Phi-3-Mini and Mistral-7B Hallucination and Factuality in Long-Context RAG Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How do Phi-3-mini and Mistral-7B-v0.1 differ in hallucination rates and factuality scores on long-context retrieval tasks evaluated by RAG benchmarks. 6 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: FARSIQA: Faithful and Advanced RAG System for Islamic Question Answering. Research question: How do Phi-3-mini and Mistral-7B-v0.1 differ in hallucination rates and factuality scores on long-context retrieval tasks evaluated by RAG benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.8/10.

3 Results

14 papers retrieved. 6 claims extracted; 2 independently verified. Quality review score: 5.8/10.

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
FARSIQA achieves a 'Answer Correctness' score of 74.3% as evaluated via LLM-as-Judge on multi-hop queries.	×	0.13
FARSIQA substantially outperforms standard baselines across metrics of relevance, correctness, and robustness.	×	0.04
The knowledge base for FARSIQA comprises approximately 431,000 unique documents from eleven reputable online Persian Isl	×	0.08
The knowledge base for FARSIQA includes approximately 304,000 question-answer pairs from authoritative Q&A websites.	×	0.07
FARSIQA achieves a remarkable 97.0% in Negative Rejection, a 40-point improvement over standard baselines.	✓	0.20
FARSIQA achieves a high Answer Correctness score of 74.3%.	✓	0.15

References

- http://arxiv.org/abs/2510.25621v1
- http://arxiv.org/abs/2510.22344v1
- http://arxiv.org/abs/2511.07328v2