Synthetic Dialogue Diversity and Safety Classifier False Positives on AdvBench

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does training on synthetic dialogue data with varying levels of semantic diversity impact the false positive rate of safety classifiers on the AdvBench dataset compared to standard instruction. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Benchmarking Adversarial Robustness to Bias Elicitation in Large Language Models: Scalable Automated Assessment with LLM-as-a-Judge. Research question: How does training on synthetic dialogue data with varying levels of semantic diversity impact the false positive rate of safety classifiers on the AdvBench dataset compared to standard instruction tuning?.

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 3.1/10.

3 Results

15 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.1/10.

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 (SLMs) are defined in this study as having a parameter count typically up to a few tens of billion	×	0.03
Gemma2 2B, Gemma2 27B, Phi-4 14B, Llama 3.1 8B, and GPT-4o mini were categorized as Small Language Models (SLMs) in the	×	0.02
Gemini 2.0 Flash, Llama 3.1 405B, Claude 3.5 Sonnet, DeepSeek V3 671B, and GPT-4o were categorized as Large Language Mod	×	0.06
A safety threshold (τ) of 0.5 was defined, where a model is considered safe if its safety score exceeds this value.	×	0.05
All Small Language Models tested locally, excluding GPT-4o mini, were run on an NVIDIA A30 GPU using the Ollama service.	×	0.03
Testing the locally run Small Language Models required a total of 10 GPU hours.	×	0.03
The estimated total cost for evaluating models accessed via API was approximately 35 USD.	×	0.02
Querying the judge LLM (DeepSeek V3) accounted for approximately 30% of the total API evaluation cost.	×	0.07
The judge evaluation control set was constructed by randomly sampling a small subset of prompts from the base prompts in	×	0.07
Five responses were manually curated for each prompt and for each class in the judge evaluation control set.	×	0.03
Five candidate large models were assessed for suitability as a judge: GPT-4o, Claude 3.5 Sonnet, Llama 3.1 405B, and oth	×	0.02

References

- http://arxiv.org/abs/2504.07887v2
- http://arxiv.org/abs/2205.14230v2
- http://arxiv.org/abs/2402.11690v1