Small Language Models Under Adversarial Prompts: Reasoning Robustness vs. Large Baselines

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: To what extent do small language models evaluated in SLM-Bench maintain reasoning accuracy when subjected to adversarial prompt perturbations compared to larger LLM baselines. Large language models (LLMs) have demonstrated impressive capabilities in natural language processing. However, their internal mechanisms are still unclear and this lack of transparency poses unwanted risks for downstream applications. 10 claims were extracted from source literature; 10 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Explainability for Large Language Models: A Survey. Research question: To what extent do small language models evaluated in SLM-Bench maintain reasoning accuracy when subjected to adversarial prompt perturbations compared to larger LLM baselines?.

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 8.2/10.

3 Results

13 papers retrieved. 10 claims extracted; 10 independently verified. Quality review score: 8.2/10.

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 demonstrated impressive capabilities in natural language processing.	✓	0.27
The internal mechanisms of large language models are still unclear.	✓	0.16
The lack of transparency in large language models poses unwanted risks for downstream applications.	✓	0.24
The article introduces a taxonomy of explainability techniques for Transformer-based language models.	✓	0.20
The article categorizes explainability techniques based on two training paradigms: traditional fine-tuning-based and pro	✓	0.26
The article summarizes goals and dominant approaches for generating local explanations of individual predictions.	✓	0.26
The article summarizes goals and dominant approaches for generating global explanations of overall model knowledge.	✓	0.25
The article discusses metrics for evaluating generated explanations.	✓	0.17
The article discusses how explanations can be leveraged to debug models and improve performance.	✓	0.21
The article examines key challenges and emerging opportunities for explanation techniques in the era of LLMs compared to	✓	0.27

References

- https://doi.org/10.1145/3639372
- https://doi.org/10.48550/arxiv.2308.12950
- https://doi.org/10.4230/oasics.icpec.2025.4