# SOVEREIGN: How does PRISM framework's retrieval efficiency (latency and throughput) compare to end-to-end multi-hop QA pi

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

**Abstract**

Large Language Models (LLMs) have demonstrated significant performance improvements across various cognitive tasks. An emerging application is using LLMs to enhance retrieval-augmented generation (RAG) capabilities. These systems require LLMs to understand user queries, retrieve relevant information, and synthesize coherent and accurate responses. Given the increasing real-world deployment of such systems, comprehensive evaluation becomes crucial. To this end, we propose FRAMES (Factuality, Retrieval, And reasoning MEasurement Set), a high-quality evaluation dataset designed to test LLMs' abil

## 1 Introduction

Analysis of: Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation. Research goal: How does PRISM framework's retrieval efficiency (latency and throughput) compare to end-to-end multi-hop QA pipeline response time across different LLM architectures (7B, 13B, 30B parameter models) on MuSiQue and 2WikiMultihopQA benchmarks, as measured by F1 scores and inference time?.

## 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

3 papers retrieved. 5 claims extracted, 2 verified. Tribunal: 6.2/10 → 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 FRAMES dataset evaluates LLMs' ability to retrieve and reason across multiple documents. | × | 0.10 |
| The baseline results show that state-of-the-art LLMs struggle, achieving 0.408 accuracy without retrieval. | ✓ | 0.17 |
| The proposed multi-step retrieval pipeline significantly improves accuracy from 0.408 to 0.66. | × | 0.14 |
| The dataset comprises challenging multi-hop questions. | ✓ | 0.16 |
| The average number of documents in the context is 2, and accuracy improves to 0.474 when double the number of articles a | × | 0.01 |

## References


- `https://arxiv.org/abs/2409.12941`
- `https://arxiv.org/abs/2212.09292`
- `https://arxiv.org/abs/2508.13992`