SOVEREIGN: What is the computational cost (in FLOPs or latency) versus F1 score trade-off when scaling context windows fr

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Large Language Models (LLMs) showcase impressive capabilities but encounter challenges like hallucination, outdated knowledge, and non-transparent, untraceable reasoning processes. Retrieval-Augmented Generation (RAG) has emerged as a promising solution by incorporating knowledge from external databases. This enhances the accuracy and credibility of the generation, particularly for knowledge-intensive tasks, and allows for continuous knowledge updates and integration of domain-specific information. RAG synergistically merges LLMs' intrinsic knowledge with the vast, dynamic repositories of exte

1 Introduction

Analysis of: Retrieval-Augmented Generation for Large Language Models: A Survey. Research goal: What is the computational cost (in FLOPs or latency) versus F1 score trade-off when scaling context windows from 128K to 256K tokens compared to iterative retrieval with reranking for multi-hop QA under adversarial distractor conditions?.

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

2 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 7.3/10 → 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.2310.07521
- https://doi.org/10.48550/arxiv.2312.10997