SOVEREIGN: To what extent does the choice of dense versus sparse retrieval method affect the correlation between BEIR rob

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to reason complex problems step by step, and retrieval-augmentation can effectively alleviate factual errors caused by outdated and unknown knowledge in LLMs.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: To what extent does the choice of dense versus sparse retrieval method affect the correlation between BEIR robustness scores and LLM multi-hop reasoning accuracy when evaluated on adversarial or out-of-distribution query variants?.

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

10 papers retrieved. 9 claims extracted, 3 verified. Tribunal: 6.3/10 → 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
The Tree of Reviews framework uses a tree-based dynamic retrieval approach for multi-hop question answering.	✓	0.29
TOR achieves state-of-the-art performance in both retrieval and response generation compared to baseline methods.	✓	0.24
The framework dynamically decides to initiate a new search, reject, or accept based on paragraphs on reasoning paths.	✓	0.30
Two tree-based search optimization strategies, pruning and effective expansion, are proposed to reduce time overhead.	×	0.03
Early RAG research employed a one-step retrieval approach that is ineffective for composite problems.	×	0.06
Self-Ask poses sub-questions before answering the main question to optimize complex composite problems.	×	0.06
Chain-like reasoning structures can deviate if an error occurs at any step in the reasoning path.	×	0.13
Tree of Thought enhances problem-solving capabilities of LLMs by introducing a tree-like structure during reasoning.	×	0.08
Static problem tree decomposition lacks assistance from external knowledge and can lead to incorrect decomposition.	×	0.04

References

- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/2104.08663v4
- http://arxiv.org/abs/2205.02303v1