To what extent does fine-tuning 7B parameter models on domain-specific datasets reduce hallucination rates compared to

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: To what extent does fine-tuning 7B parameter models on domain-specific datasets reduce hallucination rates compared to their larger counterparts when evaluated on tasks like ELI5 or SciFact. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. 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: To what extent does fine-tuning 7B parameter models on domain-specific datasets reduce hallucination rates compared to their larger counterparts when evaluated on tasks like ELI5 or SciFact?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

13 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
Large Language Models (LLMs) have a tendency to 'hallucinate,' generating text that sounds plausible but is factually in	×	0.07
Evaluating LLMs for fact-checking typically involves checking the model's output against provided evidence or reliable e	×	0.08
Methods for evaluating LLM factuality include using LLMs themselves as evaluators and building benchmarks that unify dat	×	0.06
Evaluation tasks for fact-checking are most often approached as classification problems to determine if a claim is true	×	0.07
Metrics such as accuracy, precision, recall, and F1-score are widely used for LLM fact-checking evaluation tasks.	×	0.08
In multiclass scenarios (e.g., supported, refuted, inconclusive), macro-averaged versions of classification metrics are	×	0.01
For short-form responses, token-level precision with annotated answers is a typical evaluation metric.	×	0.03
Direct comparisons between different models using token-level responses are difficult due to high dependence on the spec	×	0.03
Traditional classification metrics often reduce complex output to a binary judgment, overlooking reasoning quality or nu	×	0.03
Traditional classification metrics may be misleading in datasets with imbalanced labels.	×	0.03
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.	×	0.04
Vykopal et al. conducted a survey of approaches and techniques used in automated fact-checking using generative LLMs, in	×	0.11

References

- http://arxiv.org/abs/2508.11281v3
- http://arxiv.org/abs/2508.03860v2
- http://arxiv.org/abs/2110.06500v2