Orchestrating Human and AI Feedback: PCUI-DPO for Human-Aligned LLM Responses

Anonymous ACL submission

Abstract

This paper proposes a novel approach to Large Language Model (LLM) training that prioritizes AI-generated responses, reducing reliance on extensive human feedback. We introduce the Predicted Confidence and Uncertainty Index (PCUI) metric, offering a new dimension of LLM interpretability by capturing both confidence and uncertainty in generated text. Integrating PCUI into Direct Preference Optimization (DPO) guides the model towards favoring its own high-confidence responses during training. Notably, a confidence threshold is established using PCUI, enabling the model to prioritize AI-generated responses exceeding the threshold over human-provided feedback. This approach promotes a gradual shift towards automated LLM training with interpretability and control. We demonstrate the effectiveness of this method in text generation tasks, achieving significant improvements in performance. This work lays the groundwork for a future where AI and human feedback collaborate to create more robust and user-centric LLMs.

1 Introduction

LLMs have revolutionized various fields, from generating human-quality text to powering advanced chatbots. However, their training process remains heavily reliant on extensive human feedback, a bottleneck hindering faster model development and broader application. This paper presents a groundbreaking approach that orchestrates AI and human feedback while gradually reducing dependence on human intervention. The core of our approach lies in integrating the PCUI metric into the DPO framework. This integration demonstrably guides model training towards favoring AI-generated responses. The PCUI metric not only influences training but also serves as a valuable tool for evaluating reward models, leading to significant performance improvements. Our key contribution lies in employing the PCUI metric to establish a confidence threshold. This threshold dictates when the model prioritizes its own responses during training over those provided by human feedback.

- **PCUI Metric**: The PCUI metric goes beyond traditional evaluation methods by quantifying an LLM's confidence in its generated text alongside the inherent uncertainty. This allows for a more nuanced understanding of the model's reasoning process. The specific formulation of the PCUI metric detailed in the further section leverages advanced entropy calculation.

- **DPO with PCUI Integration**: We incorporate the PCUI metric into the DPO loss function. This modified loss function penalizes the model for situations where the predicted PCUI score for a human-provided response is higher than that of the AI-generated response. Subsequently, prioritizing high-confidence AI-generated particularly when the AI-generated response score exceeds the confidence threshold. Ensuring that only the most reliable AI-generated responses are used for further training, thereby promoting steady improvement in the model's ability to generate human-aligned text. The confidence threshold is a crucial hyperparameter that determines the degree to which the model prioritizes its own responses. By carefully calibrating the threshold, we can achieve a balance between model exploration (trying new responses) and exploitation (focusing on high-confidence outputs).

2 Related Work

Significant research has been conducted on optimizing LLM training and reducing reliance on human feedback. Techniques like RLHF (Ouyang et al., 2022a) involve human experts providing feedback to reinforcement learning agents. While effective,

RLHF can be complex to design and requires careful selection of human rewards. Our approach utilizes the PCUI metric within DPO, offering a more automated and interpretable way to incorporate human preferences. Recent work on Reinforcement Learning from AI Feedback (RLAIF) (Lee et al., 2023) addresses challenges in defining reward functions by leveraging AI-generated feedback. However, this method requires training an additional AI model for feedback generation. The PCUI metric offers a simpler and more precise way to evaluate reward models, improving their performance in our approach.

3 Predicted Confidence Uncertainty Index (PCUI) Formulation

Traditional evaluation methods for LLMs often focus solely on accuracy or human judgment. These approaches provide limited insights into the LLM's internal reasoning process. The PCUI metric addresses this gap by offering a multi-faceted assessment of LLM response quality.

- **Logit Values** (z_i): The LLM's internal workings often culminate in a set of raw scores (logits) for each possible output (represented by i). These logits embody the model's unnormalized preference for each outcome.

- **Softmax Activation**: The softmax function takes these logits (z_i) and transforms them into probabilities (P_i) for each potential response (i). The key property of the softmax function is that it ensures the probabilities sum to 1. Mathematically, this transformation is represented by the equation:

$$P_i = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{1}$$

- **Maximum Probability (Confidence)**: The confidence score within PCUI is derived from the softmax output. We identify the highest probability value P_i amongst all the generated probabilities. This maximum probability signifies the outcome the LLM is most confident about, representing its primary belief for the response.

$$\text{Confidence} = \max_i(P_i) \tag{2}$$

  where:
  - P_i is the probability for class i.

- **Uncertainty**: Entropy is a concept borrowed from information theory. In the context of PCUI, it measures the degree of uncertainty associated with the LLM's response probabilities P_i obtained from the softmax function.
  - High entropy indicates a more even distribution of probabilities across potential responses, suggesting the LLM is unsure about the most appropriate output.
  - Conversely, low entropy signifies a distribution skewed towards the most probable response (identified in the confidence calculation), implying the LLM is confident in its output.

$$\text{Entropy} = -\sum_i P_i \cdot \log(P_i + \epsilon) \tag{3}$$

  where, ϵ is a small value added to avoid logarithm of zero.

By combining the confidence and uncertainty scores, the PCUI metric provides a comprehensive assessment of LLM response quality. A high PCUI score indicates a response where the LLM is both confident (high maximum probability) and certain (low entropy). Conversely, a low PCUI score suggests a response with either low confidence or significant underlying uncertainty, potentially requiring further refinement during training. Mathematically, the PCUI score is computed as follows:

$$\text{PCUI} = \frac{\text{Confidence}}{\text{Uncertainty} + \epsilon} \tag{4}$$

where ϵ is a small constant to avoid division by zero.

3.1 Integrating PCUI into DPO

Direct Preference Optimization (DPO) (Rafailov et al., 2023) is a method used to align LLMs with human preferences by optimizing the model's outputs directly based on preference data. This work integrates the PCUI metric into the DPO framework by modifying the loss function.

3.1.1 Model Setup

- **Base Model** (ϕ): This pre-trained LLM acts as the foundation, generating initial responses to prompts.

- **Fine-Tuned Model** (ϕ_{SFT}): This ϕ_{SFT} is derived from the base model (ϕ) and undergoes further fine-tuning within the DPO framework with PCUI integration. This fine-tuning process refines the model's ability to generate human-preferred responses.

3.1.2 Dataset and DPO-PCUI Training

To establish robust PCUI-DPO models, we employed a two-step training approach:

- **Initial Training**: We leveraged the (Lambert et al., 2024) dataset, containing both human-written responses (chosen samples) and AI-generated responses (rejected samples). This dataset provided a foundation for training two distinct PCUI-DPO models namely Phi-2 and Meta-Llama-3-8B-Instruct after SFT training them.

- **Prioritizing High-Confidence AI Feedback**: During this stage, the model is primarily tailored to consider AI-generated responses only as feedback. The selection of these AI responses was guided by a confidence threshold ranging from 0.3 to 0.7. Samples with PCUI AI responses more than this threshold, signifying high confidence in the LLM's output, were prioritized over those with lower confidence PCUI scores. This strategic selection mechanism empowers the model to leverage LLMs own successes for further refinement.

The modified DPO loss function is defined as:

$$\mathcal{L}_{\text{DPO}} = -\log\left(\sigma\left(\beta\left(\text{PCUI}_{\text{chosen}} \cdot \left(\log P_{\phi_{\text{SFT}}}(\text{chosen}) - \log P_{\phi}(\text{chosen})\right)\right)\right.\right.$$
$$\left.\left. - \text{PCUI}_{\text{rejected}} \cdot \left(\log P_{\phi_{\text{SFT}}}(\text{rejected}) - \log P_{\phi}(\text{rejected})\right)\right)\right)$$

where σ is the sigmoid function and β is a scaling factor.

4 Results

To establish a strong foundation, we first evaluated existing reward models using the PCUI metric. This evaluation provided crucial insights into their effectiveness in capturing the quality of LLM responses. Table 1 shows their respective PCUI scores. As observed in the table, all models except NeuralHermes-2.5 - Mistral-7B exhibited a clear distinction between the PCUI scores for chosen and rejected responses. This indicates their ability to differentiate between high-quality and low-quality LLM outputs to some degree.

Model	Chosen	Rejected
zephyr_7b_gemma	98.15735816955566	92.8972840309143
nous_hermes_2_mistral_7B_DPO	97.862309217453	91.46125912666321
tulu-2_dpo-7b	97.45848178863525	93.49875450134277
zephyr_7B_beta	97.41134643554688	93.4740662574768
zephyr_7B_alpha	96.75858616828918	92.48956441879272
zephyr-7b_ppo	96.17967009544373	90.06326794624329
NeuralHermes-2.5 - Mistral$_7B$	56.2851011753082	55.83584904670715

Table 1: Performance metrics for Reward Models

4.1 Evaluation Metrics

To evaluate the effectiveness of the PCUI-DPO approach, we conducted a series of experiments comparing it against the baseline DPO method. Our experiments were performed on text generation tasks using the Phi-2 and Meta-Llama-3-8B-Instruct model. We also notice by setting thresholds to prefer AI generated feedback over human feedback, the model is able to drift towards generating a much more tailored human-like AI response.

4.2 Quantitative Metrics

The BLEU, ROUGE-L, and METEOR scores were calculated for both the PCUI-DPO and normal DPO models. The results, as depicted in 3 and 4, demonstrate a significant improvement in all three metrics for the PCUI-DPO models compared to the normal DPO models:

- BLEU Score: The PCUI-DPO models achieved a higher BLEU score, indicating better alignment with the reference texts and higher precision in word matching.

- ROUGE-L Score: The PCUI-DPO models outperformed the normal DPO model in terms of ROUGE-L, which measures the longest common sub sequence between the generated and reference texts, suggesting better recall.

- METEOR Score: The METEOR score, which considers synonyms and stemming, was also higher for the PCUI-DPO models, reflecting better semantic matching and robustness in the generated responses.

4.3 Qualitative Evaluation

Using GPT-4, we assessed the generated responses on various qualitative criteria. The average scores of both DPO and PCUI-DPO based models over multiple epochs for various criteria are presented in 1. The PCUI-DPO model consistently achieved higher scores across all criteria. Further, experiments with GPT-4 show that higher confidence thresholds for AI-generated feedback lead to better

performance. A threshold of 0.7 yielded the best results in coherence, relevance, correctness, and fluency of AI feedback. Whereas, a threshold of 0.3 shows only moderate improvements as the model begins to prioritize AI-generated feedback with lower confidence scores. Threshold 0.4 shows noticeable improvements in all criteria, showing better performance as the model becomes more selective with higher-confidence AI feedback. Threshold 0.5 achieves significant enhancements in all evaluation metrics, indicating a balance between inclusivity and selectivity of AI feedback is beneficial. Threshold 0.6 depicts Higher scores across all criteria as the model primarily considers feedback with very high confidence, leading to superior performance. This suggests that PCUI-DPO effectively utilizes high-confidence AI responses to improve LLM performance as depicted in 2.

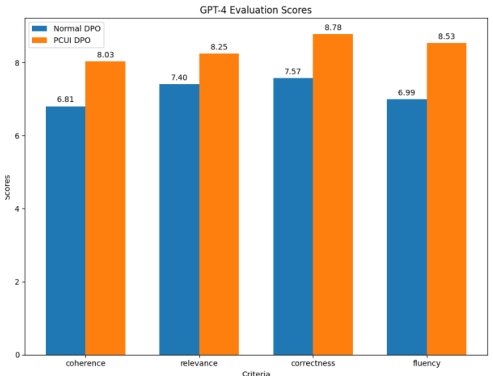

Figure 1: GPT-4 Evaluation Scores: Normal DPO vs PCUI-DPO

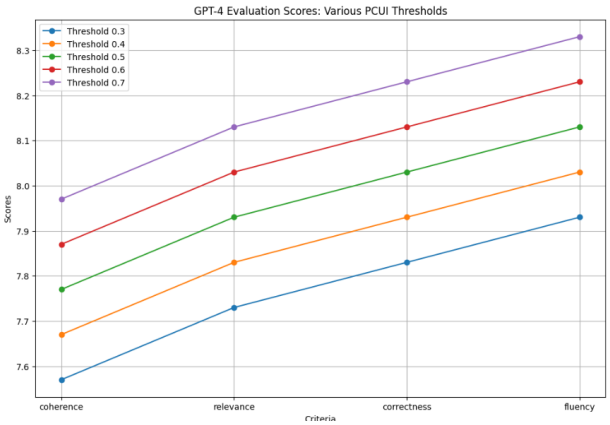

Figure 2: GPT-4 Evaluation Scores: GPT-4 Evaluation Scores: Various PCUI Thresholds

5 Conclusion

In conclusion, the PCUI-DPO approach presents a promising avenue for advancing LLM training.

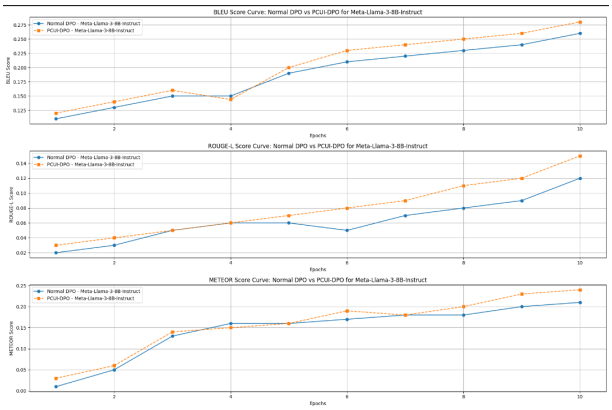

Figure 3: BLEU, ROUGE-L, and METEOR Score Curves: Normal DPO vs PCUI-DPO for Meta-Llama-3-8B-Instruct

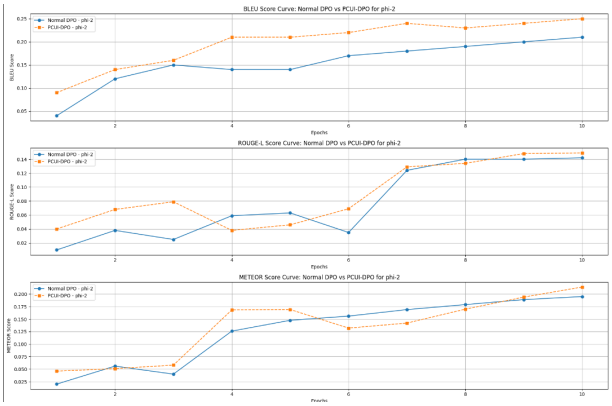

Figure 4: BLEU, ROUGE-L, and METEOR Score Curves: Normal DPO vs PCUI-DPO for Phi-2

By incorporating the PCUI metric to quantify confidence and uncertainty, we gain valuable insights into the LLM's reasoning process. This, coupled with the strategic use of AI-generated feedback through confidence thresholds, empowers the model to refine its abilities and generate human-aligned responses. The effectiveness of PCUI-DPO is demonstrably evident in the improved performance on text generation tasks, as measured by established metrics like BLEU, ROUGE-L, and METEOR. As we move forward, PCUI-DPO paves the way for a future where human and AI collaboration flourishes, fostering the development of LLMs that are not only powerful but also capable of generating human-quality text in an efficient and automated manner.

6 Appendix

6.1 Related Work

(Shen et al., 2023) surveys the technologies for aligning large language models with human values to leverage their vast potential while minimizing risks.

(Bakker et al., 2022) investigates using LLMs to aid humans in collectively finding agreement by fine-tuning to align outputs with diverse human preferences. (Wang et al., 2023) provide a comprehensive overview of alignment technologies for large language models, summarizing various methods and their effectiveness.

(Song et al., 2024) explores fine-tuning large language models with diverse data to enhance human alignment, mitigating the risk of generating toxic or offensive content. You can find the paper

(Yuan et al., 2023) proposes a new learning paradigm named RRHF for LLMs to align with human preferences using rank responses, enhancing model alignment with human feedback.

(Han et al., 2024) discusses the challenges and methods for aligning LLMs in the medical field to ensure their safety and effectiveness, addressing specific weaknesses in general-knowledge LLMs.

(Ouyang et al., 2022b) explores fine-tuning LLMs with human feedback to align with user intent across a wide range of tasks.

(Sun et al., 2023) discusses a method for self-aligning language models from scratch, reducing dependency on intensive human supervision.

These papers provide a comprehensive overview and various methods for aligning large language models with human values, preferences, and safety standards. Our work builds upon these existing techniques by introducing the novel PCUI metric and its integration with DPO. This allows for:

- Gradual Automation: We propose a progressive shift towards automated LLM training, starting with high-confidence responses (identified through PCUI) and gradually reducing reliance on human feedback. This builds upon existing approaches by offering a path towards eventual automation.

- Interpretability and Control: The PCUI metric offers a unique window into the LLM's reasoning process, enabling better model control and interpretability compared to solely relying on reward functions or human annotations, which often lack transparency.

- Threshold-Based Prioritization with Interpretability: The confidence threshold allows for fine-grained control over how much the model prioritizes its own high-confidence responses during training, addressing the challenge of defining appropriate reward functions in reward learning by leveraging the model's own confidence estimation in a more interpretable way.

By combining these elements, our research proposes a novel and impactful approach for LLM training, paving the way for a future where AI and human feedback collaborate for more efficient and robust model development

6.2 Limitations

- Data Bias: The effectiveness of PCUI-DPO hinges on the quality of training data used to establish the initial PCUI model. If the data is biased, the LLM may inherit those biases and prioritize AI-generated responses that reflect those biases, even if they are not human-aligned.

- Limited Scope: The current evaluation focuses on text generation tasks. Further research is needed to determine how PCUI-DPO generalizes to other LLM functionalities, such as question answering or summarization.

- Confidence Threshold Calibration: Finding the optimal confidence threshold can be challenging. While a higher threshold generally leads to better results in our study, it might not be universally applicable across all tasks and LLM architectures. Further research is required to develop more robust methods for calibrating the threshold effectively.

- Explainability of AI Feedback: While PCUI-DPO leverages AI-generated feedback, understanding the rationale behind these responses remains a challenge. Future work could explore techniques to make AI feedback more interpretable, allowing for more targeted improvements in the LLM.

References

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John

Aslanides, Matt Botvinick, and Christopher Summerfield. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems*, volume 35, pages 38176–38189. Curran Associates, Inc.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Medsafetybench: Evaluating and improving the medical safety of large language models. *Preprint*, arXiv:2403.03744.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Rewardbench: Evaluating reward models for language modeling. *Preprint*, arXiv:2403.13787.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *Preprint*, arXiv:2309.00267.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022a. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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. *Preprint*, arXiv:2305.18290.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *Preprint*, arXiv:2309.15025.

Feifan Song, Bowen Yu, Hao Lang, Haiyang Yu, Fei Huang, Houfeng Wang, and Yongbin Li. 2024. Scaling data diversity for fine-tuning language models in human alignment. *Preprint*, arXiv:2403.11124.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Advances in Neural Information Processing Systems*, volume 36, pages 2511–2565. Curran Associates, Inc.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *Preprint*, arXiv:2307.12966.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 10935–10950. Curran Associates, Inc.