Context Length Effects on Language Model Performance in Multi-Document Reasoning and Summarization

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does context length affect language model performance on multi-document reasoning and summarization. 19 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.6/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Parallel Context Windows for Large Language Models. Research question: How does context length affect language model performance on multi-document reasoning and summarization.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.6/10.

3 Results

16 papers retrieved. 19 claims extracted; 4 independently verified. Quality review score: 5.6/10.

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
PCW leads to statistically significant improvements in almost all cases compared to ICL.	×	0.02
PCW results in smaller standard deviations compared to ICL.	×	0.03
Prior work has not extensively experimented with information extraction in an in-context learning setting.	×	0.05
Zhao et al. (2021) reported question-answering in an in-context learning setting.	×	0.11
PCW allows in-context learning in information extraction tasks.	×	0.06
J1-Grande (17B) with PCW achieves 91.7% EM on ATIS.	×	0.01
J1-Grande (17B) with PCW achieves 69.3% EM on MIT Movies.	×	0.01
J1-Grande (17B) with PCW achieves 85.1% F1 on SQuAD.	×	0.04
J1-Grande (17B) with PCW achieves 47.4% F1 on adversarialQA.	×	0.02
PCW leads to a significant improvement in OpenBookQA with J1-Grande.	×	0.01
PCW does not significantly improve or worsen over ICL in other cases for J1-Grande.	×	0.02
PCW is expected to help aggregate information from multiple texts in question-answering settings.	×	0.05
PCW is used for parallel processing of documents related to the test example in question-answering settings.	×	0.11
PCW improves performance in Natural Questions (NQ) with J1-Grande as the number of documents per window increases.	×	0.04
PCW provides a significant boost in performance compared to single context window ICL.	×	0.07
PCW shows substantial improvements for tasks with diverse input and output spaces.	✓	0.19
PCW shows additional benefits in multi-hop questions and retrieval-augmented question answering with multiple retrieved	✓	0.24
PCW is a promising method for applying off-the-shelf LLMs in settings that require long text sequences.	✓	0.31
PCW code is publicly available at https://github.com/ai21labs/parallel-context-windows.	✓	0.34

References

- http://arxiv.org/abs/2312.00513v1
- http://arxiv.org/abs/2212.10947v3
- http://arxiv.org/abs/2505.09561v2