SOVEREIGN: How does the end-to-end latency of RAG systems with 128K context windows compare to iterative retrieval with B

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Retrieval augmented generation (RAG) with large language models (LLMs) for Question Answering (QA) entails furnishing relevant context within the prompt to facilitate the LLM in answer generation. During the generation, inaccuracies or hallucinations frequently occur due to two primary factors: inadequate or distracting context in the prompts, and the inability of LLMs to effectively reason through the facts. In this paper, we investigate whether providing aligned context via a carefully selected passage sequence leads to better answer generation by the LLM for multi-hop QA. We introduce, "Gen

1 Introduction

Analysis of: GenSco: Can Question Decomposition based Passage Alignment improve Question Answering?. Research goal: How does the end-to-end latency of RAG systems with 128K context windows compare to iterative retrieval with BM25 re-ranking in terms of exact match scores on HotpotQA when evaluated with adversarial distractors?.

2 Methodology

Multi-query arXiv search (1 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

1 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 5.0/10 → 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.

References

- https://doi.org/10.48550/arxiv.2407.10245