SOVEREIGN: What is the impact of iterative retrieval agent depth (number of retrieval rounds) on final answer F1 score an

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

In this paper, we present Hierarchical Graph Network (HGN) for multi-hop question answering. To aggregate clues from scattered texts across multiple paragraphs, a hierarchical graph is created by constructing nodes on different levels of granularity (questions, paragraphs, sentences, entities), the representations of which are initialized with pre-trained contextual encoders. Given this hierarchical graph, the initial node representations are updated through graph propagation, and multihop reasoning is performed via traversing through the graph edges for each subsequent sub-task (e.g., paragra

1 Introduction

Analysis of: Hierarchical Graph Network for Multi-hop Question Answering. Research goal: What is the impact of iterative retrieval agent depth (number of retrieval rounds) on final answer F1 score and retrieval precision across HotpotQA, 2WikiMultihopQA, and MuSiQue datasets when evaluated with controlled distractor insertion rates from 0-50%?.

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

12 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 4.7/10 → REJECT (revision_round=0). 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.

References

- https://doi.org/10.18653/v1/d18-1259
- https://doi.org/10.1109/access.2021.3140175
- https://doi.org/10.18653/v1/2020.emnlp-main.710