SOVEREIGN: How does the Tree of Reviews framework compare to standard chain-based retrieval on the MuSiQue multi-hop QA b

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Multi-hop question answering is a knowledge-intensive complex problem. Large Language Models (LLMs) use their Chain of Thoughts (CoT) capability to reason complex problems step by step, and retrieval-augmentation can effectively alleviate factual errors caused by outdated and unknown knowledge in LLMs. Recent works have introduced retrieval-augmentation in the CoT reasoning to solve multi-hop question answering. However, these chain methods have the following problems: 1) Retrieved irrelevant paragraphs may mislead the reasoning; 2) An error in the chain structure may lead to a cascade of erro

1 Introduction

Analysis of: Tree of Reviews: A Tree-based Dynamic Iterative Retrieval Framework for Multi-hop Question Answering. Research goal: How does the Tree of Reviews framework compare to standard chain-based retrieval on the MuSiQue multi-hop QA benchmark in terms of accuracy and token efficiency when using a 128K-context Llama-3 model without retrieval?.

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. 5 claims extracted, 3 verified. Tribunal: 6.7/10 → REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

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
TOR achieves state-of-the-art performance in both retrieval and response generation on three different multi-hop questio	✓	0.32
The tree structure in TOR alleviates the misleading effect of irrelevant paragraphs on the reasoning path.	✓	0.24
The diversity of reasoning path extension in TOR reduces the impact of a single reasoning error on the whole.	✓	0.22
TOR is the first retrieval framework that uses a tree-like structure to dynamically initiate requests based on external	×	0.14
Pruning and effective expansion strategies in TOR reduce time overhead and increase the diversity of path extension.	×	0.04

References

- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2604.18234v1
- http://arxiv.org/abs/2404.14464v1