Tree-Based Retrieval Stability in Multi-Hop Question Answering with Llama-3-8B-128K

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the impact of varying the number of retrieval hops (e.g., 2-hop vs. 3-hop) on the F1 score stability of the Tree of Reviews framework compared to chain-based retrieval in Llama-3-8B-128K when. 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to reason complex problems step by step, and retrieval-augmentation can effectively alleviate factual errors caused by outdated and. 9 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Tree of Reviews: A Tree-based Dynamic Iterative Retrieval Framework for Multi-hop Question Answering. Research question: What is the impact of varying the number of retrieval hops (e.g., 2-hop vs. 3-hop) on the F1 score stability of the Tree of Reviews framework compared to chain-based retrieval in Llama-3-8B-128K when evaluated on the MuSiQue benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.8/10.

3 Results

11 papers retrieved. 9 claims extracted; 3 independently verified. Quality review score: 5.8/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

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	✓	0.34
Tree of Thought (ToT) enhances the problem-solving capabilities of Large Language Models (LLMs) by introducing a tree-li	×	0.09
Asai et al. (2020) trained a retriever that dynamically retrieves information from Wikipedia graphs.	×	0.02
The decomposition of the question and the construction of the tree lack the assistance of external knowledge and informa	×	0.07
TOR is the first to propose a retrieval framework that uses a tree-like structure to dynamically initiate requests based	×	0.13
LLMs can decide dynamically whether to initiate further retrieval and what requests to generate based on external knowle	×	0.09
TOR introduces a tree structure to handle each retrieved paragraph separately, alleviating the misleading effect of irre	✓	0.30
The diversity of reasoning path extension reduces the impact of a single reasoning error on the whole.	✓	0.24
TOR proposes two tree-based search optimization strategies, pruning and effective expansion, to reduce time overhead and	×	0.07

References

- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/2604.18234v1
- http://arxiv.org/abs/1811.08772v1