow. Third, if the subject s_i corresponds to the object o_k of a preceding hop, the new subject $s_i^{(c)}$ is set to the corresponding counterfactual object $o_k^{(c)}$, ensuring the internal consistency of the reasoning chain.

Counterfactual Context Generation. To provide coherent contextual grounding, we prompt an LLM to rewrite the supporting paragraphs P_{fact} into counterfactual contexts P_{cf}. This rewriting process goes beyond simple entity substitution: the model adjusts syntax, discourse flow, and narrative structure to ensure grammatical correctness, stylistic coherence, and logical alignment with the counterfactual reasoning chain C_{cf}. The resulting passages P_{cf} seamlessly integrate fictional entities into the narrative, producing a self-consistent and contextually grounded representation. In addition, the LLM generates the counterfactual final answer A_{cf} in alignment with the newly constructed P_{cf} and reasoning chain C_{cf}, ensuring consistency across all components of the sample.

At the conclusion of this stage, each sample is represented by four components: the original question Q, the counterfactual reasoning chain C_{cf}, the rewritten contexts P_{cf}, and the counterfactual final answer A_{cf}.

3.2. Distractor Reasoning Path

To evaluate model robustness against misleading alternatives, we generate distractor reasoning chains that diverge from the correct counterfactual chain. This stage generates alternative reasoning paths that closely mimic the structural form of valid reasoning chains but deliberately diverge at specific hops, thereby producing coherent yet ultimately incorrect reasoning trajectories.

Distractor Chain Synthesis. We prompt the LLM to generate a set of distractor chains, each diverging from the correct counterfactual chain at a different hop. Given an N-hop counterfactual chain C_{cf}, we construct N distinct distractor chains $C_{dist}^{(j)}$ (where $j \in [1, N]$), with hop j serving as its initial divergence point. Each distractor hop is denoted as $h_i^{(d)} = (s_i^{(d)}, r_i^{(d)}, o_i^{(d)})$ and is generated in a hop-by-hop manner by replacing counterfactual elements with plausible but misleading alternatives. To ensure type consistency, we adopt an NER-driven strategy: the LLM analyzes the entity type of each object (e.g., person, location, date) and generates a new distractor object of the same type. Relations are modified to be semantically distinct from the originals. The subject is either preserved or replaced with the corresponding distractor object from a preceding hop, thereby maintaining coherent linkage across the reasoning chain. This approach yields distractor paths that are structurally valid and contextually plausible, yet ultimately lead to incorrect reasoning trajectories unrelated to the correct answer.

	CRiT-QA (Ours)		*MuSiQue*		*2WikiMultiHopQA*		*HotpotQA*	
	EM	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**
`LLaMA-3-8B`	16.94	23.29	24.74	36.52	35.06	44.52	38.51	52.63
`Longchat-13B-16k`	11.14	17.83	18.54	31.39	26.91	38.58	30.24	44.37
`Mistral-7B-Instruct`	18.55	31.26	27.51	46.98	35.37	50.86	44.62	63.45
`Qwen2.5-7B`	19.30	25.20	34.96	47.36	42.76	50.65	52.02	65.54
`GPT-3.5-Turbo`	30.93	40.51	39.68	55.28	50.02	62.13	57.30	72.90
`GPT-4o`	39.59	48.52	47.66	63.69	68.66	80.22	63.70	79.98
`Gemini-2.5-Flash`	38.30	46.24	55.56	69.37	71.49	79.99	66.21	81.27
`Gemini-2.5-Pro`	44.27	53.89	59.08	73.99	73.46	82.90	66.28	81.50
`Claude Haiku 4.5`	30.02	37.45	48.37	63.62	58.06	69.87	59.11	75.89
`Claude Sonnet 4.5`	40.25	48.96	54.86	70.66	68.62	77.89	64.16	80.42

Table 2: Performance comparison of LLMs across CRiT-QA and three standard multi-hop QA benchmarks.

Distractor Context Generation. To provide textual grounding for the distractor chains, we generate additional paragraphs corresponding to their unique hops. Specifically, we identify the set of all unique distractor hops that appear in any distractor chain but not in the correct counterfactual chain. For each such hop, we prompt an LLM to generate a short, coherent paragraph that provides supporting evidence. The resulting set of distractor paragraphs P_{dist} is then combined with the counterfactual paragraphs P_{cf} to form the final context $P_{final} = P_{cf} \cup P_{dist}$. This integrated context introduces multiple competing facts that are all textually plausible, compelling models to carefully discriminate between correct and misleading reasoning paths without relying on simple surface-level patterns.

4. Experiments

4.1. Experimental Setup

We evaluate LLMs on our constructed CRiT-QA dataset using a diverse set of models. For open-source models, we include `LLaMA-3-8B` (Dubey et al., 2024), `Qwen2.5-7B` (Qwen et al., 2025), `Longchat-13B-16k`[1], and `Mistral-7B-Instruct` (Jiang et al., 2023). For API-based commercial models, we evaluate `GPT-3.5-Turbo` (Brown et al., 2020), `GPT-4o` (Hurst et al., 2024), `Gemini-2.5-Flash`, `Gemini-2.5-Pro` (Comanici et al., 2025), `Claude Haiku 4.5`, and `Claude Sonnet 4.5`. To assess multi-hop QA performance, we report both Exact Match (EM) and F1 scores. In all experiments, we explicitly instruct the LLMs to derive their final answers based solely on the provided context, thereby reducing the influence of their internal parametric knowledge.

[1]https://huggingface.co/lmsys/longchat-13b-16k

4.2. Experimental Results

In Table 2, we report the performance of various LLMs on CRiT-QA in comparison to three widely used multi-hop QA datasets[2]: 2WikiMultiHopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and HotpotQA (Yang et al., 2018). We note that this comparison is not intended as a strict head-to-head evaluation across datasets, given the inherent structural differences across datasets. Rather, its purpose is to illustrate a broader trend: models that achieve strong performance on standard multi-hop benchmarks exhibit substantial performance degradation when faced with the counterfactual and distractor-based challenges introduced by CRiT-QA.

Main Results. We observe a consistent decline across all evaluated models, including both open-source and API-based commercial systems. This universal performance gap highlights a fundamental vulnerability in current reasoning capabilities. For instance, `Gemini-2.5-Pro` achieves the highest performance on CRiT-QA (44.27 EM and 53.89 F1); however, this still represents a substantial drop from its scores on MuSiQue (59.08 EM) and 2WikiMultiHopQA (73.46 EM). Similarly, `GPT-4o` and `Claude Sonnet 4.5` fall from 47.66 and 54.86 EM on MuSiQue to 39.59 and 40.25 EM on CRiT-QA, respectively.

Furthermore, this vulnerability is even more pronounced among open-sourced models. For instance, `Qwen2.5-7B` attains 52.02 EM on HotpotQA and 34.96 EM on MuSiQue but only 19.30 EM on CRiT-QA. Comparable declines are consistently observed across the remaining models as well. These results collectively underscore that

[2]All experiments were conducted on the dev sets of the respective datasets. For MuSiQue, we used the *musique_ans_v1.0_dev*, which contains only answerable QA pairs.

Setting	(Qwen2.5-7B) EM	(Qwen2.5-7B) F1	(Gemini-2.5-Pro) EM	(Gemini-2.5-Pro) F1
Oracle ($Q + P_{fact}$)	34.96	47.36	59.08	73.99
with Counterfactuals (P_{cf})	24.77 (↓10.19)	31.18 (↓16.18)	50.86 (↓8.22)	62.02 (↓11.97)
with Distractors ($P_{cf} + P_{dist}$)	19.30 (↓5.47)	25.20 (↓5.98)	47.74 (↓3.12)	57.33 (↓4.69)
without Context (only Q*)*	0.00	1.27	0.00	1.74

Table 3: Ablation results on CRiT-QA under different evidence settings.

	2-hop		*3-hop*		*4-hop*	
	EM	**F1**	**EM**	**F1**	**EM**	**F1**
LLaMA-3-8B	20.88	28.91	15.53	21.05	7.41	10.18
Longchat-13B-16k	13.84	22.69	10.13	14.87	5.93	9.11
Mistral-7B-Instruct	22.08	35.55	15.13	28.26	14.07	23.64
Qwen2.5-7B	22.80	26.69	16.97	22.49	12.84	16.44
GPT-3.5-Turbo	34.80	44.82	30.13	40.18	20.49	26.24
GPT-4o	43.84	52.78	36.71	47.87	31.85	36.53
Gemini-2.5-Flash	40.56	49.63	38.29	46.54	31.36	35.22
Gemini-2.5-Pro	44.89	55.56	45.13	55.03	40.74	46.48
Claude-4.5-Haiku	32.16	41.12	29.87	36.27	23.70	28.32
Claude-4.5-Sonnet	43.52	52.82	39.74	49.47	31.11	36.07

Table 4: Performance comparison on CRiT-QA according to hop length.

existing models still struggle with the complex multi-hop reasoning required to navigate counterfactuals and distractors.

4.3. Ablation Study

To better analyze the distinct challenges posed by CRiT-QA, we conduct an ablation study, with results presented in Table 3. We evaluate model performance by systematically decomposing the context into four settings:

- *Oracle* ($Q + P_{fact}$): The baseline configuration, where the original question Q is paired with the factual gold paragraphs P_{fact} from the source dataset.
- *with Counterfactuals* (P_{cf}): The counterfactual configuration, in which factual paragraphs P_{fact} are replaced with generated counterfactual context P_{cf}, isolating the challenge of reasoning against parametric knowledge.
- *with Distractors* ($P_{cf} + P_{dist}$): The full CRiT-QA setting, combining counterfactual context P_{cf} with multi-anchor distractor paragraphs P_{dist}.
- *without Context (only* Q*)*: A parametric-only probe to confirm that the answer cannot be recalled from memory.

The results demonstrate the compounded difficulty introduced by each component of CRiT-QA.

Impact of Counterfactuals. Transitioning from the *Oracle* setting (59.08 EM for Gemini-2.5-Pro, 34.96 EM for Qwen2.5-7B) to the counterfactual setting P_{cf} leads to substantial degradation in performance. Gemini-2.5-Pro drops by 8.22 EM points to 50.86, while Qwen2.5-7B drops by 10.19 points to 24.77. This reduction highlights the difficulty of overriding memorized factual knowledge and adhering to counterfactual evidence.

Impact of Distractors. The addition of distractor paragraphs P_{dist} in the full CRiT-QA configuration ($P_{cf} + P_{dist}$) further exacerbates the performance declines. The EM score of Gemini-2.5-Pro falls by an additional 3.12 points to 47.74, while Qwen2.5-7B drops by 5.47 points to 19.30. These results demonstrate the effectiveness of multi-anchor distractor chains in misleading models and underscore the need for more robust evidence-based reasoning.

Context Dependency. The *without context* setting confirms dataset integrity. Both models record 0.00 EM (with negligible F1 scores of 1.27 and 1.74, respectively), verifying that correct counterfactual

answers cannot be retrieved from parametric memory alone.

Overall, the ablation study demonstrates that performance degradation on CRiT-QA arises not from a single factor but from the compounding challenge of (i) reasoning under counterfactual conditions that conflict with internal knowledge and (ii) resisting plausible multi-hop distractor paths.

4.4. Analysis by Reasoning Hop Length

To further investigate the reasoning limitations of current models, we analyze performance on CRiT-QA by breaking down results according to the reasoning chain length (2-hop, 3-hop, and 4-hop), as shown in Table 4.

This analysis is particularly important given our multi-anchor distractor methodology. In our pipeline, an N-hop question is paired with N distinct distractor reasoning chains, each anchored at a different hop. Consequently, a 4-hop question requires not only a longer inference chain but also introduces a larger number of plausible distractor paragraphs P_{dist} than 2-hop or 3-hop questions. Thus, task difficulty is intentionally designed to increase with reasoning length.

The results in Table 4 empirically confirm this design, showing a consistent negative correlation between the number of hops and model performance. Across all evaluated models, accuracy is lowest on 4-hop questions. As reasoning chains lengthen and the distractors accumulate, models become increasingly prone to failure.

Even the most advanced models are not immune to this degradation. For instance, the F1 score of `GPT-4o` declines from 52.78 (2-hop) to 47.87 (3-hop) and 36.53 (4-hop). `Claude Sonnet 4.5` shows a similar trend, with its F1 score dropping from 52.82 (2-hop) to 36.07 (4-hop). `Gemini-2.5-Pro`, while maintaining relatively stable performance between 2-hop (44.89 EM) and 3-hop (45.13 EM) tasks, still exhibits a notable decrease on 4-hop questions (40.74 EM). The decline is even sharper among open-source models. The F1 score of `LLaMA-3-8B` falls from 28.91 (2-hop) to 10.18 (4-hop), and the EM of `Qwen2.5-7B` is nearly halved, dropping from 22.80 (2-hop) to 12.84 (4-hop).

These results highlight that models are not only challenged by longer inference chains but also increasingly overwhelmed by the growing density of distractor traps. The cognitive load of simultaneously validating the correct multi-hop path while rejecting multiple plausible, type-consistent distractors represents a key failure point of current LLMs.

5. Conclusion

In this paper, we address critical shortcomings in existing multi-hop QA evaluations, where high performance often conceals a reliance on parametric knowledge and dataset-specific shortcuts. We introduce CRiT-QA (Counterfactual Reasoning with Traps), a dataset designed to rigorously evaluate genuine, evidence-driven reasoning by neutralizing memorized knowledge with counterfactual entities and suppressing shallow heuristics through multi-anchor distractor traps. Our experiments empirically validate the effectiveness of CRiT-QA. We demonstrate that current LLMs experience substantial performance degradation on CRiT-QA, standing in sharp contrast to their strong results on standard benchmarks. An ablation study confirms that this degradation arises from the dual challenges of counterfactual adherence and distractor filtering, and that difficulty compounds with increased reasoning depth and distractor density. These findings underscore the persistent gap between surface-level success and genuine multi-hop reasoning. CRiT-QA thus serves as a robust diagnostic tool for exposing these weaknesses and provides a foundation for developing and evaluating next-generation, evidence-grounded LLMs.

6. Limitations & Future Directions

While CRiT-QA provides a robust framework for evaluating multi-hop reasoning, we emphasize that its primary objective is not to establish a static benchmark but to diagnose the surface-level heuristics and memorized parametric knowledge current models rely on. With this diagnostic objective, we identify the following limitations and directions for future research: (i) *Dependency on a Single Base Dataset.* CRiT-QA is currently built upon the MuSiQue dataset, inheriting its specific reasoning structures. However, its fully automated construction pipeline can be readily extended to additional multi-hop benchmarks to encompass a broader range of domains. This automation also enables periodic re-execution to generate new counterfactual entities, effectively mitigating the risk of data contamination. (ii) *Potential Generator Artifacts.* The pipeline relies on an LLM to synthesize both counterfactual contexts and distractor paragraphs, introducing the risk of stylistic artifacts. Models might exploit these as unintended shortcuts instead of performing faithful reasoning. Future work could explore more rigorous artifact-filtering mechanisms to further isolate the reasoning capabilities of evaluated models.

7. Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)] and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00556246).

8. Bibliographical References

Neeladri Bhuiya, Viktor Schlegel, and Stefan Winkler. 2024. Seemingly plausible distractors in multi-hop reasoning: Are large language models attentive readers? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2514–2528.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. 2025. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *arXiv preprint arXiv:2503.15888*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yuefei Chen, Vivek K. Singh, Jing Ma, and Ruxiang Tang. 2025. Counterbench: A benchmark for counterfactuals reasoning in large language models.

Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. 2024. Understanding the interplay between parametric and contextual knowledge for large language models. *arXiv preprint arXiv:2410.08414*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jiayu Ding, Siyuan Wang, Qin Chen, and Zhongyu Wei. 2021. Reasoning chain based adversarial attack for multi-hop question answering. *arXiv preprint arXiv:2112.09658*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. 2024. TRACE the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8472–8494, Miami, Florida, USA. Association for Computational Linguistics.

Jörg Frohberg and Frank Binder. 2022. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

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

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9.

Wangzhen Guo, Qinkang Gong, Yanghui Rao, and Hanjiang Lai. 2023. Counterfactual multihop QA: A cause-effect approach for reducing disconnected reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4214–4226, Toronto, Canada. Association for Computational Linguistics.

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.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2023. Analyzing the effectiveness of the underlying reasoning tasks in multi-hop question answering. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1163–1180, Dubrovnik, Croatia. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo Jose Taylor, and Dan Roth. 2024. A peek into token bias: Large language models are not yet genuine reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4722–4756, Miami, Florida, USA. Association for Computational Linguistics.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. Association for Computational Linguistics.

Jeonghoon Kim, Heesoo Jung, Hyeju Jang, and Hogun Park. 2024. Improving multi-hop logical reasoning in knowledge graphs with context-aware query representation learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15978–15991, Bangkok, Thailand. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Ruosen Li, Zimu Wang, Son Tran, Lei Xia, and Xinya Du. 2024. Meqa: A benchmark for multi-hop event-centric question answering with explanations. *Advances in Neural Information Processing Systems*, 37:126835–126862.

Hao Liu, Zhengren Wang, Xi Chen, Zhiyu Li, Feiyu Xiong, Qinhan Yu, and Wentao Zhang. 2025. Hoprag: Multi-hop reasoning for logic-aware retrieval-augmented generation.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707, Bangkok, Thailand. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

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.

Viktor Schlegel, Marco Valentino, Andre Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A framework for evaluation of machine reading comprehension gold standards. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5359–5369, Marseille,

France. European Language Resources Association.

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

Sudhanshu Suryawanshi, Shreyas Waghmode, Ritesh Sawant, and Megha Gupta. 2025. A knowledge graph-based rag for cross-document information extraction. In *2025 5th International Conference on Pervasive Computing and Social Networking (ICPCSN)*, pages 1401–1406. IEEE.

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

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

Jian Wu, Linyi Yang, Manabu Okumura, and Yue Zhang. 2024a. Mrke: The multi-hop reasoning evaluation of llms by knowledge edition. *arXiv preprint arXiv:2402.11924*.

Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. 2025. Cofca: A step-wise counterfactual multi-hop qa benchmark. In *The Thirteenth International Conference on Learning Representations*.

Kevin Wu, Eric Wu, and James Zou. 2024b. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *Advances in neural information processing systems*, 37:33402–33422.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

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.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023a. IfQA: A dataset for open-domain question answering under counterfactual presuppositions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8276–8288, Singapore. Association for Computational Linguistics.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023b. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.