# Context Length Scaling and Hallucination Rates in Sub-10B Theological Retrieval-Augmented Models

Assignee Research

June 8, 2026

**Abstract**

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: To what extent does the scaling of context length in retrieval-augmented generation systems affect hallucination rates in sub-10B models specialized for sensitive theological domains. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. 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: To what extent does the scaling of context length in retrieval-augmented generation systems affect hallucination rates in sub-10B models specialized for sensitive theological domains?.

## 2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

## 3 Results

15 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.2/10.

## 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 ours) outperform full-context setup. | × | 0.06 |
| Our method consistently shows strong performance across model classes, sizes, and retrievers. | × | 0.11 |
| The HELMET LongQA-based estimate is the best baseline. | × | 0.03 |
| Our results from Table 1, Table 2, and Table 3 show the effectiveness of our method in models ranging from 0.5B to 72B p | × | 0.06 |
| Qwen-2.5 $\leq$7B can run on a single 48GB GPU, while larger models would require up to 4$\times$48GB GPUs. | × | 0.01 |
| Our method often provides a significantly shorter context length estimate compared to the baselines. | × | 0.10 |
| Our estimation requires a very small sample of the dataset. | × | 0.06 |
| Qwen-2.5 72B is picked least often post-MBR decoding. | × | 0.04 |
| Llama-3.3 70B is picked most often post-MBR decoding. | × | 0.04 |
| We pick top-3 summaries per input, so a total of 276 summaries. | × | 0.02 |

## References


- http://arxiv.org/abs/2310.10378v5
- http://arxiv.org/abs/2510.22344v1
- http://arxiv.org/abs/2504.12972v1