# Robustness of Llama3.1-8B and Mistral 7B Against Powertrain Adversarial Perturbations

Assignee Research

June 4, 2026

**Abstract**

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the robustness of Llama3.1-8B against domain-specific adversarial perturbations in powertrain data compare to Mistral 7B across varying context window sizes. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.4/10. 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 the robustness of Llama3.1-8B against domain-specific adversarial perturbations in powertrain data compare to Mistral 7B across varying context window sizes?.

## 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 3.4/10.

## 3 Results

13 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.4/10.

## 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 $\tau$ is defined as 0.5, and a model is considered safe if its safety score exceeds this threshold. | × | 0.04 |
| The computational resources used for evaluating SLMs (excluding GPT-4o mini) were 10 GPU hours on an NVIDIA A30 GPU usin | × | 0.05 |
| The total cost for evaluating the remaining models via API was approximately 35 USD, with 30% of this cost attributed to | × | 0.07 |
| The control set for judge evaluation was constructed by randomly sampling a small subset of prompts from the base prompt | × | 0.07 |
| Five candidate large models—GPT-4o, Claude 3.5 Sonnet, Llama 3.1 405B, Gem—were assessed for the role of the judge. | × | 0.03 |
| The safety scores for models are presented in the benchmark tables, with values ranging from 0 to 1. | × | 0.05 |
| The refusal rate for one of the models is 0.52, and the debiasing rate is 0.11. | × | 0.02 |
| The stereotype and counter-stereotype rates for another model are 0.54 and 0.34, respectively. | × | 0.02 |

## References


- http://arxiv.org/abs/2504.07887v2
- http://arxiv.org/abs/2310.06825v1
- http://arxiv.org/abs/2604.14171v1