

# Faithfulness Degradation in 7B and 70B Models Under Context Overload on QASPER

Assignee Research

June 7, 2026

## Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How do 7B and 70B models compare in terms of answer faithfulness metrics on the QASPER benchmark when subjected to context overload beyond their trained sequence length. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: SLM-Bench: A Comprehensive Benchmark of Small Language Models on Environmental Impacts–Extended Version. Research question: How do 7B and 70B models compare in terms of answer faithfulness metrics on the QASPER benchmark when subjected to context overload beyond their trained sequence length?.

## 2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.4/10.

## 3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.4/10.

## 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.

## References

- <http://arxiv.org/abs/2410.13187v3>
- <http://arxiv.org/abs/2508.15478v2>
- <http://arxiv.org/abs/2311.09215v3>