SOVEREIGN: What is the accuracy drop on the HotpotQA multi-hop dataset when using a 128K-context Llama-3 model without re

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Prompting-based large language models (LLMs) are surprisingly powerful at generating natural language reasoning steps or Chains-of-Thoughts (CoT) for multi-step question answering (QA). They struggle, however, when the necessary knowledge is either unavailable to the LLM or not up-to-date within its parameters. While using the question to retrieve relevant text from an external knowledge source helps LLMs, we observe that this one-step retrieve-and-read approach is insufficient for multi-step QA. Here, what to retrieve depends on what has already been derived, which in turn may depend on what

1 Introduction

Analysis of: Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. Research goal: What is the accuracy drop on the HotpotQA multi-hop dataset when using a 128K-context Llama-3 model without retrieval versus a 4K-context model with 2-step retrieval, controlling for total token budget?.

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

9 papers retrieved. 7 claims extracted, 7 verified. Tribunal: 7.7/10 → 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
Prompting-based large language models (LLMs) are surprisingly powerful at generating natural language reasoning steps or	✓	0.43
LLMs struggle when the necessary knowledge is either unavailable to the LLM or not up-to-date within its parameters.	✓	0.22
Using the question to retrieve relevant text from an external knowledge source helps LLMs, but this one-step retrieve-an	✓	0.42
IRCoT is a new approach for multi-step QA that interleaves retrieval with steps (sentences) in a CoT, guiding the retrie	✓	0.47
Using IRCoT with GPT3 substantially improves retrieval (up to 21 points) as well as downstream QA (up to 15 points) on f	✓	0.38
Similar substantial gains in out-of-distribution (OOD) settings as well as with much smaller models such as Flan-T5-larg	✓	0.29
IRCoT reduces model hallucination, resulting in factually more accurate CoT reasoning.	✓	0.29

References

- https://doi.org/10.48550/arxiv.2312.10997
- https://doi.org/10.1109/access.2021.3140175
- https://doi.org/10.18653/v1/2023.acl-long.557