Small Parameter Models Exhibit Higher Hallucination Rates in High-Density Retrieval Contexts

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: To what extent do smaller parameter models (7B) suffer from hallucination rates compared to larger models (70B) when retrieval context density exceeds optimal thresholds in domain-specific QA tasks. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Hallucination Detection with Small Language Models. Research question: To what extent do smaller parameter models (7B) suffer from hallucination rates compared to larger models (70B) when retrieval context density exceeds optimal thresholds in domain-specific QA tasks?.

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 3.5/10.

3 Results

15 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
Small language models can produce accurate results for specific tasks.	×	0.12
The proposed framework improves verification accuracy by 10% over the baseline.	×	0.07
The proposed method is superior to both P(yes) and ChatGPT in detecting correct responses from partial responses.	×	0.14
The 'max' method achieves the highest F1 score of 0.99 in Fig. 5 (a).	×	0.03
The 'harmonic' method achieves the highest F1 score of 0.81 in Fig. 5 (b).	×	0.03
LLMs demonstrate substantial proficiency across a variety of NLP tasks, including language translation, sentiment analys	×	0.04
The transformer architecture enables models to effectively capture long-range dependencies within textual data.	×	0.03
LLMs are known to produce hallucinations in their outputs.	×	0.07
Detecting hallucinations is not analogous to conventional LLM measurements, such as the ROUGE metric and BLEU score.	×	0.04
The geometric mean is calculated using the formula: si(S, m = 'geo') = exp(1/|S(ri)| * Σ log(si,j)) for si,j > 0.	×	0.01
The minimum value in the dataset is found using the formula: si(S, m = 'min') = min(S(ri)).	×	0.01
The maximum value in the dataset is found using the formula: si(S, m = 'max') = max(S(ri)).	×	0.01

References

- http://arxiv.org/abs/2506.22486v1
- http://arxiv.org/abs/2509.12382v1
- http://arxiv.org/abs/2503.16581v1