SOVEREIGN: Cofca: A Step-Wise Counterfactual Multi-hop QA benchmark

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

While Large Language Models (LLMs) excel in question-answering (QA) tasks, their real reasoning abilities on multiple evidence retrieval and integration on Multi-hop QA tasks remain less explored. Firstly, LLMs sometimes generate answers that rely on internal memory rather than retrieving evidence and reasoning in the given context, which brings concerns about the evaluation quality of real reasoning abilities. Although previous counterfactual QA benchmarks can separate the internal memory of LLMs, they focus solely on final QA performance, which is insufficient for reporting LLMs' real reason

1 Introduction

Analysis of: Cofca: A Step-Wise Counterfactual Multi-hop QA benchmark. Research goal: How does the inference throughput (tokens per second) of increasing context window size from 4K to 128K compare to the throughput of adding a multi-step retrieval pipeline (e.g., 2-5 retrieval steps) for multi-hop QA on HotPotQA, measured with LLMs like Llama-3 or GPT-4?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

6 papers retrieved. 8 claims extracted, 8 verified. Tribunal: 7.5/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Large Language Models (LLMs) excel in question-answering (QA) tasks.	✓	0.25
LLMs sometimes generate answers that rely on internal memory rather than retrieving evidence and reasoning in the given	✓	0.31
Previous counterfactual QA benchmarks focus solely on final QA performance, which is insufficient for reporting LLMs' re	✓	0.36
Current factual Multi-hop QA (MHQA) benchmarks are annotated on open-source corpora such as Wikipedia.	✓	0.32
Wikipedia-based benchmarks show limitations due to potential data contamination in LLMs' pre-training stage.	✓	0.25
CofCA is a novel evaluation benchmark consisting of factual data and counterfactual data that reveals LLMs' real reasoni	✓	0.48
Experimental results reveal a significant performance gap of several LLMs between Wikipedia-based factual data and count	✓	0.30
The performance gap between factual and counterfactual data deems data contamination issues in existing benchmarks.	✓	0.22

References

- http://arxiv.org/abs/2604.09019v2
- http://arxiv.org/abs/2402.11924v5
- http://arxiv.org/abs/2404.14464v1