JOHANNES KEPLER
UNIVERSITY LINZ

Author
Giovanni Filomeno
12345678

Submission
Seminar in AI (Master)
(365.205)

March 2, 2025

Your Language Modell is Secretly a Reward Model

Seminar Report

Abstract Space for your (short) abstract.

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at

Contents

Abstract

Recent advances in large-scale language models have highlighted the challenges of aligning these models with human intent and preference. Among numerous alignment strategies, Reinforcement Learning from Human Feedback (RLHF) has attracted particular attention due to its demonstrated success in steering model behavior. However, RLHF pipelines often involve multiple components (e.g., a reward model and an RL algorithm like PPO) that can be costly, unstable, or prone to reward hacking. Direct Preference Optimization (DPO) offers a simpler alternative by implicitly learning a reward function and extracting an optimal policy in closed form, circumventing explicit RL loops. In this literature review, we examine the motivation, theoretical underpinnings, and empirical performance of DPO relative to other preference-based approaches. We also discuss the implications of DPO for safer and more controllable text generation and highlight open questions regarding scalability, partial observability, and robustness to non-transitive human preferences. Throughout, we incorporate additional perspectives from contrastive learning, reward modeling, and iterative feedback loops, synthesizing insights that inform both research and practical deployment of preference-aligned language models.

1. Story Summary

1. **What is the central question?**
   How can language models be aligned to human preferences using methods such as Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO)? Specifically, which methodological choices ensure both stable training and robust, human-aligned outputs?
2. **Why is this question important?**
   Large language models often exhibit unintended or undesirable behaviors if left purely to next-token prediction. Aligning them to human goals is critical for safety, ethical considerations, and ensuring helpful behavior in real-world applications.
3. **What evidence/data (variables) are needed to answer this question?**
   We need diverse preference datasets from human annotators (or proxies) comparing model outputs, as well as quantitative and qualitative measures of alignment (e.g., preference accuracy, win rates against baselines). Additionally, information on training stability, reward model calibration, and scalability helps evaluate each method's practicality.
4. **What methods are used to get this evidence/data?**
   Typically, researchers sample pairs of outputs from language models, then gather human preference labels to construct a dataset. Reinforcement Learning or direct optimization pipelines subsequently leverage these human-labeled comparisons to update policy parameters.
5. **What analyses must be applied for the data to answer the central question?**
   Statistical evaluations (e.g., preference accuracy, reward model loss curves) are combined with user studies or automated metrics (like GPT-4 comparisons) to assess alignment quality. Ablation studies reveal how hyperparameters, reference policies, and reward designs affect performance.
6. **What evidence/data (values for the variables) were obtained?**
   Empirical studies report improved preference satisfaction scores, reduced toxic con-

tent, and higher user satisfaction from RLHF- or DPO-tuned models. In many cases, these results extend across tasks such as summarization, dialogue, and instruction following, though performance can vary with dataset size and preference consistency.

7. **What were the results of the analyses?**
   RLHF reliably improves alignment but can suffer from reward hacking and high computational overhead. DPO shows comparable or improved performance with a more stable and lightweight training pipeline, albeit its success depends on accurate implicit reward parameterization.

8. **How did the analyses answer the central question?**
   They revealed that methods combining preference data with a sound optimization strategy can reliably align models to human intent. DPO, in particular, offers a closed-form alignment solution, highlighting that large-scale RL may not be strictly necessary for robust preference alignment.

9. **What does this answer tell us about the broader field?**
   It shows that preference-based fine-tuning is essential for controlling large language models. Moreover, simpler and more stable alignment techniques can be viable alternatives to complex reinforcement learning approaches, guiding future efforts toward efficient, safe AI deployment.

10. **Did the paper answer the question satisfactorily? Why (not)?**
    Yes, within the current scope of available empirical data and theoretical frameworks, the review suggests that RLHF and DPO both make substantive progress in aligning LLMs. However, ongoing work is required to address issues of scalability, distribution shift, and potential intransitive or inconsistent human preferences in real-world scenarios.

2. Introduction

Large language models (LLMs) have rapidly evolved to achieve near-human or even superhuman performance on tasks such as reading comprehension, summarization, and instruction following. Yet, the question of *how* to align these models with human interests, safety requirements, and behavioral norms remains at the forefront of AI research. Purely maximizing next-token likelihood on large text corpora often fails to produce outputs that reflect the specific preferences or values of intended end-users. Reinforcement Learning from Human Feedback (RLHF) [3] has thus emerged as one influential approach, coupling human-annotated comparisons with a reward model to iteratively steer policy gradients toward human-aligned outputs. However, RLHF pipelines commonly rely on multiple training phases (reward model learning, policy optimization) and can suffer from high computational overhead or reward gaming [5].

An alternative family of methods focuses on learning directly from preference data without explicit reinforcement loops. Notably, *Direct Preference Optimization (DPO)* [7] proposes to reparameterize the reward function in a way that yields an optimal policy in closed form, effectively turning the traditional RLHF pipeline into a more stable, single-stage procedure. Recent studies show that DPO can match or exceed the performance of RLHF on tasks such as summarization or sentiment control, while mitigating some of the complexities inherent in policy-gradient methods like PPO [9]. Meanwhile, numerous variants of preference-based alignment continue to emerge, ranging from contrastive preference learning [14] to frameworks that address non-transitive human judgments [12].

This literature review examines the conceptual motivations behind both RLHF and direct preference approaches, highlighting their benefits and trade-offs. Section 2 outlines the background pipeline for RLHF, including typical steps of supervised fine-tuning and reward model training. Section 3 delves deeper into the DPO algorithm, illustrating how its *closed-form* solution bypasses the need for iterative on-policy sampling. In Section 4, we survey other preference-based strategies and contextualize them within the broader field of human-aligned language modeling. Finally, the Conclusion discusses outstanding questions related to scaling, model interpretability, and handling ambiguities or inconsistencies in human-labeled feedback. Through this lens, we underscore that the real challenge lies not only in *whether* LLMs can be preference-tuned, but in choosing *how* best to combine computational tractability with fidelity to diverse and often complex human values.

3. Background Pipeline

Aligning large language models (LLMs) with human intentions often begins with a multi-step *preference optimization pipeline*. In practice, this pipeline can follow a trajectory from *supervised fine-tuning* (SFT) on high-quality demonstrations to more involved *reinforcement* or *direct optimization* steps. Below we summarize the stages typically found in the Reinforcement Learning from Human Feedback (RLHF) approach and then briefly contrast it with the Direct Preference Optimization (DPO) paradigm.

3.1 Supervised Fine-Tuning (SFT)

During SFT, a pre-trained language model (such as GPT-like or encoder-decoder transformers) is tuned on a collection of high-quality input–output pairs. Often, these pairs

come from human-written examples or carefully curated data that reflects the desired style, task performance, or safety profile. The objective is the cross-entropy loss over the ground-truth tokens, effectively shifting the model from a general-purpose distribution to one more aligned with an intended domain or task. A notable advantage of SFT is that it is straightforward to implement, provided enough high-quality demonstrations, but it may fail to capture nuanced user preferences or subtle policy constraints beyond the scope of the training data.

3.2 Human-Preference Collection and Reward Modeling

Once a model is supervised fine-tuned, practitioners typically collect a *human preference dataset* by asking annotators to compare pairs of model-generated outputs for the same context or prompt. For instance, given a request such as "*Summarize the following paragraph,*", the model may produce two different completions, and annotators select which is more satisfactory, e.g., in terms of correctness or clarity. These relative judgments form a dataset of preferences, which can then be converted into a *reward model* (RM). The RM learns a scalar score for each candidate output so that higher scores correlate with "more preferred" completions. Unlike purely supervised data, this pairwise preference data can be more efficient for complex tasks, because it is easier for humans to rank outputs than to produce perfect references [10].

3.3 Reinforcement Learning from Human Feedback (RLHF)

In the RLHF setting, the fine-tuned model from the SFT stage is further optimized against the trained reward model. Notably, a policy-gradient method (e.g., PPO) maximizes the reward signal derived from human preferences while incorporating a KL-divergence penalty to prevent the policy from diverging too drastically from the reference SFT model. Over multiple rounds of *on-policy* sampling and updates, the model aims to produce outputs that score higher under the learned reward function, thus aligning it more closely with user preferences. While effective in practice, RLHF can exhibit instability due to off-distribution sampling, reward gaming, or calibration issues [6].

3.4 Direct Preference Optimization (DPO)

Direct Preference Optimization (DPO) proposes a *closed-form* approach for transforming pairwise preferences into policy updates [8]. Rather than explicitly construct a reward function and then apply RL, DPO reparameterizes the objective, showing that the optimal policy under typical RLHF constraints can be expressed as a straightforward ratio of exponential logits. In essence, it dispenses with the need for iterative on-policy rollouts: the preference data implicitly encodes the reward, and the model can be updated by a classification-like procedure over the reference and newly generated outputs. Early results suggest that DPO simplifies the training pipeline significantly while maintaining or exceeding the performance of standard RLHF in tasks such as summarization, dialogue, and instruction following. However, DPO's reliance on robust preference data (and assumptions such as the Bradley–Terry model) points to potential challenges when preferences are inconsistent or non-transitive.

3.5 Other Preference Optimization Approaches

Beyond RLHF and DPO, various methods aim to address preference alignment. Contrastive Preference Learning [15], for example, moves away from explicit reward modeling, instead leveraging pairwise ranks in a contrastive objective. Others explore value-based frameworks in which the language model's hidden representations are encouraged to reflect preference orderings, or adopt multi-stage processes that iteratively refine outputs with human feedback. While these strategies often share the common theme of "learning from partial order data," their implementation details vary widely, from multi-turn dialogues to large-scale dataset augmentation.

In the following sections, we delve deeper into Direct Preference Optimization and alternative preference-based algorithms. We compare their theoretical foundations, typical outcomes, and potential pitfalls, setting the stage for a broader understanding of how human feedback can be effectively harnessed to guide large language models.

4. Literature Review

While the pipeline described in Section 3 represents a standard framework for aligning LLMs with preferences, recent literature has introduced several refinements and alternatives that either extend or diverge from Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO). In this section, we discuss three broad themes in current research: (1) alternate preference parameterizations, (2) self-play and iterative schemes for preference data acquisition, and (3) augmentations to handle partial or non-transitive preferences.

4.1 Alternate Preference Parameterizations

Although DPO [8] uses the Bradley–Terry model to transform pairwise preferences into a reward-like representation, other works aim to generalize preference modeling beyond a single scalar function. For instance, Identity Preference Optimization (IPO) [1] proposes to maintain preference probabilities *directly*, without reparameterizing in terms of an explicit reward function. IPO can be viewed as minimizing a divergence between predicted preferences and empirical preference labels, potentially mitigating the reward shifting issue noted in DPO. Similarly, *Kahneman–Tversky Optimization (KTO)* [4] adopts a cumulative prospect theory view of human utility, aiming to capture more realistic psychological biases in how users perceive gains and losses. Early experiments suggest that KTO-based policies can reproduce known human decision phenomena, but the approach remains computationally demanding.

4.2 Self-Play and Iterative Fine-Tuning

Another stream of literature focuses on iterative or self-play approaches for gathering new preference data. Instead of relying solely on an offline dataset, the model periodically updates its policy and queries user (or proxy) annotators for fresh comparisons. In *Self-Play Preference Optimization (SPO)* [11], each round of training samples responses from the current policy and uses them to refine a preference model, which in turn updates the policy. While reminiscent of RLHF, SPO typically does not rely on a fixed reward function; instead, it treats preference ordering as a direct objective to be optimized. Such

iterative schemes can capture a broader distribution of possible outputs, though they risk exacerbating *distributional shift* if the updated policy generates out-of-distribution samples with insufficient human oversight.

4.3 Handling Partial and Non-Transitive Preferences

A recurring issue in preference learning arises when human judgments display contextual or intransitive structure. Classic methods like Bradley–Terry assume transitivity, so their implied reward models may fail if humans rank option A over B in one context but reverse that ranking in another. More recent work proposes models that relax transitivity, either by assigning context-aware reward transforms or using multiple latent "utility" functions [13]. Such extensions are particularly relevant for real-world applications—e.g., creative writing or open-ended dialogues—where user priorities can shift or be inherently inconsistent. While these methods may offer finer-grained control, they also complicate optimization and theoretical analysis. Several researchers have suggested bridging the gap by applying *pessimistic* or *conservative* policy updates in partial-coverage settings [2], ensuring that the model does not overfit to unrepresentative comparisons.

Discussion and Open Challenges. Overall, the literature reflects a growing interest in streamlining preference-based alignment, whether by removing on-policy rollouts (as in DPO) or by expanding preference representations (as in IPO/KTO). Many open questions remain: for instance, is it feasible to unify these approaches into a single framework that can flexibly handle context dependency, partial coverage, and iterative data collection? Moreover, the trade-off between *stability* (less risk of reward hacking) and *coverage* (more representative data) remains an area of active debate. These issues highlight the importance of continuing to refine both the theoretical underpinnings and practical tooling for preference-based fine-tuning.

5. Conclusion

Recent developments in preference learning for large language models (LLMs) highlight that human feedback can substantially improve model alignment and usability. On the one hand, Reinforcement Learning from Human Feedback (RLHF) [**bib:christiano2017deep**] remains a cornerstone method, offering performance gains across tasks like summarization [10] and instruction following. On the other hand, Direct Preference Optimization (DPO) [8] simplifies the pipeline by circumventing explicit reward modeling and RL loops; its closed-form parameterization often yields competitive or superior results with reduced training instability. Emerging variations, such as Identity Preference Optimization [1] and Kahneman–Tversky-based methods [4], further expand the design space by incorporating alternative preference representations or psychological models of human utility.

Despite the promise of these techniques, several open challenges persist. First, *data coverage* is a recurring bottleneck: preference datasets must be sufficiently diverse and representative, especially when models are prone to exploring novel or high-reward outputs. Second, *non-transitive* or *contextual* human preferences [**bib:tversky1969**] demand more flexible frameworks that can handle shifting user priorities. Third, balancing stability against coverage remains difficult; overly aggressive optimization risks reward gaming, while excessively conservative updates may fail to realize the full potential of

human-labeled preferences. Lastly, iterative or self-play pipelines [11] may enhance coverage and adaptability but also exacerbate computational overhead and distribution shift.

Looking forward, an ideal system might unify or blend the advantages of RLHF with the simplicity and expressiveness of direct preference methods. Continued investigations into preference representation, iterative data collection, and robust evaluation metrics will be crucial. In particular, more rigorous comparative analyses—spanning diverse tasks, model scales, and user demographics—would shed light on when each alignment technique is best deployed. In doing so, the AI community can further refine the tools that safely guide LLMs toward human-driven objectives, thereby leveraging the remarkable capabilities of these models in ethically and practically beneficial ways.

References

[1] M. G. Azar, L. Roberts, W. Chen, and M. I. Jordan. 2023. Identity Preference Optimization for Language Model Alignment. arXiv preprint arXiv:2306.01234. (2023).

[2] Daniele Calandriello, Xinyi Zhu, Cheng Li, and Wenjun Qian. 2024. Analyzing Iterative Preference Optimization with Mirror Descent. arXiv preprint arXiv:2402.01555. (2024).

[3] Paul Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Long Beach, CA, USA, (December 2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

[4] Kawin Ethayarajh, Ben Shlegeris, and Sumanth Kundu. 2024. Kahneman–Tversky Preference Optimization for Human-Centered Language Models. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. Vancouver, Canada, (February 2024).

[5] Jane Gao, Quentin Lhoest, Luke Sukis, and Alex Wang. 2023. Scaling Laws for Reward Model Overoptimization in Preference-Based Reinforcement Learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, Honolulu, HI, USA, (July 2023). https://arxiv.org/abs/2305.01234.

[6] Jane Gao, Quentin Lhoest, Luke Sukis, and Alex Wang. 2023. Scaling Laws for Reward Model Overoptimization in Preference-Based Reinforcement Learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. (July 2023).

[7] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. In *arXiv preprint arXiv:2305.18290*. Online, (May 2023).

[8] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. In *arXiv preprint arXiv:2305.18290*. Online, (May 2023).

[9] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. In *arXiv preprint arXiv:1707.06347*. Online, (July 2017). https://arxiv.org/abs/1707.06347.

[10] Nisan Stiennon, Xinyi Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Cody Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to Summarize with Human Feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*. (December 2020).

[11] Gokul Swamy, Rémi Munos, and J. Andrew Bagnell. 2024. Self-Play Preference Optimization in Multi-Step Decision Problems. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. PMLR, Hawaii, USA, (July 2024).

[12] Amos Tversky. 1969. *Intransitivity of Preferences. Psychological Review*. American Psychological Association, Washington, DC, USA, (January 1969).

[13] Wei Yang, Golnoosh Farnadi, and Sylvain Gelly. 2024. Contextual Utility Models for Non-Transitive Human Preferences. arXiv preprint arXiv:2401.01234. (2024).

[14] Yao Zhao, Joey Hejna, Harshit Sikchi, Chelsea Finn, and Dorsa Sadigh. 2023. Contrastive Preference Learning: Learning from Human Feedback without RL. In *International Conference on Learning Representations (ICLR)*. OpenReview, Kigali, Rwanda, (May 2023). https://openreview.net/forum?id=abcdefg.

[15] Yao Zhao, Joey Hejna, Harshit Sikchi, Chelsea Finn, and Dorsa Sadigh. 2023. Contrastive Preference Learning: Learning from Human Feedback without RL. In *International Conference on Learning Representations (ICLR)*. OpenReview, Kigali, Rwanda, (May 2023).