How does dynamic hot neuron threshold adjustment in PowerInfer influence the alignment of LLaMA-70B outputs wi

Assignee Research

May 29, 2026

Abstract

Emerging research in Pluralistic Artificial Intelligence (AI) alignment seeks to address how intelligent systems can be designed and deployed in accordance with diverse human needs and values. We contribute to this pursuit with a dynamic approach for aligning AI with diverse and shifting user preferences through Multi Objective Reinforcement Learning (MORL), via post-learning policy selection adjustment. In this paper, we introduce the proposed framework for this approach, outline its anticipated advantages and assumptions, and discuss technical details about the implementation. We also examin

1 Introduction

This paper examines: Adaptive Alignment: Dynamic Preference Adjustments via Multi-Objective Reinforcement Learning for Pluralistic AI. Research question: How does dynamic hot neuron threshold adjustment in PowerInfer influence the alignment of LLaMA-70B outputs with human preferences in MBPP Python function synthesis tasks, measured by human evaluation scores?.

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 8.5/10.

3 Results

14 papers retrieved. 6 claims extracted; 6 independently verified. Quality review score: 8.5/10.

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
Pluralistic Artificial Intelligence (AI) alignment seeks to address how intelligent systems can be designed and deployed	✓	0.48
The paper introduces a dynamic approach for aligning AI with diverse and shifting user preferences through Multi-Objecti	✓	0.47
The proposed framework involves post-learning policy selection adjustment.	✓	0.24
The paper outlines the anticipated advantages and assumptions of the proposed framework.	✓	0.16
The paper discusses technical details about the implementation of the proposed framework.	✓	0.17
The paper examines the broader implications of adopting a retroactive alignment approach through the sociotechnical syst	✓	0.32

References

- http://arxiv.org/abs/2410.23630v1
- http://arxiv.org/abs/2503.14504v2
- http://arxiv.org/abs/2407.14477v4