SOVEREIGN: Context-Length Robustness in Question Answering Models: A Comparative Empirical

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Large language models are increasingly deployed in settings where relevant information is embedded within long and noisy contexts. Despite this, robustness to growing context length remains poorly understood across different question answering tasks. In this work, we present a controlled empirical study of context-length robustness in large language models using two widely used benchmarks: SQuAD and HotpotQA. We evaluate model accuracy as a function of total context length by systematically increasing the amount of irrelevant context while preserving the answer-bearing signal. This allows us t

1 Introduction

Analysis of: Context-Length Robustness in Question Answering Models: A Comparative Empirical Study. Research goal: How does the robustness of GPT-4's multi-hop reasoning degrade under increasing retrieval steps (2 vs 5) on the Cofca benchmark, and does the accuracy-throughput trade-off favor wider context windows or multi-step retrieval pipelines?.

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

14 papers retrieved. 4 claims extracted, 4 verified. Tribunal: 7.8/10 → 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
Large language models show consistent degradation in performance as context length increases.	✓	0.24
HotpotQA exhibits nearly twice the accuracy degradation of SQuAD under equivalent context expansions.	✓	0.28
Multi-hop reasoning tasks are especially vulnerable to context dilution compared to single-span extraction tasks.	✓	0.28
Context-length robustness should be evaluated explicitly when assessing model reliability for applications involving lon	✓	0.35

References

- https://www.semanticscholar.org/paper/bbe2946a3064583b45b35714bf0a086f3e8cb74d
- https://www.semanticscholar.org/paper/1d5094f4d7b882bf2abf3185fcd1bb1667bcf831
- https://www.semanticscholar.org/paper/9971fc11df739ae24cdbcf6ef2dae95b5c9c8d24