Synthetic Soft-Label Pre-Training Duration and Bias Resistance Degradation in LLMs

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the correlation between synthetic soft-label pre-training duration and the degradation of bias elicitation resistance in LLMs evaluated via LLM-as-a-Judge frameworks. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. 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 correlation between synthetic soft-label pre-training duration and the degradation of bias elicitation resistance in LLMs evaluated via LLM-as-a-Judge frameworks?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

11 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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
The safety threshold τ is defined as 0.5 for evaluating model safety.	×	0.06
Models are considered safe if their safety score exceeds the threshold τ = 0.5.	×	0.03
Small Language Models (SLMs) are defined as those with a parameter count typically up to a few tens of billions.	×	0.03
Gemma2 2B, Gemma2 27B, Phi-4 14B, Llama 3.1 8B, and GPT-4o mini are categorized as Small Language Models (SLMs).	×	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
All SLMs, excluding GPT-4o mini, were tested locally on an NVIDIA A30 GPU using the Ollama service, requiring a total of	×	0.02
The total cost for accessing the remaining models via API was approximately 35 USD.	×	0.01
Querying the judge LLM (DeepSeek V3) accounted for approximately 30% of the total cost.	×	0.07
Five candidate large models—GPT-4o, Claude 3.5 Sonnet, Llama 3.1 405B, Gem—were assessed for the judge evaluation.	×	0.03

References

- http://arxiv.org/abs/2504.07887v2
- http://arxiv.org/abs/2502.06193v3
- http://arxiv.org/abs/2408.13006v2