SOVEREIGN: Tree of Reviews: A Tree-based Dynamic Iterative Retrieval Framework for Multi-ho

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 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to reason complex problems step by step, and retrieval-augmentation can effectively alleviate factual errors caused by outdated and unknown knowledge in LLMs.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: What is the accuracy drop on counterfactual multi-hop QA (e.g., Cofca) when using a 128K-context Llama-3 model without retrieval compared to 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

11 papers retrieved. 12 claims extracted, 12 verified. Tribunal: 8.2/10 → 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
Multi-hop question answering is a knowledge-intensive complex problem.	✓	0.30
Large Language Models (LLMs) use their Chain of Thoughts (CoT) capability to reason complex problems step by step.	✓	0.30
Retrieval-augmentation can effectively alleviate factual errors caused by outdated and unknown knowledge in LLMs.	✓	0.28
Recent works have introduced retrieval-augmentation in the CoT reasoning to solve multi-hop question answering.	✓	0.36
Retrieved irrelevant paragraphs may mislead the reasoning in chain methods.	✓	0.24
An error in the chain structure may lead to a cascade of errors.	✓	0.22
The Tree of Reviews (ToR) framework uses a tree structure where the root node is the question and other nodes are paragr	✓	0.31
The ToR framework dynamically decides to initiate a new search, reject, or accept based on the paragraphs on the reasoni	✓	0.31
The tree structure in ToR handles each retrieved paragraph separately, alleviating the misleading effect of irrelevant p	✓	0.30
The diversity of reasoning path extension in ToR reduces the impact of a single reasoning error on the whole.	✓	0.25
Experiments were conducted on three different multi-hop question answering datasets.	✓	0.25
Compared to baseline methods, ToR achieves state-of-the-art performance in both retrieval and response generation.	✓	0.27

References

- http://arxiv.org/abs/2510.06426v1
- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/1912.02145v1