Hybrid Retrieval Reduces Hallucinations in Mistral-7B Across Legal and Scientific Domains

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does the hybrid retrieval approach perform in mitigating hallucinations in Mistral-7B when applied to domain-specific benchmarks beyond religious texts, such as legal or scientific corpora,. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Hallucination to Truth: A Review of Fact-Checking and Factuality Evaluation in Large Language Models. Research question: How does the hybrid retrieval approach perform in mitigating hallucinations in Mistral-7B when applied to domain-specific benchmarks beyond religious texts, such as legal or scientific corpora, measured by F1 score and hallucination rate?.

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 4.0/10.

3 Results

16 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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
LLMs tend to 'hallucinate,' or generate text that sounds plausible but is factually incorrect.	×	0.03
Evaluating LLMs for fact-checking involves checking the model's output against provided evidence or reliable external so	×	0.09
Traditional classification metrics like accuracy, precision, recall, and F1-score are widely used for evaluating fact-ch	×	0.06
Macro-averaged versions of classification metrics are employed in multiclass scenarios such as classifying statements as	×	0.01
Token-level precision with annotated answers is typical for short-form responses.	×	0.06
Lexical overlap metrics such as BLEU-4, METEOR, and chrF assess surface-level similarity in text generation tasks.	×	0.03
ROUGE evaluates the extent to which summaries or explanations capture the core content.	×	0.02
Semantic similarity metrics like BERTScore are used for evaluating text generation tasks.	×	0.05
Dataset characteristics such as domain specificity, annotation quality, and multilingual coverage affect the performance	×	0.07
Vykopal et al. conducted a survey of approaches and techniques used in automated fact-checking using generative LLMs, su	×	0.07

References

- http://arxiv.org/abs/2404.07220v2
- http://arxiv.org/abs/2503.16581v1

- http://arxiv.org/abs/2508.03860v2