SOVEREIGN: How does the inference efficiency (tokens/sec and memory usage) of a 70B-parameter LLM agent compare when usin

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Retrieval-Augmented Generation (RAG) methods enhance LLM performance by efficiently filtering relevant context for LLMs, reducing hallucinations and inference cost. However, most existing RAG methods focus on single-step retrieval, which is often insufficient for answering complex questions that require multi-step search. Recently, multi-step retrieval approaches have emerged, typically involving the fine-tuning of small LLMs to perform multi-step retrieval. This type of fine-tuning is highly resource-intensive and does not enable the use of larger LLMs. In this work, we propose Q-RAG, a novel

1 Introduction

Analysis of: Q-RAG: Long Context Multi-step Retrieval via Value-based Embedder Training. Research goal: How does the inference efficiency (tokens/sec and memory usage) of a 70B-parameter LLM agent compare when using a 2-step retrieval-augmented generation pipeline versus a long-context-only approach on the MuSiQue multi-hop QA dataset?.

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, 0 verified. Tribunal: 1.2/10 → REJECT (revision_round=0). 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
Q-RAG achieves the highest average performance on BabiLong in ultra-long contexts ranging from 1 to 10 million tokens.	×	0.06
The majority of models perform worst on the QA3 subtask of BabiLong.	×	0.02
Q-RAG demonstrates superior generalization to long contexts compared to other specialized long-context methods.	×	0.09
BabiLong and RULER require long contexts.	×	0.11
MuSiQue and HotPotQA use short contexts.	×	0.06
Some baselines require at least 8×A100 GPUs to fine-tune.	×	0.03

References

- http://arxiv.org/abs/2511.07328v2
- http://arxiv.org/abs/2508.06447v2
- http://arxiv.org/abs/2507.23334v2