Language Model Performance on Multi-Document Reasoning and Summarization Across Context Lengths

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does context length affect language model performance on multi-document reasoning and summarization v15. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Estimating Optimal Context Length for Hybrid Retrieval-augmented Multi-document Summarization. Research question: How does context length affect language model performance on multi-document reasoning and summarization v15.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

4 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 4.2/10.

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
All RAG-based systems (baselines and the proposed method) outperform the full-context setup on the SummHay dataset.	×	0.07
The HELMET LongQA-based estimate is the best performing baseline.	×	0.02
Neither RULER-based nor HELMET-based estimates are consistently the best in all settings.	×	0.04
The proposed method was evaluated on models ranging from 0.5B to 72B parameters.	×	0.03
Qwen-2.5 models with $\leq$7B parameters can run on a single 48GB GPU.	×	0.02
Larger models (above 7B) require up to 4$\times$48GB GPUs.	×	0.02
The proposed method often provides a significantly shorter context length estimate compared to baselines.	×	0.08
The proposed method requires task-specific additional inference time compute to determine the optimal context length.	×	0.11
Benchmarks such as RULER and HELMET also require compute to compute task averages.	×	0.07
The proposed estimation method requires a very small sample of the dataset.	×	0.06
In the system pooling comparison using a GTE 7B retriever, Qwen-2.5 72B was picked least often.	×	0.01
In the system pooling comparison using a GTE 7B retriever, Llama-3.3 70B was picked most often.	×	0.01
A total of 276 silver summaries were generated by picking the top-3 summaries per input post-MBR decoding.	×	0.01
Llama-3.3 70B contributed 79 silver summaries in the post-MBR decoding count.	×	0.01
Qwen-2.5 72B contributed 33 silver summaries in the post-MBR decoding count.	×	0.01
ProLong 512K contributed 59 silver summaries in the post-MBR decoding count.	×	0.01

References

- http://arxiv.org/abs/2503.06692v5
- http://arxiv.org/abs/2212.14815v3
- http://arxiv.org/abs/2504.12972v1