Adversarial Prompting Effects on Cross-Domain Alignment Stability in Multilingual LLMs

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: To what extent does adversarial prompting impact the cross-domain alignment stability of multilingual LLMs when measured by refusal rate consistency in technical domains versus general conversational. The growing integration of Large Language Models (LLMs) into critical societal domains has raised concerns about embedded biases that can perpetuate stereotypes and undermine fairness. Such biases may stem from historical inequalities in training data, linguistic imbalances, or. 6 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: Benchmarking Adversarial Robustness to Bias Elicitation in Large Language Models: Scalable Automated Assessment with LLM-as-a-Judge. Research question: To what extent does adversarial prompting impact the cross-domain alignment stability of multilingual LLMs when measured by refusal rate consistency in technical domains versus general conversational benchmarks?.

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 4.0/10.

3 Results

13 papers retrieved. 6 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
The safety threshold τ is defined as 0.5, and a model is considered safe if its safety score exceeds this threshold.	×	0.04
All SLMs, excluding GPT-4o mini, were tested locally on an NVIDIA A30 GPU using the Ollama service, requiring a total of	×	0.04
For the remaining models, accessed via API, the total cost was approximately 35 USD, with querying the judge LLM account	×	0.03
The models assessed include Gemma2 2B, Gemma2 27B, Phi-4 14B, Llama 3.1 8B, GPT-4o mini for SLMs, and Gemini 2.0 Flash,	×	0.05
The control set for judge evaluation was constructed by randomly sampling a small subset of prompts from the base prompt	×	0.06
Five candidate large models—GPT-4o, Claude 3.5 Sonnet, Llama 3.1 405B, Gem—were assessed for the judge evaluation.	×	0.06

References

- http://arxiv.org/abs/2504.07887v2
- http://arxiv.org/abs/2604.23270v1
- http://arxiv.org/abs/2410.15308v2