SOVEREIGN: To what extent does the marginal accuracy improvement of extending context windows from 32K to 128K tokens in

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Retrieval-Augmented Generation (RAG) has quickly grown into a pivotal paradigm in the development of Large Language Models (LLMs). Although existing research mainly emphasizes accuracy and efficiency, the trustworthiness of RAG systems remains insufficiently explored. RAG can improve LLM reliability by grounding responses in external and up-to-date knowledge, reducing hallucinations. However, unreliable retrieval or improper knowledge utilization may still lead to undesirable outputs. To address these concerns, we propose a unified framework, Trust-RAG Compass, that assesses the trustworthines

1 Introduction

Analysis of: Trustworthiness in Retrieval-Augmented Generation Systems: A Survey. Research goal: To what extent does the marginal accuracy improvement of extending context windows from 32K to 128K tokens in RAG-based multi-hop reasoning on HotPotQA saturate after 3 retrieval steps, and how does this saturation differ between 7B and 70B parameter LLMs when controlling for retrieval precision?.

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. 4 claims extracted, 4 verified. Tribunal: 8.3/10 → 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
Retrieval-Augmented Generation (RAG) can improve LLM reliability by grounding responses in external and up-to-date knowl	✓	0.31
Trust-RAG Compass framework assesses the trustworthiness of RAG systems across six key dimensions: factuality, robustnes	✓	0.32
TRC Bench (Trust-RAG Compass Benchmark) provides evaluation benchmark for trustworthiness of RAG systems across six dime	✓	0.22
RAG systems can have undesirable outputs due to unreliable retrieval or improper knowledge utilization	✓	0.23

References

- https://doi.org/10.48550/arxiv.2406.07887
- https://doi.org/10.48550/arxiv.2409.10102
- https://doi.org/10.48550/arxiv.2310.07521