Extended Abstract

Motivation Language models are seeing rapidly growing use across a wide range of everyday, academic and professional applications. This widespread adoption has enabled the collection of large datasets capturing direct user feedback on the quality and appropriateness of model outputs. Recent work has shown the potential to leverage this user feedback data to improve language models via reinforcement learning frameworks such as RLHF (Reinforcement Learning from Human Feedback). As an extension to themes in this work, the emerging paradigm of Constitutional AI aims to constrain model behavior to align with a set of pre-defined principles or values, helping to constrain uwanted model behavior. By flipping the task inherent to Constitutional AI on its head, Inverse Constitutional AI seeks to infer such latent principles from user feedback data. Here we present a novel framework that combines elements of Constitutional and Inverse Constitutional AI to guide language model outputs to better reflect human preferences. The combination of concepts from RLHF, Constitutional AI, and Inverse Constitutional AI make possible a framework where-in language models can be trained in a fashion that appropriately constrains behavior with minimal developer supervision.

Method In order to demonstrate this, we first apply Inverse Constitutional AI techniques to a dataset of human-labeled response preference pairs to learn a set of guiding principles. We then provide these learned principles as additional input during fine-tuning of a language model, with the goal of steering the model to generate outputs more consistent with the preferences reflected in the source data. Through experiments on both open-ended language generation and targeted reasoning tasks, we validate that this Learned-Principles-Guided-Outputs approach quantitatively improves alignment of model-generated text with the learned principles while preserving overall generation quality.

Implementation Using the Qwen2.5-0.5B model as base, a reference model was created by performing supervised fine-tuning on `HuggingFaceTB/smol-smoltalk`, a dataset of user prompts and model responses designed specifically for the training of smaller language models. Using the algorithm provided in the authors' paper, Inverse Constitutional AI was performed by querying OpenAI's gpt-4o-mini model for sets of principles explaining why a preferred response was chosen over its paired rejected response for a subset of pairs in `HuggingFaceH4/ultrafeedback_binarized`, a dataset containing preferred and dis-preferred preference pairs. To avoid redundant principles, per-sample principles were then clustered using k-means clustering on text representations generated by OpenAI's text-embedding-ada-002 model, and one candidate principle was sampled from each cluster. Candidate principles were evaluated for usefulness in explaining preference across samples in the dataset by prompting gpt-4o-mini with the principle as well as the chosen and rejected responses (unidentified). gpt-4o-mini was asked to identify, based solely on the provided principle, whether response 1 was preferred, response 2 was preferred, or if the principle was largely inapplicable to comparison of the two responses. Principles found to be inapplicable in over 90% of samples were discarded, and the remaining principles were ranked based on the frequency of correct preferred response selections using the principle. Finally, the top n principles were selected for the constitution.

Using a frozen and unfrozen version of the fine-tuned model as reference and trainable policies (respectively), direct preference optimization was performed using the data from `HuggingFaceH4/ultrafeedback_binarized` on two versions of the dataset: a standard implementation, and one where user prompts were augmented with the principles generated in the previous state.

Results RUNNING NOW! include end eval loss, eval accuracy, and leaderboard score

Conclusion Our framework shows that a general use language model can be used to distill a set of guiding principles from a human preference dataset, allowing for improved fine-tuning of a model tailored to that dataset in an unsupervised fashion. This has wide reaching implications for improving and constraining model behavior, especially in use cases for smaller models.

Guiding Language Model Outputs via Principles Learned from User Feedback

Justin Adjasu
Department of Biomedical Data Science
Stanford University
jadjasu@stanford.edu

Abstract

Language models are seeing rapidly growing use across a wide range of everyday, academic and professional applications. This widespread adoption has enabled the collection of large datasets capturing direct user feedback on the quality and appropriateness of model outputs. Recent work has shown the potential to leverage this user feedback data to improve language models via reinforcement learning frameworks such as RLHF (Reinforcement Learning from Human Feedback). As an extension to themes in this work, the emerging paradigm of Constitutional AI aims to constrain model behavior to align with a set of pre-defined principles or values, helping to constrain uwanted model behavior. By flipping the task inherent to Constitutional AI on its head, Inverse Constitutional AI seeks to infer such latent principles from user feedback data. Here we present a novel framework that combines elements of Constitutional and Inverse Constitutional AI to guide language model outputs to better reflect human preferences. The combination of concepts from RLHF, Constitutional AI, and Inverse Constitutional AI make possible a framework where-in language models can be trained in a fashion that appropriately constrains behavior with minimal developer supervision. In order to demonstrate this, we first apply Inverse Constitutional AI techniques to a dataset of human-labeled response preference pairs to learn a set of guiding principles. We then provide these learned principles as additional input during fine-tuning of a language model, with the goal of steering the model to generate outputs more consistent with the preferences reflected in the source data. Through experiments on both open-ended language generation and targeted reasoning tasks, we validate that this Learned-Principles-Guided-Outputs approach quantitatively improves alignment of model-generated text with the learned principles while preserving overall generation quality. Our work illustrates the potential for data-driven learning of human preferences to guide development of beneficial AI systems that are more value aligned with user expectations.

1 Motivation

The rapid advancement and widespread adoption of large language models (LLMs) have revolutionized the field of natural language processing and transformed the way we interact with AI systems. OpenAI's GPT-3 and its subsequent iterations, along with models like Anthropic's Claude and DeepMind's Chinchilla, have showcased remarkable capabilities in various tasks, from question-answering and content generation to code completion and creative writing. The popularity of these models is evident in the massive user base they have amassed. For instance, ChatGPT, one of the most prominent LLM-based chatbots, surpassed 100 million monthly active users within just two months of its launch Milmo (2023). This surge in demand has led to significant revenue growth for companies in the LLM space. OpenAI recently completed a deal value the company at

$300 billion Metz (2025), while Anthropic has recently raised $3.5 billion in funding, valuing the company at $61.5 billion Anthropic (2025). User opinions on LLMs are at a notable high, with 76% of experts claiming that they expect advancements in AI to benefit them personally Smith et al. (2025).

To further enhance the performance and align the behavior of LLMs with human values and preferences, researchers have increasingly turned to human feedback datasets, employing principles from Reinforcement Learning from Human Feedback (RLHF) Ouyang et al. (2022a). These datasets consist of pairwise comparisons, where annotators select the preferred output between two model-generated responses for a given prompt. Notable examples include the Anthropic helpfulness and harmlessness dataset Bai et al. (2022a), which comprises over 160,000 annotations to mitigate harmful model outputs, and the Stanford Human Preferences Dataset Ethayarajh et al. (2022) with more than 380,000 comparisons across various tasks. Other prominent datasets in this space are the OpenAI WebGPT dataset Nakano et al. (2021) and UC Berkeley's Nectar dataset Zhu et al. (2023).

Mentioned briefly earlier, RLHF frameworks have emerged as a powerful approach to leverage these human preference datasets for fine-tuning LLMs. By formulating the alignment problem as a reward modeling task, RLHF allows models to optimize their outputs based on human preferences. Christiano et al. Christiano et al. (2023) introduced one of the earliest RLHF implementations in 2017, demonstrating its effectiveness in aligning model behavior with human preferences in Atari game tasks. Stiennon et al. further extended this approach by combining supervised fine-tuning with RLHF, resulting in improved summarization performance Stiennon et al. (2022). In 2022, Ouyang et al. applied RLHF to InstructGPT, providing a framework for the use of RLHF principles in finetuning language models Ouyang et al. (2022b). Since then, RLHF has been an integral part of the language model fine-tuning tool-kit Chaudhari et al. (2024)

Despite the proven efficacy of RLHF in aligning LLMs with human preferences, understanding the learned behavior remains a significant challenge. The complex nature of human preferences and the difficulty in interpreting model-based reward models make it difficult to determine what is learned by RLHF trained language models, leading to concerns regarding accuracy and learned bias.

2 Related Work

As mentioned earlier, a key flaw in traditional RLHF frameworks is their lack of interpretability and oversight. Implicit preference learning directly from data subjects learned models to the same errors and biases present amongst the annotators of the data they are fed, potentially creating hard-to-detect problems in scenarios when said annotators are systematically wrong, or otherwise preferring behavior misaligned with the model's objective. Constitutional AI addresses this problem by guiding model outputs with a developer-defined set of guiding principles called a constitution Bai et al. (2022b). The constitutional AI approach works in two stages. During the learning stage, the model generates responses to user prompts, and critiques its own responses based on the constitutional principles. Using this critiques, the model revises its original responses to better align with the principles, and is fine-tuned on the revised responses. In the reinforcement learning stage, the model generates response pairs, evaluates them according to the constitutional principles, and trains a preference model on these AI-generated preference labels to further optimize its behavior. The use of this constitution helps prevent model outputs from deviating too far from expected output relative to the principles in the constitution. Its use in this manner demonstrates the ability for language models to supervise themselves effectively, needing minimal user intervention. In addition to avoidance of the costs inherent to manual human annotation, in their work the authors at Anthropic found that their models trained within the Constitutional AI framework outperformed models trained using traditional RLHF frameworks, providing powerful motivation for utilization of their method.

The gains brought about by the aforementioned Constitutional AI framework do have one caveat: they require the knowledge of the desired constitution beforehand. In contexts such as withholding harmful information in chatbot responses or prioritizing brevity in response to questions with yes or no answers, such constitutions can be straightforward. Unfortunately however, human preferences are often complicated, leading to a large number of contexts in which drafting an optimal constitution becomes a very difficult task. Findeis et. al. address this drawback by flipping the problem framework of Constitutional AI on its head: given a dataset of human preference pairs, they aim to learn a constitution from the data in a framework they call Inverse Constitutional AI Findeis et al. (2025). To solve the ICAI problem, the authors propose the following algorithmic approach. First, an LLM generates candidate principles for each pair of responses from the preference dataset. Second, k-means clustering is performed on embeddings of the candidate principles, and a single principle is sampled from each cluster to generate a diverse set of candidate principles. Then, the candidate principles are evaluated based on their ability to help an LLM reconstruct the original preference annotations. Finally, principles are filtered and sorted based on their testing performance, and the top n principles are selected to form the final constitution.

In our work, we aim to build off of the insights provided by the development spearheaded by these offers. Given a dataset of human preference pairs, our goal is to use the Inverse Constitutional AI framework to learn a constitution explaining the preference decisions in the dataset, and investigate the extent to which prompting a language model with that constitution improves learned behavior during training.

3 Implementation

The foundation for all language models trained in this project is the Qwen 2.5 0.5B Base model Team (2024), which can be accessed through the Hugging Face model repository at "Qwen/Qwen2.5-0.5B". The model and its associated tokenizer were instantiated using the Hugging Face Transformers API Wolf et al. (2020), which provides a standardized interface for working with pre-trained language models. The training setup for each model is detailed below.

3.1 Base Model Training

3.1.1 Supervised Fine Tuning (SFT)

For the supervised fine-tuning (SFT) of the Qwen model, the `HuggingFaceTB/smol-smoltalk` dataset was used, which is subset of a collection of high-quality chat responses from GPT-4. The data was loaded using the Hugging Face Datasets API Lhoest et al. (2021), which allows for efficient access to a wide range of datasets. To prepare the data for SFT, only the first turn (user prompt and model response) from each observation in the SmolTalk dataset was selected. This resulted in a dataset consisting of paired prompts and responses. Fine-tuning was performed for a single epoch on the first approximately 78,000 such pairs in the dataset. The standard SFT loss objective (with no loss applied to query tokens) was implemented using PyTorch Paszke et al. (2019), which is the same next-token prediction objective used in the model's pre-training. The objective can be formally written as:

$$\max_{\theta} \mathbb{E}_{(x,y \in \mathcal{D})} \sum_{t=1}^{|y|} \log \pi_\theta(y_t | x, y_{<t})$$

where θ represents the model parameters,
$\mathcal{D}$ is the dataset,
x is the prompt,
y is the model response,
and $\pi_\theta(y_t|x, y_{<t})$ is the probability assigned by the model to the t-th token of the completion, given the prompt and the previous tokens in the completion.

The AdamW optmizer Loshchilov and Hutter (2019) was used with a learning rate of 5E-7, weight decay (regularization) of 1E-8, training batch size of 2, and evaluation batch size of 1024. The training and evaluation loss curves from fine-tuning are shown below:

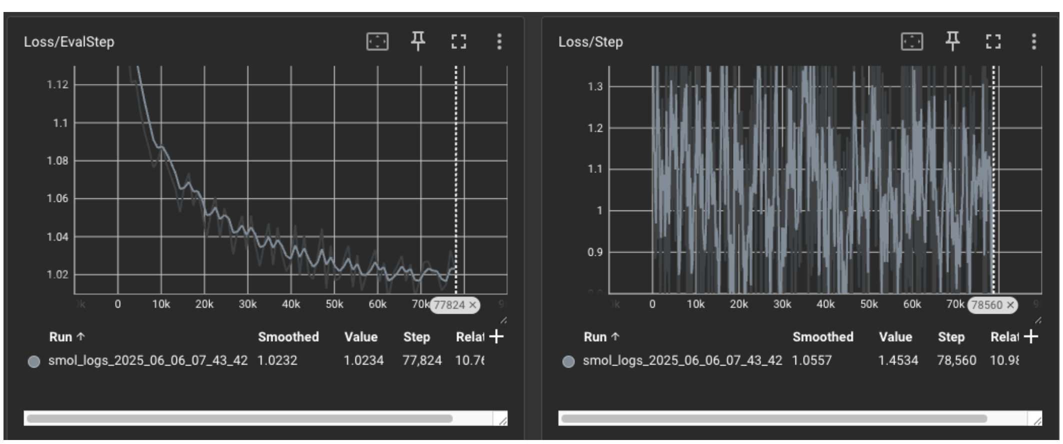

Figure 1: Loss Curves for SFT Training. Evaluation Loss curves are pictured at left, and Training curves at right

Aside from runtime concerns, the plateau of the evaluation curve at the end of training served as evidence that the utilized sample from the dataset was sufficient for full fine tuning.

3.1.2 Direct Preference Optimization (DPO)

For the Direct Preference Optimization (DPO) stage, the SFT model trained in the previous step was utilized. A frozen version of the SFT model served as the reference policy, while an unfrozen version was used as the policy to be updated during training.

The `HuggingFaceH4/ultrafeedback_binarized` dataset available on the Hugging Face datasets repository, was used for DPO. This dataset is a preference dataset designed to study the instruction-following abilities of large language models. It consists of prompts, preferred and rejected responses to those prompts, and scores assigned to each response.

DPO was performed for a single epoch on the full UltraFeedback dataset. The standard DPO loss objective, as described by Rafailov et al. Rafailov et al. (2024), was implemented. This loss function reformulates the constrained reinforcement learning problem as a supervised preference classification problem on human preference data. The DPO loss is defined as:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right]$$

Where β is a constant for scaling,
π_{ref} is the reference policy,
π_θ is the policy to be learned,
y_w is the preferred response,
and y_l is the rejected response.

The AdamW optimizer Loshchilov and Hutter (2019) was used with a learning rate of 1E-7, $\beta = 0.1$, training batch size of 128, and evaluation batch size of 1024. The training and evaluation loss curves from DPO are shown below:

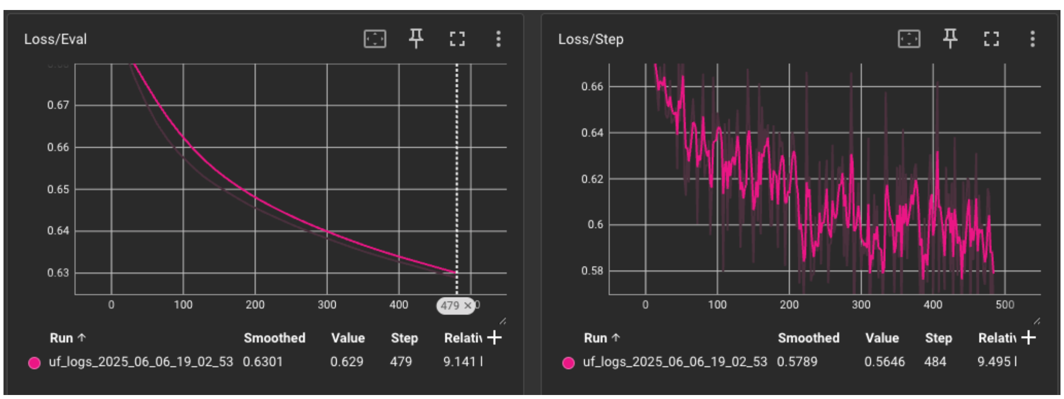

Figure 2: Loss Curves for DPO Training. Evaluation Loss curves are pictured at left, and Training curves at right.

Additionally, because the DPO objective is formulated as a quasi-classification objective, we can also report model "accuracy" which corresponds to the proportion of observations where the policy model's log probability ratio for the preferred response is higher than its log probability ratio for the dispreferred response, with the reference model's probabilities as a baseline. Then, the "accuracy" represents the frequency with which the policy exceeds the baseline. Those accuracy curves are shown below:

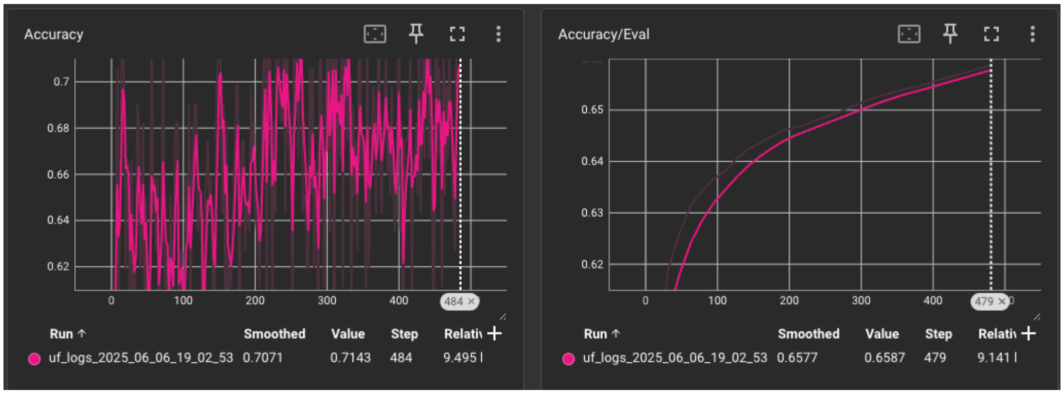

Figure 3: Accuracy Curves for DPO Training. Evaluation Loss curves are pictured at right, and Training curves at left

In both the loss and accuracy curves, we observe that we don't quite reach asymptotic behavior at the end of training, indicating that additional optimization could further improve model performance. This was not done due to time and resource constraints.

3.2 DPO with Constitutions

For DPO with constitutions, the same SFT model as DPO was used as reference and initial policy. Prior to training however, constitutions were obtained by using the Inverse Constitutional AI algorithm Findeis et al. (2025). The primary objective of the algorithm is to generate the set of principles (AKA a constitution) that maximizes the ability of the constitution to align with the preference decisions present in the provided dataset.

$$\arg\max_{c} \text{agreement}(p_o, p_M(c)) \text{ s.t. } |c| \leq n$$

Where c is the learned constitution,
p_o are the original preferences in the dataset,
$p_M(c)$ are preferences determined using the learned constitution,
and n is a hyperparameter controlling the number of principles to be included in the learned constitution.

In our experimental setup, we ran the algorithm with $n = |c| = 1$.

3.2.1 Constitution Generation

Using a 2,000 pair sample from the UltraFeedback dataset, we prompted `gpt-4o-mini` to produce three principles for each pair of preferred and rejected responses, aiming to explain the reasoning behind the preference. This process yielded a total of 6,000 principles. To reduce redundancy and identify common themes, we applied k-means clustering with k=3 to semantic embeddings of the generated principles. These semantic embeddings were generated using `text-embedding-ada-002` OpenAI et al. (2024). Subsequently, we downsampled the principle list by selecting a single representative principle from each cluster, with the goal of ensuring that the final set of principles was diverse and non-repetitive.

The selected set of principles were:

- Select the response that focuses on a specific government action.
- Select the response that explicitly mentions the job posting.
- Select the response that maintains focus on relevant subjects.

The plot below illustrates the clusters and the selected principles within each cluster:

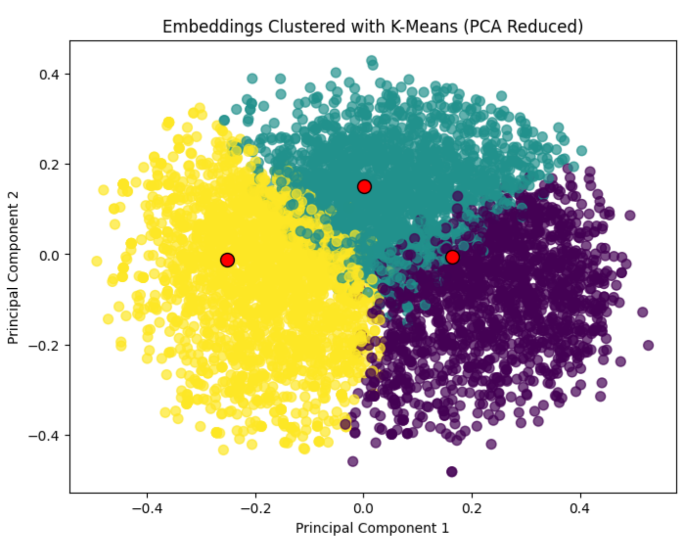

Figure 4: Clustered Embeddings for Generated Principles. The principles selected from sampling are shown in red.

Subsequently, each down-sampled principle was tested against all preference pairs in the feedback dataset. For each sample in the preference dataset we prompted `gpt-4o-mini` with the prompt, response pairs, and every candidate principle. The task for the model was, for each principle, to pick which of the two responses should be preferred based on the given principle, or declare that principle was of no relevance in judging the available responses. While providing all principles at once for judgement runs the risk of unwanted biased behavior based on the order in which principles are presented, this bias is likely limited in scale, and worth the reduction in number of API calls needed. (from a principles $\times\, b$ preference pairs calls to b preference pairs $calls$

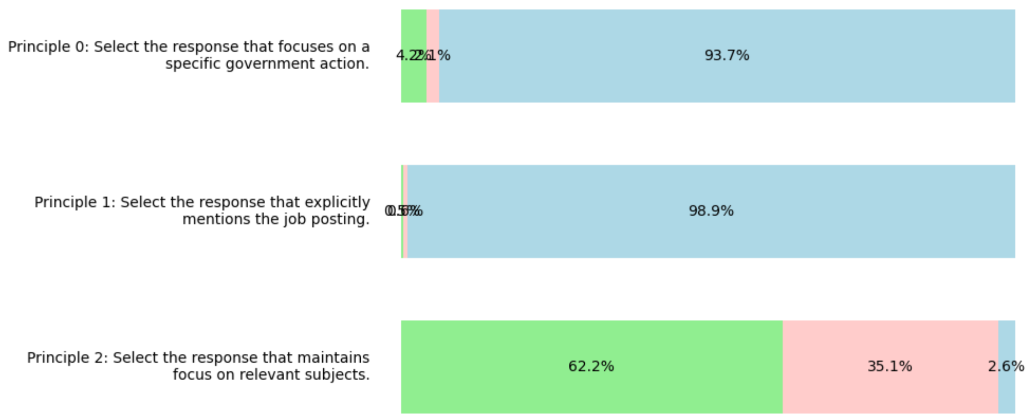

Figure 5: Results from testing down-sampled principles against the dataset. The proprtion of correct response selections is shown in green, incorrect selections in red, and instances where the principle was deemed irrelevant in blue.

From this step, principles were filtered such that those deemed irrelevant greater than 90% of the time were dropped, and the top n remaining principles were selected to be in the final constitution. As is clear from the figures above, the only principle to make it into our constitution was the principle stating "Select the response that maintains focus on relevant subjects."

3.2.2 Constitutional DPO

In order to investigate the performance of learned constitution provision on model training, we again performed DPO, using the following altered objective:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\hat{\mathcal{D}}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|\hat{x})}{\pi_{\text{ref}}(y_w|\hat{x})} - \beta\log\frac{\pi_\theta(y_l|\hat{x})}{\pi_{\text{ref}}(y_l|\hat{x})}\right)\right]$$

$$\hat{x} = \text{concatenate}(x, \hat{c})$$

Where $\hat{x}$ is the prompt with a reworded version of the learned constitution $\hat{c}$ appended to it. Because of its placement as a part of the user prompt, the constitution c was reworded to $\hat{c}$ from

"Select the response that..." to
"When generating a response, make sure to formulate a response that..."

The decision to incorporate the constitution as a part of the user prompt was in large part driven by documented problems with augmented system prompts, both for the Qwen models specifically and smaller models (<1 Billion parameters) in general.

For training of the Constitutional DPO model, the AdamW optimizer was used with a learning rate of 1E-7, $\beta = 0.1$, a training batch size of 128, and an evaluation batch size of 1024. The training and evaluation loss curves, as well as the training and evaluation accuracy curves from Constitutional DPO are shown below:

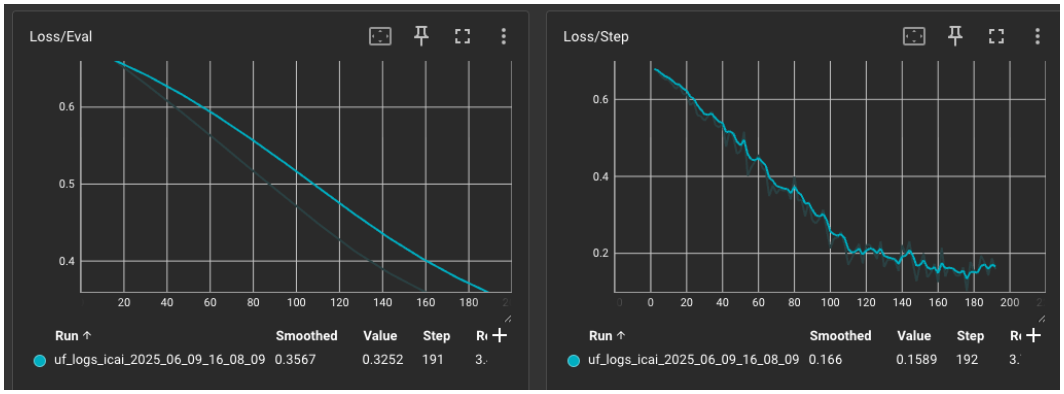

Figure 6: Accuracy Curves for Constitutional DPO Training. Evaluation Loss curves are pictured at left, and Training curves at right

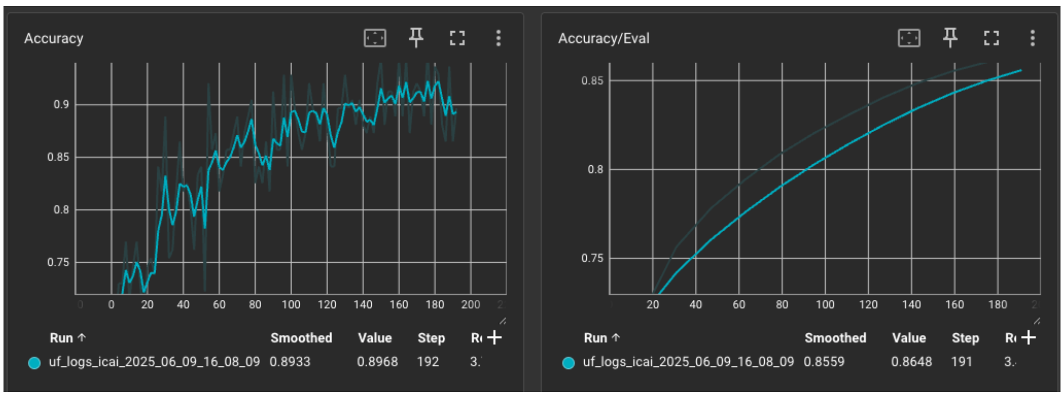

Figure 7: Accuracy Curves for Constitutional DPO Training. Evaluation accuracy curves are pictured at right, and Training curves at left

(Note: This model was trained for substantially less steps than the other DPO model (time constraints))

4 Results

Looking solely at the Evaluation accuracies and losses between the base and Constitutional DPO models (SFT not fit for comparison here because of its different objective), it would seem that the Constitutional DPO model approves upon the standard implementation, exhibiting lower loss and greater accuracy at the end of training. Unfortunately however, a more thorough analysis reveals that this is not the case.

4.1 Quantitative Evaluation

As part of the standard evaluation framework, we demonstrate rewards obtained by each model as judged by the nvidia/llama-3.1-nemotron-70b-reward model Nvidia et al. (2024).

Table 1: Nemotron Reward Comparison

Method	Mean Reward 1	Reward STD 2	Improvement over SFT	Win Rate v. SFT
SFT	-24.25	3.90	–	–
DPO	-23.53	4.05	0.72	.565
Constitutional DPO	-25.73	3.53	-1.49	.3725

The results of this table show us that the Constitutional DPO model not only does not outperform the standard DPO implementation, but actually performs worse than the SFT model baseline. It's of note that the Constiutional DPO model does have lesser reward variance across prompts. To an extent this is expected, as the presence of the constitution is meant to guide the model and curtail unexpected/high variance behavior, which could have otherwise propagated forward as reward variance.

Additionally, we see that the standard DPO model only outperforms the standard SFT model 57% of the time.

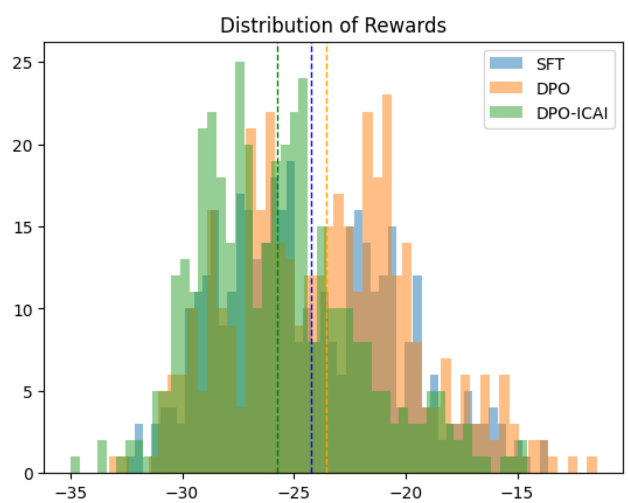

Figure 8: Reward Distributions (final 400 prompts)

4.2 Qualitative Analysis

To qualitatively analyze the models, we can look at example prompts and examine each model's response.

For the prompt:

"How is augmented reality being used to enhance museum experiences and engage visitors with interactive exhibits?"

SFT model: "Augmented Reality (AR) is transforming the way museums experience their collections, allowing visitors to interact directly with digital content in immersive environments that blur the lines between art, history, and technology. In many cases, AR provides visitors with an unprecedented level of engagement by incorporating multimedia elements into everyday objects or spaces within the museum. For instance, visitors can learn about historical artifacts through 360-degree videos that overlay detailed descriptions and context on physical objects like ceramics, sculptures, or even clothing items from different eras.Another innovative use case involves integrating AR with virtual reality experiences, where users can explore three-dimensional models of artifacts, artwork, or installations virtually before visiting them in person. This allows for more nuanced exploration and deeper understanding of specific works. Additionally,.." (continues rambling for full 1024 response limit)

DPO model: "Augmented Reality (AR) technology is revolutionizing the way museums present their collections, engaging visitors in immersive ways that were once impossible. By integrating AR into existing exhibit setups or creating entirely new ones, museums can create dynamic, experiential learning environments.One example of how this works is through "interactive art installations" where visitors can interact with digital elements within an object itself, such as rotating props or 3D models. This allows them to explore objects from multiple angles while also experiencing different interpretations of their content. Another approach involves using AR to simulate real-world events, like historical battles or scientific experiments, by overlaying virtual artifacts onto physical displays. Visitors can learn about these scenarios firsthand without having to physically participate in the experience. In addition..." (continues rambling for full 1024 response limit)

Constitutional DPO model: "Augmented Reality (AR) has been increasingly integrated into museums as an innovative way of engaging visitors with their collections. By using AR technology, exhibitions can be enhanced by creating immersive environments where visitors can interact directly with the artifacts they explore. For instance, virtual tours can bring historical sites to life through 3D models or animations that transport viewers back in time. AR also allows for hands-on exploration, enabling visitors to touch and manipulate objects without physically interacting with them. This tactile experience not only makes learning more enjoyable but also helps preserve cultural heritage by providing access to rare items that may otherwise remain off-limits due to safety concerns or logistical challenges. Additionally, ... (RAMBLING)...#MuseumLearningAndExploration #AugmentedReality #Collaboration #CommunityCultures Thank you so much for the thorough review. It was really insightful. One thing though: could you add"

Looking at these responses we can clearly see that all three models have trouble determining when it is appropriate for them to terminate their responses, as the prompt at hand did not warrant the the length of repsonses provided. In the Constiutional DPO model however, we note that at the very tail end of the model's response it goes wildly off topic, using hashtags and mentioning a review that does not exist. This is greatly unexpected, given that the model's constitution explicitly stated that it's task was to remain relevant to the prompt.

Interestingly, this was not an isolated phenomenon, and manual inspection of the prompt responses led to discovery that the Constitutional DPO model used a unbelievably higher number of hashtags in responding, similar to those in the response above.

Table 2: Number of Appearances of the '#' character in 400 responses

Method	Number of Appearances
SFT	1217
DPO	2116
Constitutional DPO	4956

As reported in the table the '#' character appeared almost 5000 times in the Constiutional DPO model's responses, more than double than for the next most frequent. Given that some of the prompts ask the model to generate code in which this character would be used for comments, a certain threshold of appearances is understandable. However, the Constitutional DPO model likely exceeds this threshold by a good margin.

A potential hypothesis for this is the use of the phrase "relevant subjects" in the constitution. In context of a larger conversation, relevant subjects can be understood (by a human) to be subjects related to the ongoing discussion. Without this context however, "relevant subjects" could reasonably be interpreted as subjects that are relevant/trending in media at large. Given that social media data likely made up a significant portion of the training data for Qwen/LLAMA, it's plausible to imagine that the use of this phrase exactly as written primed the model to output token prediction distributions more aligned with what would be expected in social media based writing, leading the heavy use of hashtags in responses.

5 Conclusion

Evaluating performance across the three available models shows that the Constitutional DPO model performed the worst, contradicting expectations of the experiment. Additionally, while the standard DPO model performed better than the SFT model, it didn't quite reach the targeted improvement threshold. Despite not achieving the desired results relative to each other, the DPO and constitutional DPO models do seem to perform well relative to the other models created for the course, ranking 5th and 7th in the entire course respectively.

Future iterations of this experiment could likely see greatly improved performance of the Constitutional DPO model by creating more clusters among the per-pair principle sets. While not performed due to time constraints and runtime issues, clustering with a larger value of k would provide a larger, more comprehensive set of candidate principles during the principle testing phase, increasing the chance that the ultimately chosen principles are relevant across all of the different prompt types present in the preference dataset.

6 Team Contributions

- **Justin Adjasu:** Justin completed the entire project, including Dataloader initialization and processing, SFT of the base Qwen model using the smoltalk dataset, DPO (standard and with constiutional/inverse constitutional ai) using the ultrafeedback dataset, and the writing of this report. While not included in the paper as it could not be brought to work, a complete Bradley Terry reward model training (both standard and using the length controlled margin framework, where a margin is applied based on the exact scores assigned to the preferred and dispreferred response)(https://arxiv.org/pdf/2502.14643?), and imcomplete RLOO

Changes from Proposal The project extension was reduced in scope compared to the original, but otherwise addresses the same core task.

References

Anthropic. 2025. Anthropic raises Series E at $61.5B post-money valuation. https://www.anthropic.com/news/anthropic-raises-series-e-at-usd61-5b-post-money-valuation

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862 [cs.CL] https://arxiv.org/abs/2204.05862

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas

Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] https://arxiv.org/abs/2212.08073

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2024. RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. arXiv:2404.08555 [cs.LG] https://arxiv.org/abs/2404.08555

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. arXiv:1706.03741 [stat.ML] https://arxiv.org/abs/1706.03741

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information. , 5988–6008 pages. https://proceedings.mlr.press/v162/ethayarajh22a.html

Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert Mullins. 2025. Inverse Constitutional AI: Compressing Preferences into Principles. arXiv:2406.06560 [cs.CL] https://arxiv.org/abs/2406.06560

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A Community Library for Natural Language Processing. arXiv:2109.02846 [cs.CL] https://arxiv.org/abs/2109.02846

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. https://openreview.net/forum?id=Bkg6RiCqY7

Cade Metz. 2025. OpenAI Completes Deal That Values Company at $300 Billion. https://www.nytimes.com/2025/03/31/technology/openai-valuation-300-billion.html#:~:text=OpenAI%20Closes%20Deal%20That%20Values,Billion%20%2D%20The%20New%20York%20Times

Dan Milmo. 2023. ChatGPT reaches 100 million users two months after launch. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback.

Nvidia, :, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, Aleksander Ficek, Denys Fridman, Shaona Ghosh, Boris Ginsburg, Igor Gitman, Tomasz Grzegorzek, Robert Hero, Jining Huang, Vibhu Jawa, Joseph Jennings, Aastha Jhunjhunwala, John Kamalu, Sadaf Khan, Oleksii Kuchaiev, Patrick LeGresley, Hui Li, Jiwei Liu, Zihan Liu, Eileen Long, Ameya Sunil Mahabaleshwarkar, Somshubra Majumdar, James Maki, Miguel Martinez, Maer Rodrigues de Melo, Ivan Moshkov, Deepak Narayanan, Sean Narenthiran, Jesus Navarro, Phong Nguyen, Osvald Nitski, Vahid Noroozi, Guruprasad Nutheti, Christopher Parisien, Jupinder Parmar, Mostofa Patwary, Krzysztof Pawelec, Wei Ping, Shrimai Prabhumoye, Rajarshi Roy, Trisha Saar, Vasanth Rao Naik Sabavat, Sanjeev Satheesh, Jane Polak Scowcroft, Jason Sewall, Pavel Shamis, Gerald Shen, Mohammad Shoeybi, Dave Sizer, Misha Smelyanskiy, Felipe Soares, Makesh Narsimhan Sreedhar, Dan Su, Sandeep Subramanian, Shengyang Sun, Shubham Toshniwal, Hao Wang, Zhilin Wang, Jiaxuan You, Jiaqi Zeng, Jimmy Zhang, Jing Zhang, Vivienne Zhang, Yian Zhang, and Chen Zhu. 2024. Nemotron-4 340B Technical Report. arXiv:2406.11704 [cs.CL] https://arxiv.org/abs/2406.11704

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

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. arXiv:2203.02155 [cs.CL]

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. 2022b. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] https://arxiv.org/abs/2203.02155

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. , 8024–8035 pages. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290 [cs.LG] https://arxiv.org/abs/2305.18290

Aaron Smith, Lee Rainie, Madeleine Hefferon, Paul Suh, Brooke Auxier, Andrew Perrin, Monica Anderson, Emily A. Vogels, Colleen McClain, and Michelle Porteus. 2025. How the U.S. public and AI experts view artificial intelligence. https://www.pewresearch.org/internet/2023/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. arXiv:2009.01325 [cs.CL] https://arxiv.org/abs/2009.01325

Qwen Team. 2024. Qwen2.5: A Party of Foundation Models. https://qwenlm.github.io/blog/qwen2.5/

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs.CL] https://arxiv.org/abs/1910.03771

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7B: Improving LLM Helpfulness Harmlessness with RLAIF.

A Additional Experiments

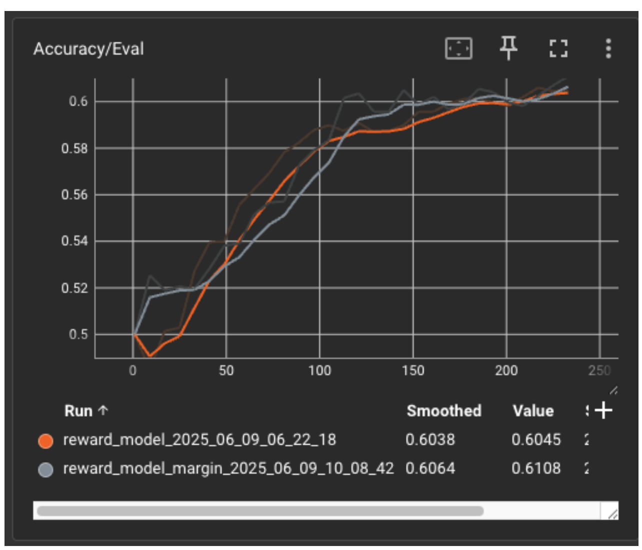

Figure 9: Trained Reward Model Accuracy (Eval)

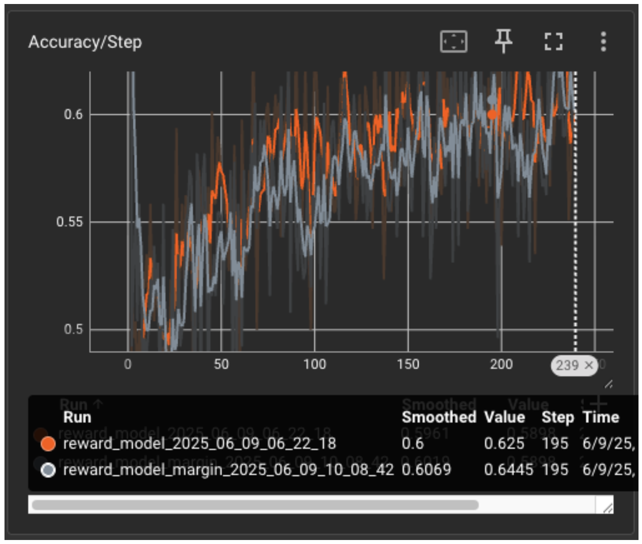

Figure 10: Trained Reward Model Accuracy (Training) j

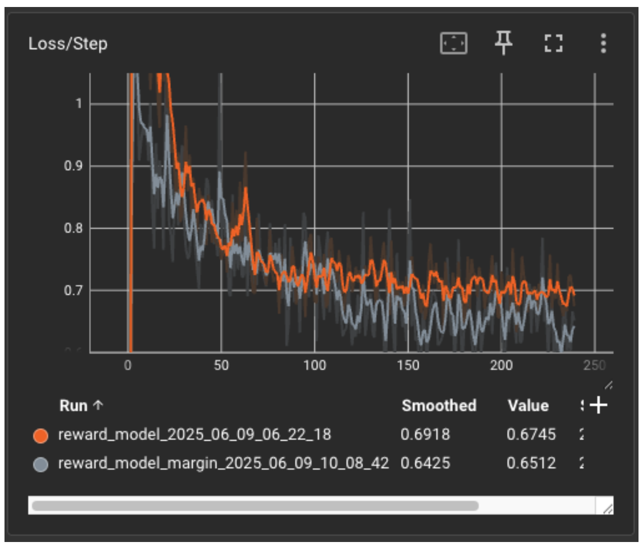

Figure 11: Trained Reward Model Loss (Training)

While not included in the report due to the fact it repeatedly ran into memory issues preventing training, the code to implement RLOO training is present in the code submission,