Scaling Model Parameters and Robustness in SFT-Aligned vs DPO-Aligned Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 5 peer-reviewed papers addressing the following research question: What is the impact of scaling model parameters on the robustness of SFT-aligned versus DPO-aligned models when evaluated on StressTest adversarial subsets. 6 claims were extracted from source literature; 6 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Comprehensive Review of Neuro-symbolic AI for Robustness, Uncertainty Quantification, and Intervenability. Research question: What is the impact of scaling model parameters on the robustness of SFT-aligned versus DPO-aligned models when evaluated on StressTest adversarial subsets?.

2 Methodology

Systematic literature search across multiple databases yielded 5 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.8/10.

3 Results

5 papers retrieved. 6 claims extracted; 6 independently verified. Quality review score: 8.8/10.

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
AI systems are increasingly deployed in high-stakes domains such as healthcare, autonomous systems, finance, and critica	✓	0.30
Purely data-driven 'black-box' models have limitations in handling distribution shifts, ambiguous inputs, and human over	✓	0.26
Neuro-symbolic AI combines the learning capabilities of neural networks with the reasoning strengths of symbolic AI	✓	0.31
Neuro-symbolic systems offer enhanced interpretability, verifiability, and control	✓	0.26
This paper presents a comprehensive survey of neuro-symbolic AI through the lenses of robustness, uncertainty quantifica	✓	0.33
The paper systematically reviews state-of-the-art techniques for modeling robustness, quantifying uncertainty, and enabl	✓	0.24

References

- https://openalex.org/W7155574602
- https://doi.org/10.1007/s13369-025-10887-3
- https://doi.org/10.36227/techrxiv.177160514.47512518/v1