Robustness of Large Language Models to Adversarial Prompts via Reward Shaping on AdvBench

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the effect of different reward shaping functions on the robustness of LLMs against adversarial prompts in the AdvBench dataset. 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: What is the effect of different reward shaping functions on the robustness of LLMs against adversarial prompts in the AdvBench dataset?.

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 3.1/10.

3 Results

16 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	×	0.02
Gemma2 2B, Gemma2 27B, Phi-4 14B, Llama 3.1 8B, and GPT-4o mini are categorized as Small Language Models (SLMs) in this	×	0.02
Gemini 2.0 Flash, Llama 3.1 405B, Claude 3.5 Sonnet, DeepSeek V3 671B, and GPT-4o are categorized as Large Language Mode	×	0.07
A safety threshold (τ) of 0.5 was defined for the evaluation, where a model is considered safe if its safety score excee	×	0.06
All tested SLMs, excluding GPT-4o mini, were run locally on an NVIDIA A30 GPU using the Ollama service.	×	0.01
The local evaluation of SLMs (excluding GPT-4o mini) required a total of 10 GPU hours.	×	0.02
The estimated total cost for evaluating models accessed via API was approximately 35 USD.	×	0.01
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	×	0.08
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 the role of judge: GPT-4o, Claude 3.5 Sonnet, Llama 3.1 405B, and others m	×	0.02

References

- http://arxiv.org/abs/2504.07887v2
- http://arxiv.org/abs/2410.20971v2
- http://arxiv.org/abs/2407.04295v2