Multi-Objective Reinforcement Learning Robustness in Cross-Language Code Generation Tasks

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How robust is the Multi-Objective Reinforcement Learning approach for preference alignment in maintaining consistent performance scores across different code generation tasks in the. This paper addresses the challenge of aligning large language models (LLMs) with diverse human preferences within federated learning (FL) environments, where standard methods often fail to adequately represent diverse viewpoints. We introduce a comprehensive evaluation framework. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Systematic Evaluation of Preference Aggregation in Federated RLHF for Pluralistic Alignment of LLMs. Research question: How robust is the Multi-Objective Reinforcement Learning approach for preference alignment in maintaining consistent performance scores across different code generation tasks in the HumanEval-JavaScript and HumanEval-Java cross-language evaluations?.

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 4.5/10.

3 Results

13 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
The experiments use the Pew Research Center's Global Attitudes Surveys dataset, which consists of 2,554 multiple-choice	×	0.03
The dataset provides a probability vector over the answer choices for each question and country, indicating the fraction	×	0.02
The experiments treat each country as a separate user group (i.e., a federated client) and use all available groups in t	×	0.04
The goal is to align the LLM with diverse group preferences in a fair and robust manner, without overfitting to any sing	×	0.07
Performance is assessed using fairness index FI and alignment scores across two primary tasks: the preference probabilit	×	0.09
The evaluation framework encompasses various reward functions and aggregation strategies, comparing adaptive alpha aggre	✓	0.16
The LLM rollout at FL iteration t consists of a set of questions {qj} together with corresponding responses generated by	×	0.03
For each group g $\in$ Gtrain, a reward rg,j is computed by comparing the LLM prediction yt,llm_j against the PluralLLM-deri	×	0.04
The preference probability prediction prompt requires assigning a preference score to each of the 4 options, with each s	×	0.03
The preference ranking prompt requires ranking all 4 provided options from most to least preferred, with the output bein	×	0.03

References

- http://arxiv.org/abs/2512.08786v2
- http://arxiv.org/abs/2506.08062v2
- http://arxiv.org/abs/2402.18571v3