Gemini 1.5 Pro Performance Degradation on Qasper with Context Position Shifts

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: How does the performance of Gemini 1.5 Pro on the Qasper dataset degrade as the position of relevant information shifts from the beginning to the middle versus the end of a 500k token context window. 8 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs. Research question: How does the performance of Gemini 1.5 Pro on the Qasper dataset degrade as the position of relevant information shifts from the beginning to the middle versus the end of a 500k token context window?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.1/10.

3 Results

1 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 8.1/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
In traditional RAG framework, the basic retrieval units are normally short, typically 100-word Wikipedia paragraphs.	✓	0.27
The imbalanced 'heavy' retriever and 'light' reader design in traditional RAG can lead to sub-optimal performance.	✓	0.26
The loss of contextual information in short, chunked units may increase the likelihood of introducing hard negatives dur	✓	0.29
LongRAG processes the entire Wikipedia corpus into 4K-token units by grouping related documents.	✓	0.28
By increasing the unit size, LongRAG significantly reduces the total number of units, reducing the burden on the retriev	✓	0.19
LongRAG achieves an EM of 62.7% on NQ and 64.3% on HotpotQA without requiring any training.	✓	0.22
LongRAG's performance on NQ and HotpotQA is on par with the (fully-trained) state-of-the-art (SoTA) model.	✓	0.17
LongRAG processes each individual document as a single (long) unit for non-Wikipedia-based datasets like Qasper and Mult	✓	0.32

References

- https://doi.org/10.48550/arxiv.2406.15319