SOVEREIGN: PRISM: Agentic Retrieval with LLMs for Multi-Hop Question Answering

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Retrieval plays a central role in multi-hop question answering (QA), where answering complex questions requires gathering multiple pieces of evidence. We introduce an Agentic Retrieval System that leverages large language models (LLMs) in a structured loop to retrieve relevant evidence with high precision and recall. Our framework consists of three specialized agents: a Question Analyzer that decomposes a multi-hop question into sub-questions, a Selector that identifies the most relevant context for each sub-question (focusing on precision), and an Adder that brings in any missing evidence (fo

1 Introduction

Analysis of: PRISM: Agentic Retrieval with LLMs for Multi-Hop Question Answering. Research goal: Does the robustness of LLMs to irrelevant context in multi-hop QA generalize across different reasoning benchmarks (e.g., MuSiQue, 2WikiMultihopQA) when using large context windows versus iterative retrieval, as measured by F1 and precision under controlled distractor insertion?.

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. 6 claims extracted, 6 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
The Agentic Retrieval System leverages large language models (LLs) in a structured loop to retrieve relevant evidence wi	✓	0.41
The framework consists of three specialized agents: a Question Analyzer, a Selector, and an Adder	✓	0.26
The iterative interaction between Selector and Adder agents yields a compact yet comprehensive set of supporting passage	✓	0.27
The system achieves higher retrieval accuracy while filtering out distracting content	✓	0.25
Downstream QA models can surpass full-context answer accuracy while relying on significantly less irrelevant information	✓	0.31
Experiments on four multi-hop QA benchmarks (HotpotQA, 2WikiMultiHopQA, MuSiQue, and MultiHopRAG) demonstrate that the a	✓	0.29

References

- https://www.semanticscholar.org/paper/cb882a8440581ee70d1d5006bd6e64dbd19919ec
- https://www.semanticscholar.org/paper/8f531abb2339936d3b4b20bc2af573ea35b55b1b
- http://arxiv.org/abs/2510.14278v1