Comparative Analysis of Preference-Informed Alignment Techniques for Language Model Alignment

Stanford CS224N Custom Project

Soo Wei Koh
SCPD
Stanford University
soowei@stanford.edu

Abstract

This study conducts a thorough comparative analysis of various techniques for aligning language models with human preferences, focusing on Reinforcement Learning from Human Feedback (RLHF), Direct Preference Optimization (DPO), and including newer methodologies like Kahneman-Tversky Optimization (KTO). The study evaluates these techniques on the Mistral-7B model across dialogue and summarization tasks and the AlpacaEval2.0 leaderboard to gauge their effectiveness in aligning LMs with human preferences. The findings reveal that DPO outperforms other methods across tasks, underscoring its efficacy. Meanwhile, Sequence Likelihood Calibration (SLiC) consistently under-performs, highlighting the challenges in calibration-focused approaches. This research not only scrutinizes the strengths and limitations of each method but also explores the incorporation of human decision-making theories into language model training. Through this exploration, we aim to shed light on the complexities of model alignment and propose directions for future enhancements, including a novel experiment with a Tanh specification for the value function in KTO following the use of the logistic function in Ethayarajh et al. (2024).

1 Key Information to include

- Mentor: Tathagat Verma
- External Collaborators (if you have any): NA
- Sharing project: NA

2 Introduction

Large Language Models (LLMs) like GPT have revolutionized the field of natural language processing, demonstrating remarkable capabilities in generating human-like text, comprehending complex contexts, and performing tasks without explicit programming. Yet, their prowess in content generation, derived from expansive datasets, introduces significant challenges in ensuring their outputs align with human preferences.

The complexity of aligning LLMs with human preferences stems from the diverse expectations of users, the enormity of the data these models are trained on, and the potential for generating biased or erroneous content. This discrepancy often reflects the underlying biases in the training data, complicating efforts to ensure that model outputs adhere to ethical standards and meet user expectations across varied applications.

To address this, methods such as Supervised Fine-Tuning (SFT), Reinforcement Learning from Human Feedback (RLHF), and Direct Preference Optimization (DPO) have been developed. Each approach aims to better align LLMs with human values, leveraging different strategies to mitigate

the models' limitations. SFT, despite incorporating Helpful, Honest, and Harmless (3H) data, has struggled to achieve the desired levels of safety and groundedness. RLHF, as highlighted by Bai et al. (2022), offers a targeted alignment through human feedback but demands substantial resources and extensively annotated data. On the other hand, DPO and related methods propose more efficient use of preference data to closely align models with human preferences, showcasing the ongoing challenge of capturing the full spectrum of human values.

Aligning LLMs with human preferences transcends a mere technical challenge; it is fundamental to developing AI that genuinely serves humanity. Imagine AI that not only discerns subtle human nuances but also tailors its interactions to individual preferences. The quest for refining preference optimization methods is crucial for creating AI systems that embody human ethics, cater to our diverse needs, and elevate their effectiveness.

In this study, we undertake a comprehensive evaluation of various preference-informed alignment techniques applied to the Mistral-7B model, aiming to critically assess their strengths and limitations. Our investigation includes a novel experimentation with the Kahneman-Tversky Optimization (KTO) model, where we introduce a Tanh specification as an alternative to the logistic function employed in the original KTO framework (Ethayarajh et al., 2024). The findings from our evaluations reveal that DPO method consistently surpasses other models in achieving alignment with human preferences. Conversely, the Sequence Likelihood Calibration (SLiC) method, despite its innovative formulation calibrating the likelihood of model-generated sequences, falls short of expectations.

3 Related Work

3.1 Reinforcement Learning with Human Feedback

The RLHF method, as described by Ouyang et al. (2022), uses human feedback to train language models to align with human preferences. This involves creating a reward model based on human feedback, then fine-tuning the language model with reinforcement learning.

The training process for LLMs is typically outlined as follows:

- **Pre-training:** The model's initial training phase involves learning foundational language patterns from a vast corpus, optimizing for next-token prediction. This is achieved by minimizing the cross-entropy loss.
- **Supervised Fine-tuning:** After pretraining, the model is fine-tuned on task-specific datasets with supervised learning. This process aims to align the model's outputs with provided responses from the training data, and thus significantly enhance the model's performance in generating appropriate responses for a specific task.
- **Preference-Informed Fine-tuning:** The model is further fine-tuned by integrating human preference alignment to achieve improved alignment. This was initially typically done through reinforcement learning and reward modeling, albeit challenged by its necessity for extensive feedback and sophisticated reward modeling.

Reward Modeling: Training reward models often involves utilizing a dataset comprised of paired comparisons between two responses generated for the same input. The Bradley-Terry model is typically employed, providing a probabilistic framework for estimating the likelihood that one output is preferred over another.

Leveraging a dataset D that consists of human preferences (x, y_w, y_l), where x is the input, y_w is the human-preferred output, and y_l is the less preferred output, we have

$$p^*(y_w \succ y_l|x) = \sigma(r^*(x, y_w) - r^*(x, y_l)), \tag{1}$$

where σ represents the logistic function, and $r^*(x, y)$ is the "true" reward function estimating the human preference for output y given input x. The model is optimized by minimizing the negative log-likelihood of observed human preferences

$$L_R(r_\phi) = \mathbb{E}_{(x,y_w,y_l)\sim D}[-\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))]. \tag{2}$$

This allows the reward model to approximate r^* by learning to predict the probability that one output is preferred over another, as judged by humans.

Reinforcement Learning: Pursuing a straightforward goal of maximizing rewards derived from human preferences can inadvertently compromise linguistic quality and adherence to norms. Instead a KL divergence penalty is typically incorporated to ensure that the fine-tuned model π_θ does not significantly deviate from π_{ref}, thus preserving linguistic integrity. Accordingly, the RLHF loss function can be formalized as

$$L_{\text{RLHF}}(\theta) = -\mathbb{E}_{(x,y)\sim\pi_\theta}[r_\phi(x,y)] + \beta D_{KL}(\pi_\theta(y|x)\|\pi_{\text{ref}}(y|x)), \tag{3}$$

where $\beta > 0$ is a hyperparameter controlling the strength of the KL divergence penalty, $D_{KL}(\pi_\theta(y|x)\|\pi_{\text{ref}}(y|x))$, given by the divergence between the fine-tuned model π_θ and the reference model π_{ref}.

PPO: Proximal Policy Optimization (PPO) initially designed as an online algorithm has inherent challenges for use in RLHF, such as slow processing due to the necessity of sampling generations, and potential instability particularly in distributed settings.

However the PPO-Clipped variant has become a preferred method within the RLHF framework for LLMs. The appeal lies in several key attributes, including sample efficiency, stability and robustness from the clipping mechanism and the ease of implementation and customizations.

Our implementation adheres to that in Ethayarajh et al. (2024), incorporating insights from Baheti et al. (2023), notably maintaining a static reference model to measure updates, utilizing pre-existing dataset avoiding real-time preference generation, and directly assigning +1 for preferred outputs (w) and -1 for less preferred outputs (l) eliminating reward function learning. The adapted PPO loss function is presented as follows

$$L_{\text{PPO}}(\theta) = \mathbb{E}_t\left[\min\left(\frac{\pi_\theta(a_t|x_t)}{\pi_{\text{ref}}(a_t|x_t)}A_t, \text{clip}\left(\frac{\pi_\theta(a_t|x_t)}{\pi_{\text{ref}}(a_t|x_t)}, 1-\epsilon, 1+\epsilon\right)A_t\right)\right] + \beta D_{KL}(\pi_\theta||\pi_{\text{ref}}) + \eta H(\pi_\theta), \tag{4}$$

where:

- $\pi_\theta(a_t|x_t)$ represents the probability of taking action a_t given state x_t under the policy parameterized by θ.
- π_{ref} is the reference model used to generate the ratio of new to old policy probabilities ensuring that updates are measured against a fixed baseline
- A_t is the advantage at time t, which measures how much better taking action a_t is compared to the average action at state x_t.
- The clip function ensures that the ratio $\frac{\pi_\theta(a_t|x_t)}{\pi_{\text{old}}(a_t|x_t)}$ does not exceed the interval $[1-\epsilon, 1+\epsilon]$, which helps in controlling the update step and prevents the policy from changing too drastically.
- $\beta D_{KL}(\pi_\theta||\pi_{\text{ref}})$ is the KL divergence penalty between the current policy and a reference policy π_{ref}, scaled by β to ensure that the updated policy does not deviate too significantly from the reference policy, maintaining stability in the learning process.
- $\eta H(\pi_\theta)$ is the entropy of the policy π_θ, scaled by η. to ensure that the policy to explore by penalizing overly deterministic behavior.

3.2 Direct Preference Optimization

In response to the complexities involved in RLHF, Rafailov et al. (2023) introduced DPO as a streamlined approach for aligning LLMs with human preferences. DPO parameterizes the reward model to facilitate the direct extraction of the optimal policy through a simple classification loss. This approach not only makes the process computationally more efficient but also more stable and easier to implement.

Rafailov et al. (2023) developed closed-form loss functions that effectively maximize the margin between preferred (y_w) and less preferred (y_l) model generations. While there have been other attempts to streamline preference optimization, such as SLiC by Zhao et al. (2023), DPO has gained traction due to its mathematical congruence with RLHF, expressed as:

$$L_{DPO}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}\left[-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{5}$$

where π_θ denotes the policy of the model being optimized, π_{ref} the reference model, and β a scaling factor enhancing the discriminative capability of the model.

The DPO loss function is designed to optimize the policy by comparing the probabilities of the chosen action (the one taken by the policy) versus the rejected action (the one not taken), relative to a reference policy. This comparison is scaled by a factor beta and passed through a logistic function, with the goal of maximizing the probability of the chosen action while minimizing that of the rejected action.

By directly optimizing for human preferences without the intermediate step of explicit reward model training, DPO presents a promising avenue for efficiently and effectively aligning LMs with the nuanced landscape of human preferences.

3.3 Sequence Likelihood Calibration

The SLiC method, as defined in the paper by (Zhao et al., 2023), introduces a calibration loss which is designed to ensure that the probability of the chosen action is not just relatively higher than the rejected action, but also that this probability is sufficiently confident above a threshold. The SLiC loss could be represented as

$$L_{SLIC}(\pi_\theta) = \mathbb{E}\left[\max(0, \beta - \log \pi_\theta(y_w|x) + \log \pi_\theta(y_l|x)) - \lambda \log \pi_\theta(y_w|x)\right] \tag{6}$$

where π_θ is the policy being optimized, β is the threshold for confidence calibration, and λ is the regularization coefficient that balances the calibration with the entropy of the chosen action. While DPO is more about discriminating between chosen and rejected actions relative to a reference, SLIiC is about calibrating the confidence of the policy's decisions against a fixed threshold while maintaining entropy.

3.4 Identity Preference Optimization

Without adequate regularization or mechanisms to ensure diversity in the training data, there is a risk of the DPO model overfitting to the specific preferences represented in the dataset. This can make the model less flexible and potentially less effective when encountering new types of inputs or preferences.

To address the above limitations inherent in DPO, Azar et al. (2023) introduces IPO, incorporating a regularization term to strike a balance between optimizing for human preferences and ensuring model generalizability. IPO amends the DPO loss by introducing a regularization term that constrains deviations from the reference model, thereby preventing overfitting.

The IPO loss function is formulated as

$$L_{\text{IPO}}(\theta, y_w, y_l) = \left(\log\left(\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}\right) - \log\left(\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right) - \frac{1}{2\beta}\right)^2$$

where β is the scaling factor used to adjust the sensitivity of the policy optimization process to the differences in log odds.

The loss L_{IPO} is calculated by taking the square of the difference between the log odds of the policy's probabilities for the chosen and the least favored actions relative to the reference policy, offset by a margin scaled by $\frac{1}{2\beta}$. This ensures that the optimized policy π_θ does not deviate excessively from the behavior of the reference model π_{ref}, promoting stability and robustness during the training phase.

3.5 Kahneman-Tversky Optimization

Ethayarajh et al. (2024) introduced KTO leveraging prospect theory, which describes how humans perceive outcomes in a biased manner, such as being loss-averse. KTO aligns LLMs by directly maximizing the utility of generations, and instead of relying on preference data, requiring only a binary signal of whether an output is desirable or not. KTO's utility-based approach is particularly advantageous in scenarios involving unpaired or sparse data.

The KTO loss function is formulated in the paper as

$$L_{KTO}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{xy\sim\mathcal{D}}[w(y)(1 - v_{KTO}(x, y; \beta))]$$

where

$$r^{KTO}(x,y) = \beta \log\left(\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right),$$

$$z_{\text{ref}} = \mathbb{E}_{x'\sim D}\left[\beta \text{KL}\left(\pi_\theta(y'|x')\|\pi_{\text{ref}}(y'|x')\right)\right],$$

$$v^{KTO}(x,y;\beta) = \begin{cases}\sigma(r^{KTO}(x,y) - z_{\text{ref}}) & \text{if } y \sim y_{\text{desirable}}|x,\\ \sigma(z_{\text{ref}} - r^{KTO}(x,y)) & \text{if } y \sim y_{\text{undesirable}}|x\end{cases},$$

$$w(y) = \begin{cases}\lambda_D & \text{if } y \sim y_{\text{desirable}}|x,\\ \lambda_U & \text{if } y \sim y_{\text{undesirable}}|x\end{cases}.$$

v_{KTO} is the value function based on the Kahneman-Tversky option valuation. $r^{KTO}(x,y)$ calculates the log ratio of the probabilities of generating y given x under the policy π_θ and the reference model π_{ref}, scaled by β. This ratio measures the relative confidence of the policy in generating y compared to the reference model.

The value function applies a logistic to the adjusted ratio of log probabilities, either $r^{KTO}(x,y) - z_{\text{ref}}$ for desirable outcomes or $z_{\text{ref}} - r^{KTO}(x,y)$ for undesirable ones. This determines the value of each generation based on its desirability and the policy's divergence from the reference.

The weight function $w(y)$ assigns weights to the loss based on whether the outcome y is considered desirable or undesirable, with λ_D and λ_U being the weights for desirable and undesirable outcomes, respectively.

The inclusion of β as a parameter in the utility function enables fine-tuning of the model's sensitivity to these aspects, offering a nuanced control over how strongly the model prioritizes avoiding losses over securing gains, in alignment with human behavioral tendencies identified in prospect theory.

Ethayarajh et al. (2024) set v_{KTO} to be the logistic function σ to make it easier to optimize. The loss-aversion coefficient is replaced with two hyperparameters λ_D, λ_U that weight the losses for desirable and undesirable outputs respectively.

In this study, we introduce the Tanh function for the value function v_{KTO}, which increases sensitivity around zero and make the model more responsive to smaller discrepancies or errors compared to the logistic function.

3.6 Relative Preference Optimization

In a recent work, Yin et al. (2024) introduces Relative Preference Optimization (RPO) leveraging both paired and non-paired preference data. RPO employs a contrast matrix for comparing preferred and rejected responses, enabling nuanced distinctions across both identical and related prompts. Weighting strategies are employed to adjust the impact of comparison pairs based on prompt similarities. This method allows for context-sensitive preference learning, prioritizing comparisons from thematically similar prompts. Through these innovative approaches, Yin et al. (2024) shows that the method improves model generalization and alignment with human preferences, ensuring a more nuanced and robust alignment across varied scenarios.

4 Approach

Our methodology focuses on assessing and comparing different techniques for fine-tuning LLMs. Our aim is to gauge the efficacy of these strategies in crafting model outputs that better resonate with human preferences and values.

4.1 Baselines

Our experimental groundwork is the Mistral-7B-v0.1 model, renowned for its robust performance and cutting-edge capabilities within the research community. This model serves as our foundational pre-trained baseline. For comparative analysis, we employ the SFT model as the baseline for dialogue and summarization tasks, evaluating the improvements offered by the advanced alignment techniques.

4.2 Contributions

This research conducts a comprehensive comparative analysis of various preference-informed optimization methodologies, systematically applying them to uniform preference datasets. Our analysis delves into the distinct advantages and challenges inherent to each method, illuminating how they each contribute to aligning LLMs with the complex and varied spectrum of human preferences. Notably, this study pioneers the exploration of a tanh specification for Kahneman-Tversky Optimization (KTO), investigating its impact on model alignment efficiency compared to the logistic function used in Ethayarajh et al. (2024). Furthermore, our investigation broadens to include performance evaluations on the AlpacaEval2.0 leaderboard, offering insights into the methods' scalability and effectiveness across a standardized set of criteria.

4.3 Implementation

For the training implementation, we leverage and modify the existing code-base at relative-preference-optimization repository on GitHub as a foundation for our work. This base has itself significantly built upon the DPO and KTO repositories.

5 Experiments

This section delineates the methodology and datasets underpinning our comprehensive evaluation of preference-informed alignment techniques. By leveraging diverse datasets and employing a rigorous evaluation framework, we aim to uncover the nuances of each alignment strategy's effectiveness in refining LLMs to mirror human preferences.

5.1 Data

In our study, we draw upon two pivotal datasets, each selected for its relevance to specific aspects of open-ended text generation tasks.

- Anthropic's Helpful and Harmless (HH) Dataset (Bai et al., 2022): This dataset was utilized for assessing single-turn dialogue performance of the models. The dataset contains 170k dialogues, each comprising a human query and paired model responses rated for helpfulness and harmlessness.
- OpenAI's Summarization Dataset (Stiennon et al., 2020): This dataset was utilized for the summarization task, each input x in the dataset is a substantive forum post, and the task for the model is to generate a concise summary y.

Both datasets are in paired preference format. The SFT phase was informed by preferred responses from this dataset. For PPO and KTO that utilize unpaired binary data, we convert preference data $y_w \succ y_l$ by assuming that y_w is drawn from the desirable distribution and y_l from the undesirable one.

5.2 Evaluation Method

We conducted an assessment of the preference-informed alignment techniques using the validation sets from Anthropic's Helpful and Harmless (HH) Dataset for dialogues and the OpenAI Summarization Dataset for summarization tasks. Following Yin et al. (2024), we also incorporated the AlpacaEval2.0 leaderboard (Li et al., 2023) into the evaluation framework to asses the model's adaptability and overall capability in following instructions.

The main metric for our evaluation was the win rate, with GPT-4 serving as the evaluation tool. This metric quantitatively gauged the preference rate of our model's responses in comparison to those generated by the SFT targets (i.e. the alignment datasets). We use GPT-4 to assess whether the response from the aligned model is superior to the SFT target within the given context. Through this win rate comparison, we determine the extent to which the outputs from the aligned models surpassed the SFT target, aligning with the specified evaluation standards.

5.3 Experimental details

The training utilized 4 Nvidia A100 GPUs, with a batch size of 64, optimized with RMSProp optimizer. The initial phase involves training SFT models, succeeded by the application of preference-informed alignment strategies.

All models are aligned under identical settings on the same data, save for hyperparameters unique to them. We maintain a consistent beta value ($\beta = 0.1$) which, in DPO, IPO, and KTO, acts as a scaling factor that modulates the sensitivity of the model to human preferences, influencing the balance between preference alignment and model generalizability.

The detailed hyperparameters are presented in Table 1. We train the models for 1 epoch. The number of samples employed for calculating the win rate is established at 128.

Table 1: Hyperparameters.

Hyperparameters	Value
Batch size	64
GPUs	4
Learning rate	5e-7
Epochs	1
Max prompt length	256
Max prompt length + Max response length	512
Optimizer	RMSprop
β	0.1
Sampling temperature	0
GPT judge	gpt-4-0125-preview
AlpacaEval judge	alpaca_eval_gpt4_turbo_fn

5.4 Training and Evaluation Details

Table 2 encapsulates key observations gleaned from the training logs, highlighting each method's unique learning dynamics, challenges encountered, and indications of potential areas for refinement.

Table 2: Training Logs

Method	**Observations**
RLHF	Exhibits instability with fluctuating policy entropy and stagnant critic loss, indicating challenges in effective learning and policy development. Requires finetuning of parameters for better stability.
DPO	Demonstrates progressive optimization and effective discrimination between preferred outcomes, with sustained learning despite occasional volatility. Minor instabilities suggest a need for occasional adjustments.
SLIC	Struggles with effectively differentiating between varying reward decisions, showing minimal improvement and potential issues with generalization due to variability in rewards and gradient norms.
IPO	Shows modest learning with consistent, narrow reward ranges and slight improvements over time. The lack of significant progress in accuracies points to a need for strategic adjustments in training.
KTO (logistic)	Indicates stable learning with successful differentiation between desirable and undesirable outputs. Observations of increasing gradient norms across runs highlight potential convergence issues.
KTO (Tanh)	Reveals effective policy refinement and a more consistent training process compared to the logistic variant, with less variability in rewards and loss, indicating a smoother optimization process and potentially superior performance.

5.5 Results

Table 3 offers a comparative analysis of win rates for alignment methods applied to Mistral-7B model, addressing tasks across the Anthropic-HH dataset, OpenAI Summarization dataset, and the AlpacaEval2.0 leaderboard.

Table 3: Win rate on Anthropic-HH and OpenAI Summarization datasets and AlpacaEval2.0 leaderboard

Method	Anthropic-HH	OpenAI Summarization	AlpacaEval2.0
SFT	51.8	20.7	13.0
PPO	59.6	44.2	13.6
DPO	71.4	53.8	31.0
SLIC	57.3	23.9	17.4
IPO	65.8	43.6	20.7
KTO(Logistic)	59.5	40.7	16.8
KTO(Tanh)	59.5	42.1	18.1

SFT: The base performance established by SFT's performance is relatively lower across all datasets, reflecting its limited capacity to align model outputs closely with human preferences without preference-informed training, though it has a more than 50% win rate over the baseline SFT Target.

PPO: PPO exhibits an improvement over SFT in both the Anthropic-HH and OpenAI Summarization tasks, highlighting the benefits of its reinforcement learning approach allowing PPO to more dynamically align the model's outputs with human preferences. However, its performance on the AlpacaEval2.0 leaderboard, indicates limitations in capturing the full range of preferences and nuances required for the task. Moving beyond binary reward signals used in this study to a more nuanced or continuous reward scale is likely to enhance PPO's learning efficacy.

DPO: DPO achieves the highest win rates across all evaluations, underlining its efficacy in closely aligning model outputs with human preferences. DPO's strategy of directly optimizing based on preference data seems to allow for a more nuanced understanding and generation of preferred outcomes, leading to significantly better performance across diverse evaluation benchmarks.

SLiC: SLiC demonstrates moderate improvements over SFT , with its calibration approach falling short in the summarization task which requires deep content understanding. SLiC's focus on calibrating the model's output probabilities to match human preferences might not fully capture the complexity of preferences in tasks with higher demands on content generation and structure. Experimenting with different settings of the parameters to optimize the balance between calibration strength and model regularization could enhance SLiC's capability.

IPO: IPO's strong performance suggests its effectiveness in modeling preferences with an emphasis on regularization to prevent overfitting. However, its relatively weaker performance for the summarization task indicates indicates potential areas for refinement, especially in tasks that require capturing nuanced or complex preferences. Developing task-specific regularization strategies may help its performance in tasks with complex preference structures.

KTO (Logistic): The model exhibits moderate performance across the datasets. The logistic function's characteristic curve might limit the model's sensitivity to changes in preferences, especially when preferences are subtle or when the difference between desirable and undesirable outcomes is not stark. This can affect the model's ability to finely tune its outputs according to human judgments.

KTO (Tanh): The Tanh variant of KTO shows a slight improvement in performance over its logistic counterpart, particularly in handling nuances and variability in human preferences. The Tanh function's symmetric output range might have offered a more nuanced sensitivity to preference distinctions, potentially providing a more flexible and responsive utility modeling approach.

6 Analysis

The analysis of preference-informed alignment techniques offers insights into the strengths, weaknesses, and potential areas for improvement in aligning machine learning models with human preferences.

DPO Outperforms Other Methods: DPO's simplicity and directness in optimizing for human preferences make it highly effective across various tasks. Its approach of directly optimizing the margin between preferred and less preferred outcomes leads to superior performance, highlighting the benefit of a straightforward, preference-focused optimization strategy.

Challenges and Innovations of KTO: The performance of KTO models, inspired by human decision-making biases, particularly loss aversion, indicates that the complexity of accurately modeling and implementing these human biases might limit their effectiveness compared to more direct optimization methods.

Importance of Model Flexibility and Adaptability: The performance variations across tasks underscore the importance of model flexibility and adaptability. Models that can dynamically adjust their parameters or strategies based on specific task requirements or feedback characteristics (like DPO and IPO) tend to perform better across diverse benchmarks.

SLiC's Calibration Challenge: SLiC's focus on calibrating model outputs to align with human preferences stumbles in complex content generation tasks, suggesting that calibration alone may not suffice. This revelation points towards the necessity of integrating calibration with additional optimization or feedback mechanisms to enhance content generation capabilities.

Sensitivity to Hyperparameters Settings: Models, especially those like KTO and SLiC, which rely on specific parameters highlighted the sensitivity of model outcomes to hyperparameter settings. The systematic tuning and possibly dynamic adjustment of these parameters based on task or performance feedback could improve outcomes. For models like IPO, which incorporate regularization to balance preference alignment with generalizability, finding the right balance is key. Overemphasis on regularization could dampen the model's ability to learn from specific feedback, affecting performance.

Adaptive Modeling and Continuous Alignment: The dynamic nature of human preferences and the complexity of tasks also underscore the importance of continuous model evaluation, feedback integration, and iterative refinement. Models that can adapt based on ongoing feedback and performance assessments are better positioned to maintain alignment with human preferences over time.

7 Conclusion

This research underscores the importance of direct, nuanced preference modeling, the potential benefits of incorporating human decision-making insights into model optimization, and the importance of model flexibility and adaptability. The findings suggest avenues for future research, including the development of more sophisticated feedback mechanisms, exploration of hybrid model approaches, and continuous model adaptation strategies to enhance alignment with human preferences across a wide range of tasks.

8 Future Work

Future research should concentrate on refining loss functions and optimization strategies to more accurately reflect the complexity of human preferences, by incorporating advanced insights from fields such as behavioral economics and psychology. Alongside, developing more sophisticated feedback mechanisms is essential, particularly those that enable nuanced and contextual understanding of user inputs, thereby enhancing model alignment with human judgments. Moreover, emphasizing continuous adaptation and learning will ensure that models remain dynamically aligned with evolving language patterns and user expectations over time. Together, these focus areas could significantly advance the capability of language models to understand and generate language that resonates deeply with human users, paving the way for more interactive, adaptive, and personalized AI systems.

References

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*.

Ashutosh Baheti, Ximing Lu, Faeze Brahman, Ronan Le Bras, Maarten Sap, and Mark Riedl. 2023. Improving language models with advantage-based offline policy gradients. *arXiv preprint arXiv:2305.14718*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. *GitHub repository*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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. *arXiv preprint arXiv:2305.18290*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *NeurIPS*, 33:3008–3021.

Yueqin Yin, Zhendong Wang, Yi Gu, Hai Huang, Weizhu Chen, and Mingyuan Zhou. 2024. Relative preference optimization: Enhancing llm alignment through contrasting responses across identical and diverse prompts. *arXiv preprint arXiv:2402.10958*.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.