# Llama-3 Alignment Reduces Spurious Texture Sensitivity Over Vicuna SFT in Distribution Shifts

Assignee Research

June 7, 2026

**Abstract**

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: Does the alignment strategy used in Llama-3 reduce sensitivity to background texture spurious features more effectively than the SFT approach in Vicuna, as measured by performance drops on shifted. 8 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Thinking Beyond Distributions in Testing Machine Learned Models. Research question: Does the alignment strategy used in Llama-3 reduce sensitivity to background texture spurious features more effectively than the SFT approach in Vicuna, as measured by performance drops on shifted distributions in MLNeedle?.

## 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 8.3/10.

## 3 Results

11 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 8.3/10.

## 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 |
|---|---|---|
| Testing practices within the machine learning community have centered around assessing a learned model's predictive perf | ✓ | 0.38 |
| Machine learning test datasets are often drawn from the same distribution as the training dataset. | ✓ | 0.20 |
| Recent work on robustness and fairness testing within the ML community has pointed to the importance of testing against | ✓ | 0.38 |
| Recent efforts on robustness and fairness testing focus on estimating the likelihood of the model making an error agains | ✓ | 0.33 |
| The current view of testing actively discourages researchers and developers from looking into other sources of robustnes | ✓ | 0.34 |
| Corner cases in machine learning models may have severe undesirable impacts. | ✓ | 0.23 |
| Decades of work within software engineering testing have focused on assessing software systems against various stress co | ✓ | 0.36 |
| Software engineering testing has historically assessed systems against stress conditions as opposed to solely focusing o | ✓ | 0.26 |

## References


- http://arxiv.org/abs/2305.12100v3
- http://arxiv.org/abs/2112.03057v1
- http://arxiv.org/abs/2403.03375v3