Multi-Language LLM Alignment with Human Feedback Enhances Robustness in Adversarial Code Generation

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the alignment of multi-language LLMs with human feedback (e.g., using DPO or RLAIF) improve their robustness on adversarial code generation benchmarks like CodeGenBench, evaluated via BLEU. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating the Limits of Large Language Models in Multilingual Legal Reasoning. Research question: How does the alignment of multi-language LLMs with human feedback (e.g., using DPO or RLAIF) improve their robustness on adversarial code generation benchmarks like CodeGenBench, evaluated via BLEU and code execution success rates?.

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 3.3/10.

3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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.

References

- http://arxiv.org/abs/2509.22472v1
- http://arxiv.org/abs/2311.08588v3
- http://arxiv.org/abs/2412.15453v1