2nd INTERNATIONAL CONFERENCE ON
Artificial Intelligence
31 May - 1 June 2026

IICAI
Second International Interdisciplinary Conference on Artificial Intelligence

From RLHF to DPO and Beyond: A Review of Preference Optimization Methods for Aligning Large Language Models

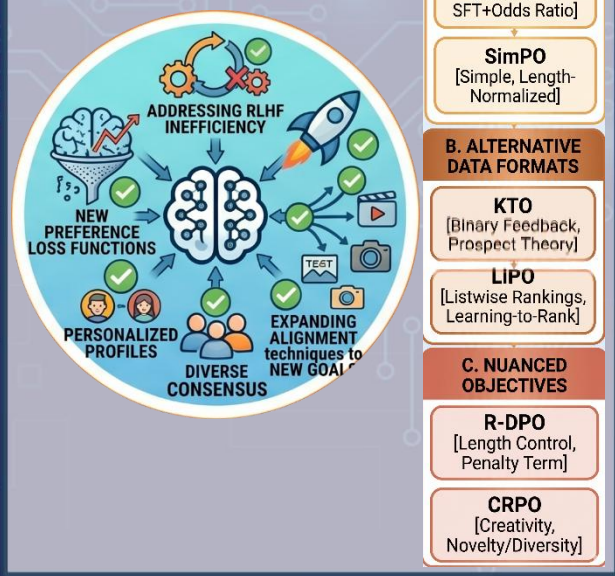
Poorya Saneei[1], Parham Rahimi[1] , Behrouz Minaei-Bidgoli*[1]
1. Department of Computer Engineering Iran University of Science and Technology Tehran, Iran
poorya_saneei@iust.ac.ir, pa_rahimi@alumni.iust.ac.ir, b_minaei@iust.ac.ir

Abstract

Large Language Models (LLMs) inherently lack alignment with human values. While Reinforcement Learning from Human Feedback (RLHF) initially addressed this, its complexity, instability, and high computational cost limited large-scale application. This paper traces the mathematical evolution of alignment methods, focusing on the shift from RLHF to Direct Preference Optimization (DPO), which elegantly reparametrizes alignment into a supervised learning problem. We review how subsequent algorithms overcome DPO's limitations by eliminating the reference model for better efficiency, expanding feedback formats, and optimizing for complex goals like creativity and multilingual alignment. The field is decisively moving towards modular, stable, and highly efficient reference-free algorithms.

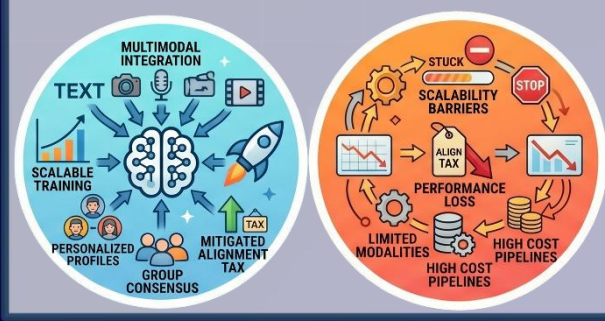
Research method

We conducted a systematic **literature review focusing on major archival journals** (ACL, NeurIPS, ICML) from 2022 to 2025.

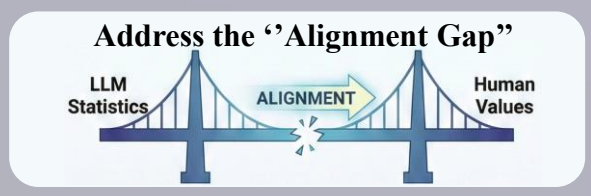


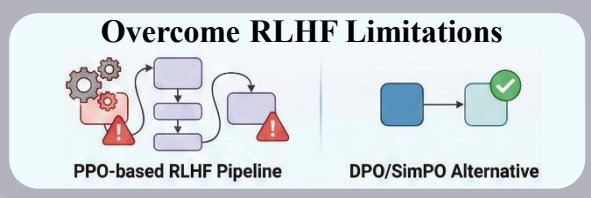
conclusion

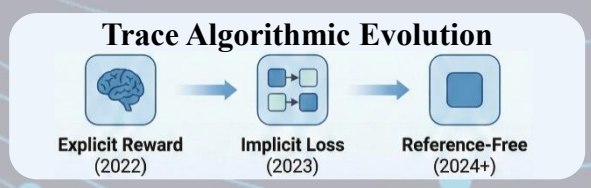
1. The field of LLM alignment has decisively shifted from **complex RLHF pipelines to simple**, stable direct optimization methods.
2. There is a strong prioritization of computational **efficiency** through reference-free models and monolithic pipelines.
3. Future research must address **scalability**, mitigate the "**alignment tax**", and pivot toward **multimodal**, personalized, and group-wise alignment strategies.



The purpose of the research







Practical or simulation results

- **Simplified Architecture**: DPO replaces complex reinforcement learning with a *direct, supervised loss function*.
- **Memory Efficiency**: Reference-free algorithms (CPO, ORPO, SimPO) significantly *lower GPU* resource demands.
- **Versatile Objectives**: Optimization now adapts to *diverse feedback formats* (binary, listwise) and specific goals like creativity or conciseness.
- **Global Applicability**: Techniques like MAPO and CALM use anchors and self-alignment to improve safety and reduce language hallucinations in *low-resource languages*.

References

[1] L. Ouyang et al., "Training language models to follow instructions with human feedback," NeurIPS, 2022.
[2] R. Rafailov et al., "Direct preference optimization: Your language model is secretly a reward model," NeurIPS, 2023.
[3] J. Hong et al., "ORPO: Monolithic Preference Optimization without Reference Model," EMNLP, 2024.
[4] Y. Meng et al., "SimPO: simple preference optimization with a reference-free reward," NeurIPS, 2024.
[5] K. Ethayarajh et al., "Model alignment as prospect theoretic optimization (KTO)," ICML, 2024.
10.18653/v1/2025.findings-acl.1088.