JOHANNES KEPLER
UNIVERSITY LINZ

Author
Giovanni Filomeno
12345678

Submission
Seminar in AI (Master)
(365.205)

February 26, 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 **3**

1. Story Summary **3**

2. Introduction **5**

3. Example **5**

4. Conclusion **6**

References **6**

Appendix A. An Appendix **7**

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

Space for your introduction.

3. Example

This is an example for another section in your thesis. It should primarily show you some basic LaTeX tricks. For instance, section 2 references to your introduction. You can also embed figures, tables, etc. in your work. Figure 1 is an example for a figure (i.e. photos, diagrams, and other artwork). Note that figures (just like tables, see Table 1) are floating elements. They are either placed at the top or bottom of a page (or sometimes stand on their own page), but not in between your text. You can influence their placement with the placement modifiers "t", "b", and "p". Bottom placement ("b") is sometimes more tricky, as LaTeX does not consider this nice in some situations. You can convince LaTeX to honor your placement expectation with "!b" then. Since you do not place figures/tables in between your running text, you always need to reference them by their label (e.g. Figure 1 or Table 1).

You will often reference and quote other works. The source of these citations needs to be marked with the \cite{} command. There are different ways to cite the work of others. These are a few examples:

- The Tor directory authority moria1 shows a voting behavior for the HSDir flag that significantly deviates from that of other directory authorities [1]. Be aware that the \cite command is part of the sentence and, hence, comes before the period.
- Roland [3] proposes a novel attack concept against NFC secure elements in mobile phones.

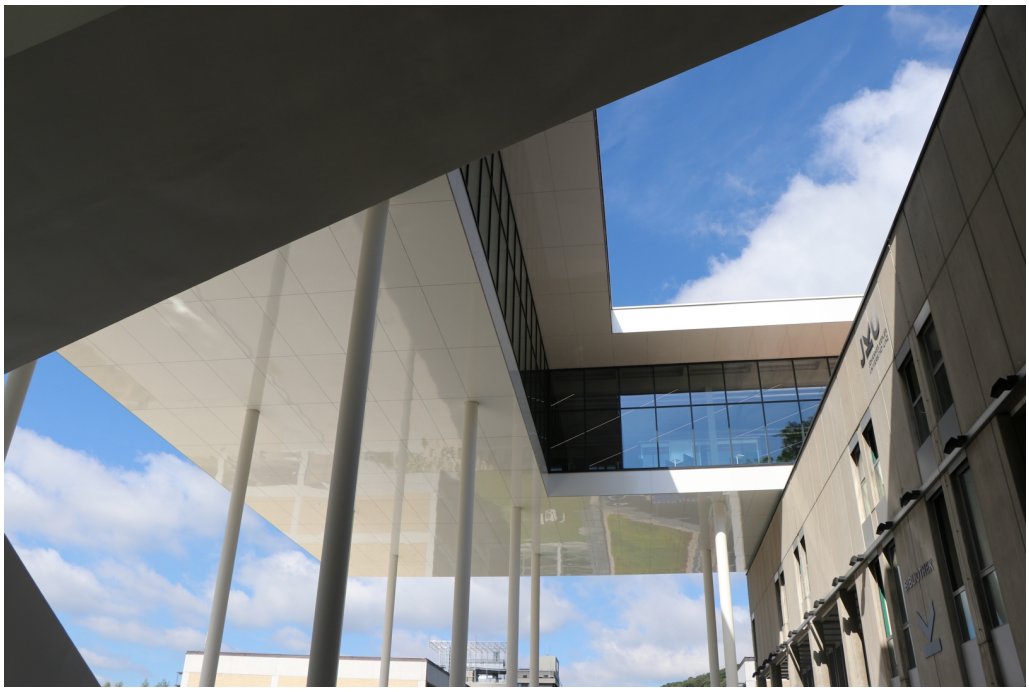

Figure 1: A figure. Be aware that figures have their caption below the artwork

Table 1: A table. Be aware that tables have their caption above the table

Option	Description
`phdthesis`	PhD thesis
`mathesis`	Master's thesis
`diplomathesis`	Diploma thesis
`bathesis`	Bachelor's thesis
`seminarreport`	Seminar report
`techreport`	Technical report

- Roland and Langer [4] uncovered a flaw in legacy-support of the MasterCard contactless payment protocols that allows an attacker to clone certain credit cards.
- Höller et al. [2] designed an experiment to measure the usage of Tor V3 onion services in a privacy-conscious way. Here we used "et al." because the cited work has more than two authors (\citeauthor will automatically take care of this).
- Höller et al. [1] conclude:

  > *Ultimately, the high fluctuations in the hidden service directory were caused by a mixture of several issues. First the changed voting behavior of three directory authorities reduced the amount of obtainable votes to six. If any of the remaining six relays went offline – which tends to happen during ongoing DOS attacks – relays needed to obtain five out of five available votes. So any individual measurement failure regarding either bandwidth or uptime led to a withdrawn HSDir flag.*

- You sometimes also paraphrase from multiple sources. For that purpose, the \cite command accepts a list of multiple comma-separated references. Do not add spaces inbetween them. Various analyses of the Tor network have been performed recently [1, 2].

4. Conclusion

Space for your summary, central conclusions, and an outlook on potential future work.

References

[1] Tobias Höller, Michael Roland, and René Mayrhofer. 2021. Analyzing inconsistencies in the Tor consensus. In *The 23rd International Conference on Information Integration and Web Intelligence (iiWAS2021)*. ACM, Linz, Austria, (November 2021), 10 pages. DOI: 10.1145/3487664.3487793.

[2] Tobias Höller, Michael Roland, and René Mayrhofer. 2021. On the state of V3 onion services. In *Proceedings of the ACM SIGCOMM 2021 Workshop on Free and Open Communications on the Internet (FOCI '21)*. ACM, Virtual, (August 2021), pp. 50–56. DOI: 10.1145/3473604.3474565.

[3] Michael Roland. 2015. *Security Issues in Mobile NFC Devices. T-Labs Series in Telecommunication Services*. Springer, Cham, (January 2015). ISBN: 978-3-319-15487-9. DOI: 10.1007/978-3-319-15488-6.

[4] Michael Roland and Josef Langer. 2013. Cloning Credit Cards: A combined pre-play and downgrade attack on EMV Contactless. In *7th USENIX Workshop on Offensive Technologies*. USENIX, Washington, DC, USA, (August 2013). https://www.usenix.org/conference/woot13/workshop-program/presentation/roland.

Appendix A. An Appendix

Space for an appendix. You can have more than one appendix section. Appendices are, of course, optional.