Llama3-70B and Codestral-7B Alignment with Human Security Judgments Across Fine-Tuning Iterations

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the alignment of Llama3-70B with human security review judgments (measured by EM score on SECURITYBENCH) evolve compared to Codestral-7B across different iterations of instruction fine-tuning. Large language models (LLMs), initially developed for generative AI, are now evolving into agentic AI systems, which make decisions in complex, real-world contexts. Unfortunately, while their generative capabilities are well-documented, their decision-making processes remain. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Teaching AI to Handle Exceptions: Supervised Fine-Tuning with Human-Aligned Judgment. Research question: How does the alignment of Llama3-70B with human security review judgments (measured by EM score on SECURITYBENCH) evolve compared to Codestral-7B across different iterations of instruction fine-tuning?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

12 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
At many prompt-level exception intensities, the LLM refusal rate is close to 1, meaning the LLM-generated decision is al	×	0.03
The Claude refusal rates reported represent a weighted average of responses from Opus 4, Sonnet 4, and Haiku 3.5.	×	0.02
In a scenario where a friend needed flour for a birthday cake but the price was 1 cent above their stated limit, 92.3% o	×	0.02
In a scenario involving a red light while someone had fainted on the sidewalk ahead, 83.3% of human participants indicat	×	0.01
In a scenario evaluating welfare benefits for a family earning one dollar above the income threshold, 79.1% of human par	×	0.02
The study compared LLM responses to responses from 303 human participants.	×	0.04
In the Grocery Prices scenario, the first level of exception involves flour costing $25 against a $10 limit.	×	0.01
In the Grocery Prices scenario, the sixth level of exception involves flour costing $10.01 against a $10 limit.	×	0.01
Off-the-shelf LLMs demonstrate rigid adherence to policy, while human decision-makers exhibit flexibility based on situa	×	0.05

References

- http://arxiv.org/abs/2408.07888v2
- http://arxiv.org/abs/2503.02976v3
- http://arxiv.org/abs/2312.10793v3