Impact of CHARM Calibration on Reward Model-Human Preference Correlation Across Reasoning Benchmarks for Qwen2.5

Assignee Research

June 11, 2026

Abstract

Reward models (RMs) play a crucial role in Reinforcement Learning from Human Feedback by serving as proxies for human preferences in aligning large language models. However, they suffer from various biases which could lead to reward hacking. In this paper, we identify a model preference bias in RMs, where they systematically assign disproportionately high scores to responses from certain policy models, leading to unfair judgments. To mitigate this bias, we propose a calibration method named CHatbot Arena calibrated Reward Modeling (CHARM) that leverages Elo scores from the Chatbot Arena to con

1 Introduction

This paper examines: CHARM: Calibrating Reward Models With Chatbot Arena Scores. Research question: How does CHARM calibration impact the correlation between reward model scores and human preferences across diverse reasoning benchmarks like MMLU and GSM8K for Qwen2.5 variants?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.4/10.

3 Results

14 papers retrieved. 22 claims extracted; 20 independently verified. Quality review score: 7.4/10.

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
CHARM provides a simple, effective, and broadly applicable approach to building more reliable and fair reward models.	✓	0.33
CHARM's code is available at https://github.com/HexagonStar/CHARM.	✓	0.17
Benchmarks such as MT-Bench, Alpaca-Eval, and Arena-hard employ LLM-as-a-judge systems to assess the quality of model re	✓	0.27
Ultrafeedback and RLAIF use LLMs for preference annotation which correlate training signals with evaluation measures.	✓	0.21
ChatBot Arena employs a crowdsourced, pairwise comparison system where users challenge two anonymous models with prompts	✓	0.27
ChatBot Arena uses human preference data to compute a dynamic Elo rating for each model, which serves as a widely recogn	✓	0.25
Park et al. (2024) identified six distinct types of bias in evaluation models and leveraged LLMs to construct a debiased	✓	0.30
Li et al. (2025) found that judge models may develop bias, favoring content generated by themselves or closely related L	✓	0.30
Dubois et al. (2024) proposed a regression-based method to mitigate length bias.	✓	0.30
Huang et al. (2025) introduced a post hoc calibration technique for reward models.	✓	0.22
Chatbot Arena represents a universal distribution of real-world prompts.	✓	0.17
An ideal RM should produce scores that strongly correlate with the platform's Elo ratings.	✓	0.23
AlpacaEval consists of 805 carefully curated questions.	✓	0.17
AlpacaEval exhibits a 98% Spearman correlation with Chatbot Arena.	✓	0.15
Five popular RMs were evaluated on a diverse set of policy models with varying Elo scores.	✓	0.20
RM Scores Correlate Positively with Human Preferences.	✓	0.17
CHARM achieves improved evaluation accuracy on RM-Bench and the Chat-Hard domain of RewardBench.	✓	0.23
CHARM produces scores more closely aligned with Elo rankings.	✓	0.17
CHARM improves downstream post-training performance.	×	0.12
CHARM reconstructs a debiased preference dataset to mitigate the preference bias in reward modeling.	✓	0.18
The Elo rating system provides a probabilistic	×	0.11

References

- http://arxiv.org/abs/2508.04149v2
- http://arxiv.org/abs/2402.18571v3
- http://arxiv.org/abs/2504.10045v2