Extended Abstract

Motivation Large language models (LLMs) have become widely used, yet their instruction-following ability is still under question. Despite the exponentially larger corpora of training data used, and cutting-edge reinforcement learning (RL) techniques to align LLMs with human preference, many still output irrelevant answers or hallucinating. Many scholars have pointed out the need for better and more targeted post-training using a wider range of RL methods or clearer and more diverse reward signals. But how to improve the instruction-following capabilities of LLMs is still an open research question.

Method We attempted to improve upon a 0.5-billion-parameter Qwen 2.5 model's performance by using multi-objective Direct Preference Optimization (DPO), focusing on improving the instruction-following, truthfulness, coherence, and verbosity of the model's answers. This process not only improves the quality of model responses but also reduces training time, as a much smaller dataset is needed to produce higher-quality responses. First, we trained and evaluated the Qwen model using Supervised Fine-tuning (SFT) and DPO. Then, we generated 3 separate datasets using mixed-reward signals each from Ultrafeedback and HelpSteer datasets, that respectively optimized for the instruction-following and truthfulness, and the coherence and verbosity of model responses.

Implementation We completed one epoch of SFT using the SmolTalk dataset, which consists of more than 1.1 million prompt-answer pairs taken from high-quality chat responses from GPT-4o, optimizing for next-token prediction. We completed one epoch of DPO training using the Binarized UltraFeedback dataset, which consists of more than 61,100 chosen-reject answer pairs to user prompts, optimizing for the likelihood of producing the "chosen" responses over the "reject" responses. Then, we crafted new datasets using UltraFeedback and HelpSteer, selecting chosen-response pairs based on relevant reward signals. For UltraFeedback, we selected the pairs based on the balance of instruction-following and truthfulness scores, and for HelpSteer, we selected them based on the balance of coherence and verbosity scores. For each dataset, we created three versions of subsets using different balances of the two reward signals.

Results We found that while traditional post-training techniques such as SFT and DPO are able to significantly improve base model performance, multi-objective DPO using a dataset of mixed reward signals can offer a greater boost in the helpfulness and quality of responses, as scored by a Llama 3.1 Nemotron 70B Reward Model. This boost is as much as a 25% higher win rate of model responses against the SFT model. Further, the multi-objective datasets are more lightweight and required much less training time.

Discussion We observe small differences in the effectiveness of different balances of reward signals, specifically the coherence-verbosity split. While our dataset may not be sufficiently diverse, we hypothesize that a dataset with higher emphasis on verbosity scores is able to achieve higher performance because verbosity is a common pitfall of LLMs. Ultimately, we were able to see the DPO training on a 10,000-datapoint multi-objective UltraFeedback subset comparable to another full epoch of training on the more than 61,000-datapoint Binarized UltraFeedback.

Conclusion Traditional RL methods like SFT and DPO are strong baselines, while multi-objective DPO further improves performance, showing the value of mixing high-quality, diverse reward signals. Further experimentation is needed on the effect of different tradeoffs and on aligning models to human preferences.

Improving LLM Instruction-Following Capabilities with Multi-objective Reinforcement Learning

An Doan
Department of Computer Science
Stanford University
andoan@stanford.edu

Felicity Huang
Department of Computer Science
Stanford University
huangfe@stanford.edu

Linda Liu
Department of Computer Science
Stanford University
ylinliu@stanford.edu

Abstract

Large language models (LLMs) have become widely used, yet their instruction-following ability is still under question. Despite the large amounts of training data and traditional reinforcement learning (RL) techniques to align LLMs with human preference, many still output irrelevant answers or unfactual ones, a behavior known as hallucination. We sought to investigate how multi-objective reinforcement learning, which optimizes for not one but multiple reward signals, can improve the instruction-following capabilities of LLMs. We trained a Qwen 2.5 0.5B base model using Direct Preference Optimization (DPO) and found that using a dataset that optimizes for two reward signals — such as instruction-following, truthfulness, coherence, and verbosity scores — can significantly boost the model's performance compared to traditional methods such as Supervised Fine-tuning (SFT) and DPO. Future studies could focus on evaluating whether multi-objective RL can reduce LLM hallucination, a promising research area.

1 Introduction

The development of large language models (LLMs) has fundamentally transformed society. While LLM agents such as ChatGPT and Claude see millions of users worldwide, they are still prone to "hallucinations," or outputting irrelevant and sometimes made-up facts in response to user prompts. This significantly erodes these models' usability and credibility. Much of recent language model developments focus on scaling model parameters and pretraining data, but not enough attention is given to enhancing their instruction-following capabilities.

We sought to investigate whether we can use reinforcement learning techniques to improve the instruction-following ability of a Qwen 2.5 base model with 0.5 billion parameters. To do so, we employed two classic methods, Supervised Fine-tuning (SFT) and Direct Preference Optimization (DPO). Then, to further improve the model's behavior, we employed a multi-objective DPO training with reward signals that balanced the coherence and verbosity, as well as instruction-following ability and truthfulness, of the model responses. We hypothesize that using data with clearer reward signals can lead to improved models' responses that clearly encapsulate these signals.

Through our experiments, we found that while traditional and widely embraced RL techniques like SFT and DPO can improve upon a pretrained LLM, more targeted instruction-following training using a carefully curated multi-objective reward signal can offer more significant boosts in performance.

Training on a smaller multi-objective dataset is able to give our model results comparable to another full epoch of training. Future research should focus on examining how reward signals such as instruction-following and truthfulness can not only lead to more helpful model responses, but also reduce hallucination and enhance model usability, with the goal of producing ethical AI agents and agents that don't lie.

2 Related Work

In 2019, Radford et al., a group of OpenAI researchers, unveiled the GPT-2 model. The largest GPT-2 model, a 48-layer decoder-only Transformer model with more than 1.5B parameters trained on 40GB of text data, can achieve state-of-the-art results on many tasks in a zero-shot fashion.Radford et al. (2019) Ever since GPT-2 demonstrated the potential of unsupervised multitask learning, researchers worked on scaling the parameters and pretraining data size to produce even better Transformer-based models. OpenAI, for instance, produced the 175B-parameter GPT-3 trained on 600GB of text data, exhibiting few-shot learning capabilities that rivaled state-of-the-art results on some tasks; the 1.8T-parameter GPT-4 trained on 1PB of text, reaching professional-level performance on several benchmarks; and more.Brown et al. (2020)OpenAI et al. (2024) Each new generation of models outperformed the past generation.

A fundamental part of LLM training is the pretraining-finetuning paradigm. After models are pretrained on large corpora of text, they often undergo finetuning for specific tasks and further reinforcement learning (RL) training to align their behavior with human preferences. Reinforcement Learning from Human Feedback (RLHF) is an effective line of methods that researchers have used to improve the truthfulness and reduce toxicity of GPT-3 outputs.Ouyang et al. (2022) It includes first fitting a reward model based on human preferences and then fine-tuning the LLM using reinforcement learning to maximize this estimated reward.Li et al. (2023) A more light-weight and intuitive method is Direct Preference Optimization (DPO), which simply works with chosen-reject pairs of responses to a given prompt and maximizes the LLM's likelihood of outputting the chosen answer, using another LLM as a reference.Rafailov et al. (2024)

While fine-tuning and traditional RL techniques have demonstrated success in AI alignment, recent literature highlighted their inefficiency and limitations. Scholars have noted that instruction-following is still a significant limitation of generalist LLMs, despite the pretraining-finetuning paradigm. This is especially seen in LLMs' tendency to hallucinate, or produce unfactual answers that look plausible.Huang et al. (2025) Xu et al. (2025) argued that hallucination — more formally, inconsistencies between a computable LLM and a computable ground truth function — is unavoidable when LLMs are used as general problem-solvers. This means that many LLMs output responses that not only deviate from users' intent but also are misleading.

Scholars have tried to explain why traditional RL techniques fail to produce AI models that represent diverse human values and perspectives. Recent work by Sorensen et al. (2024) and Singh et al. (2025) emphasizes limitations of optimizing LLMs towards a single notion of correctness. To alleviate this limitation, Sorensen et al. (2024) proposed using benchmarks that contain more than one objective (e.g., helpfulness vs. harmlessness), trading off benchmarks, or using a panel of simulated or real annotators to assess a model's output from diverse viewpoints.

How to best align LLMs with human preference is still an open research question. Following Sorensen et al., we sought to examine how training the Qwen model with mixed reward signals emphasizing instruction-following and truthfulness, as well as coherence and verbosity, can improve training results and efficiency.

3 Method

3.1 Supervised Fine-tuning (SFT)

3.1.1 Methodology

The first step in our training pipeline is Supervised Fine-tuning (SFT) of the Qwen 2.5 0.5B base model, yielding a reference policy π_{ref} to guide downstream preference optimization. SFT is implemented as a standard next-token prediction task, with loss applied only to the response (completion) tokens rather than the prompt (instruction) tokens, in alignment with established practices for instruction tuning.

The objective function is defined as

$$\max_{\theta} \sum_{t=1}^{T} \log \pi_\theta(y_t \mid x, y_{<t})$$

where x is the prompt and $y = (y_1, \ldots, y_T)$ is the expert response. The objective maximizes the log likelihood of the expert distribution, given the prompt. To ensure proper masking, we construct token-level labels and attention masks such that only tokens corresponding to the response contribute to the loss.

We tokenized data using the HuggingFace tokenizer for Qwen 2.5, applied truncation and padding, and used PyTorch `DataLoader` for efficient batch sampling. Training was conducted over one epoch with the AdamW optimizer, a learning rate of 1×10^{-6}, and batch size of 1 to fit within GPU memory. For the SmolTalk SFT model, we stopped training once loss approximately plateaued due to the size of the dataset, since it was unlikely that the model would learn from all the information and for the sake of time constraints with training. For the WarmStart Countdown SFT model, we ran 15 epochs over the 1K dataset until loss had approximately plateaued, shown in the figure below.

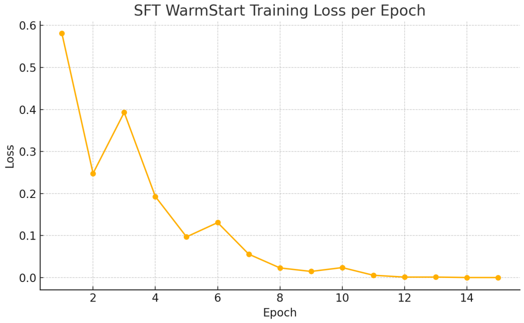


Figure 1: SFT WarmStart training loss

3.1.2 Datasets

SmolTalk: To fine-tune the Qwen model on instruction-following, we used the train split of the SmolTalk dataset, which contains over 1.1 million high-quality instruction–response pairs generated by GPT-4o.Allal et al. (2025) This dataset is particularly suited for initializing instruction-following capabilities in small language models.

WarmStart: For the Countdown reasoning task, we used the WarmStart dataset, which provides prompt–response pairs from a variety of cognitive reasoning strategies including backtracking and verification.Singh (2025) This dataset served as a warm-start for math reasoning models prior to preference optimization using rule-based reward functions. This warm-start phase helped bias the model towards structured, verifiable reasoning patterns essential for downstream reinforcement learning.

3.2 Direct Preference Optimization (DPO)

3.2.1 Methodology

Direct Preference Optimization has been demonstrated to be a state-of-the-art RL method that is known for its lightweight nature and ability to outperform RLHF.Rafailov et al. (2024) The key to DPO is the comparison of pairs of a "winning" and "losing" example, so that the model learns by comparing the two outputs. This leads to a higher probability of producing preferred outputs over disfavored ones, in comparison to a frozen reference policy. DPO optimizes for the objective by using the following loss function, where π_θ is the DPO trained policy and π_{ref} is the frozen reference model:

$$L_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{1}$$

In the above, (x, y_w, y_l) is the prompt, "winning," and "losing" pair; β controls how much to diverge from π_{ref}; and σ defines the sigmoid function.

3.2.2 Dataset: Binarized Ultrafeedback

We used the supervised fine-tuned Qwen 2.5 model as our reference model and trained a base Qwen 2.5-0.5B model on the train split of the UltraFeedback Binarized dataset, which consists of 61,135 tuples featuring a prompt, a chosen response, and a rejected response.Cui et al. (2023) The chosen response to a given prompt was selected by taking the best of four model responses to the prompt based on scores assigned by GPT-4 using criteria like helpfulness and honesty. The rejected response was sampled at random from the remaining three responses. The chosen response serves as a "winning" example, and the rejected response is our "rejected" example.

We hypothesize that DPO will be able to significantly improve the model's instruction-following capability compared to the SFT model, due to the method's effectiveness at aligning model performance with human preferences.

3.3 Multi-objective DPO

3.3.1 Methodology

We trained the DPO-finetuned Qwen model on HelpSteer and UltraFeedback datasets that focus on pairs of reward signals, or objectives, at a time. We hypothesize that the multi-objective training can significantly improve model performance, and that the model would perform even better than if it had gone through a second epoch of training on the full Binarized UltraFeedback dataset.

To clearly examine whether training on a smaller dataset with clearer reward signals can lead to a greater boost in model performance, specifically greater instruction-following capabilities and more truthful responses, we used subsets of the unbinarized UltraFeedback dataset. We created three subsets consisting of 10,000 (prompt, chosen response, rejected response) tuples, where the chosen response has the highest instruction-following and truthfulness scores, and the rejected response has the lowest. As mentioned above, unbinarized UltraFeedback consists of four responses that are scored for helpfulness, honesty, instruction-following, and truthfulness for each prompt.Cui et al. (2023) The dataset was generated using the weighted sum scalarization reward:

$$\boldsymbol{r}_{total} = \boldsymbol{\lambda} \times \boldsymbol{r}_{instruction} + \boldsymbol{(1 - \lambda)} \times \boldsymbol{r}_{truthfulness}$$

We explored 3 instruction-truthfulness ratios, with $\lambda = 0.3, 0.5$ and 0.7. For each, we used the corresponding reward function to generated preference datasets from the original HelpSteer dataset and trained DPO upon the new datasets.

Then, we further explored the tradeoff between coherence and verbosity in multi-objective DPO using another dataset, HelpSteer. HelpSteer contains over 35,000 prompt-response pairs with human-annotated scores for helpfulness, correctness, coherence, complexity, and verbosity. Wang et al. (2023) We used a similar weighted sum scalarization reward.

$$\boldsymbol{r}_{total} = \boldsymbol{\lambda} \times \boldsymbol{r}_{coherence} + \boldsymbol{(1 - \lambda)} \times \boldsymbol{r}_{verbosity}$$

Similarly, we explored 3 coherence-verbosity ratios, with $\lambda = 0.3, 0.5$ and 0.7 and trained DPO on the datasets generated using their corresponding reward functions.

3.4 RLOO

3.4.1 Methodology

We also experimented with how REINFORCE Leave-One-Out (RLOO), an emerging RL method, can improve the Qwen model's mathematical reasoning capabilities. RLOO is a cutting-edge method that uses multiple trajectory samples and unbiased baselines and consistently outperforms RL methods such as Proximal Policy Optimization.Ahmadian et al. (2024) The RLOO objective encourages the model to assign higher probability to responses that yield higher rewards compared to other responses sampled from the same policy:

$$\frac{1}{k}\sum_{i=1}^{k}\left[R(y^{(i)},x)-\frac{1}{k-1}\sum_{j\neq i}R(y^{(j)},x)\right]\nabla_\theta \log \pi_\theta(y^{(i)}|x)$$

For each response $y^{(i)}$, the reward it receives is compared to the average reward of the other $k-1$ responses, creating a relative reward signal. If a response performs better than its peers, the model is nudged to increase the likelihood of generating it in the future. Conversely, if it performs worse, its probability is reduced.

For the Countdown task, we trained SFT on the WarmStart dataset on the Qwen base model, then subsequently RLOO on the finetuned model, using $k = 4$ and batch size of 16 and the rule-based reward function as provided by the project guidelines.

3.4.2 Datasets

To train RLOO on mathematical reasoning, we used the Countdown dataset, which had over 490,000 rows of a target number and a sequence of numbers to use to obtain the target.Pan (2025) We trained our model to respond to the following prompt: "Given the numbers: nums and a target of target, write a step-by-step solution using +, -, *, or /."

4 Experimental Setup

4.1 Training

For each dataset, we tokenized each sample using the Qwen 2.5 0.5B base model tokenizer. We applied truncation and right padding using the HuggingFace transformers utilities, constructed attention masks that excluded prompt tokens from loss, and batched the samples using a PyTorch DataLoader.

For each of our methods, we completed training with $\beta = 0.1$ and a learning rate of 10^{-6}. We used the AdamW optimizer and scaled dot product attention.Loshchilov and Hutter (2019)

4.2 Evaluation

Instruction Following: We evaluated all three methods on a held-out test set of 2,000 prompts of the UltraFeedback dataset. To do so, we generated responses to these prompts using our trained model and a reference model. Then, we used the Llama 3.1 Nemotron 70B Reward Model to score our model's responses and the reference model responses based on quality.Wang et al. (2024) We calculated a "win rate" as the percentage of prompts where our trained model scored higher than the reference model. This served as our benchmark for the instruction-following capability of our trained models.

For the Qwen 2.5 model trained with SFT, we used the Qwen 2.5 0.5B Instruct model as the reference model. To evaluate DPO and multi-objective DPO, we used our SFT model as the reference model.

Math Reasoning: On the verifier-based datasets, we evaluated our trained models by generating responses to a subset of the dataset problems and evaluating the responses using a two-stage reward, format score and verification score, as provided by the rule-based reward function.

Table 1: Experimental methods, reference models and baselines for evaluation

Method	Reference Model	Baseline win rate
Supervised Fine-tuning (SFT)	Qwen 2.5 0.5B Instruct	30%
Direct Preference Optimization (DPO)	Qwen 2.5 0.5B + SFT	60%
Multi-objective DPO	Qwen 2.5 0.5B + SFT	> 60%

4.3 Baseline

As explained in Table 1, we hoped that our Qwen model trained using SFT can achieve a 30% win rate over the Qwen 2.5 0.5B Instruct model, and that our DPO-trained model can achieve a 60% win rate over the SFT-trained model. For our extension, multi-objective DPO, we expected it to significantly outperform DPO and use less training time.

5 Results

5.1 Quantitative Evaluation

5.1.1 Instruction Following

Table 2: Win rates of our various training methods against reference models on the instruction-following task

Method	Data	Win-rate
SFT	SmolTalk	40.85%
DPO (epoch 1)	UltraFeedback	65.55%
DPO (epoch 2)	UltraFeedback	91%
Multi-objective DPO	UltraFeedback, 30/70 instruction/truthfulness split	92%
Multi-objective DPO	UltraFeedback, 50/50 instruction/truthfulness split	92.8%
Multi-objective DPO	UltraFeedback, 70/30 instruction/truthfulness split	92.35%
Multi-objective DPO	HelpSteer, 30/70 coherence/verbosity split	88.55%
Multi-objective DPO	HelpSteer, 50/50 coherence/verbosity split	86.75%
Multi-objective DPO	HelpSteer, 70/30 coherence/verbosity split	81.05%

Table 2 above shows the win rates of our various training methods against reference models. We see that SFT is able to achieve a 40.85% win rate against the Qwen Instruct model, and that DPO was able to improved SFT win-rate by around 25%. This shows us that the traditional pretraining-finetuning paradigm offers models solid baseline performance, and confirms the effectiveness of DPO in aligning language models with human preferences.

However, multi-objective DPO offered a bigger boost, improving it by a further 25% when we used UltraFeedback subsets that emphasized the instruction-following and truthfulness scores and 20% when we used HelpSteer subsets that emphasized the coherence and verbosity scores. Interestingly, despite the multi-objective subsets being smaller datasets (around 10,000 training examples), training DPO on them offered performance levels similar to the level after training a second epoch on the full Binarized UltraFeedback dataset, which has more than 61,000 training examples. This suggests that multi-objective RL can lead to more efficient training.

We can see that all ratios of the subsets offered significant improvement, confirming the value in giving the model clearer reward signals in training data. For instruction/truthfulness, we did not observe a significant performance difference between different λ values, whereas for coherence/verbosity, we observed that a higher emphasis on verbosity compared to coherence offered the greatest performance boost.

5.1.2 Math Reasoning

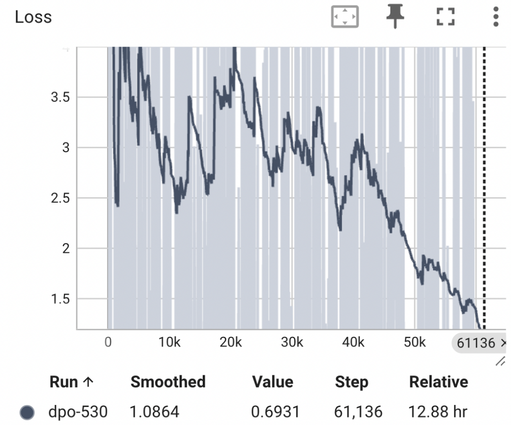
Our model's math reasoning performance is limited and requires further examination.

Table 3: Model performance on the mathematical reasoning task (evaluated using Countdown Task-3-to-4)

Method	Average Score	Correct/Total
SFT-WarmStart	0.184518678	1357/8352
RLOO	0.223214285	50/224

5.2 Qualitative Analysis

We believe that multi-objective DPO produced better results because there are limitations within DPO itself. While multi-objective DPO was able to significant improve model performance, we observed that there might have been inefficiency within the naive DPO training process itself. As shown in Figure 2 below, our training loss was volatile and not monotonically decreasing. This suggests that the model could have benefited from more epochs of training to fully digest the training data, or more targeted training on a subset of the data.



Run ↑	Smoothed	Value	Step	Relative
● dpo-530	1.0864	0.6931	61,136	12.88 hr

Figure 2: DPO Epoch 1 training loss

RLOO experienced big instabilities in its loss curves. Increasing the batch size as well as K, the number of samples might help. Examining input/outputs showed that many responses were receiving rewards of 0 because it wasn't solving the problem within 1024 tokens or producing an answer in the wrong format. Having more samples could provide more options and useful advantage values, and re-generating responses when rewards are too similar could further advance the model. Further experimentation of RLOO is needed.

6 Discussion

It is difficult to draw conclusions on the comparative effectiveness of different balances of our multi-objective reward signals because of the limitations of our dataset. The three datasets we generated from HelpSteer had a 70-80% overlap, which likely contributed to the minimal differences in model performance between the three. Assuming that the performance differences we observe in the three HelpSteer subsets is statistically significant, entries that achieved a higher verbosity score may be linked with inherently more desirable responses in general, as compared to coherence.

Nevertheless, the effectiveness of our multi-objective training was clear. Our UltraFeedback subsets consisted of fewer than $1/6$ of the full UltraFeedback dataset, yet training on the subsets gave our model a comparable win rate to the DPO model after a second full epoch of training on the full dataset.

These results emphasize the importance of reward signal clarity in preference optimization. Despite training on smaller subsets of UltraFeedback and HelpSteer, the models fine-tuned with multi-objective DPO outperformed those trained on the full dataset. This suggests that clearer, more targeted preference signals can be more valuable than large volumes of noisier data. In particular, our results indicate that aligning reward signals with specific user intents — such as coherence or verbosity — can significantly boost alignment quality without increasing training cost.

Moreover, the high overlap among preference pairs across datasets with different λ values in HelpSteer may have limited the diversity of learned behaviors. This highlights a key challenge in constructing multi-objective datasets: ensuring that different scalarizations meaningfully shift training emphasis. Future work could explore selecting preference examples with greater diversity to better isolate the effects of different reward objectives.

7 Conclusion

Traditional RL methods such as SFT and DPO are solid baseline techniques to align LLM behavior with human preference. Multi-objective DPO offered a further boost in win-rate, highlighting the benefit of a higher quality dataset and mixing reward signals. Further examination of HelpSteer dataset may be needed to draw statistically significant conclusions on the effects of different coherence-verbosity ratios. Our findings reinforce the idea that more data is not always better — especially in the context of alignment via preference learning. Instead, training on compact datasets with well-defined and orthogonal reward signals can lead to more aligned, efficient, and higher-quality model behavior. Multi-objective DPO, in particular, presents a scalable and flexible framework for tailoring model responses to multiple axes of quality, offering a promising path toward personalized or domain-specific LLM alignment.

8 Team Contributions

- **An Doan:** Implemented SFT, created SmolTalk and UltraFeedback dataloaders for the preferencetask and WarmStart and CountDown datasets for the validation task. Trained and evaluated SFT models and multi-objective models.
- **Felicity Huang:** Implemented RLOO, created multi-objective HelpSteer datasets, trained and evaluated models.
- **Linda Liu:** Implemented DPO, created multi-objective UltraFeedback dataset, trained and evaluated extension models.

Changes from Proposal Due to changes in the default project requirements, we switched from implementing RLOO on UltraFeedback to Countdown. Our Bradley-Terry reward model trained on the preference dataset obtained 65% but could not be used.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs. *arXiv preprint arXiv:2402.14740v1* (2024). `https://doi.org/10.48550/arXiv.2402.14740` Version 1, submitted 22 Feb 2024.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model. arXiv:2502.02737 [cs.CL] `https://arxiv.org/abs/2502.02737`

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] `https://arxiv.org/abs/2005.14165`

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. UltraFeedback: Boosting Language Models with High-quality Feedback. arXiv:2310.01377 [cs.CL]

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43, 2 (Jan. 2025), 1–55. `https://doi.org/10.1145/3703155`

Zihao Li, Zhuoran Yang, and Mengdi Wang. 2023. Reinforcement Learning with Human Feedback: Learning Dynamic Choices via Pessimism. arXiv:2305.18438 [cs.LG] https://arxiv.org/abs/2305.18438

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG] https://arxiv.org/abs/1711.05101

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

Jiayi Pan. 2025. Countdown-Tasks-3to4: A dataset of arithmetic "countdown" problems with 3–4 operands. https://huggingface.co/datasets/Jiayi-Pan/Countdown-Tasks-3to4. Accessed: 2025-06-09.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9–32. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.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

Anikait Singh. 2025. cog_behav_all_strategies: Cognitive Behavior Strategies Dataset. https://huggingface.co/datasets/Asap7772/cog_behav_all_strategies. Accessed: 2025-06-09.

Anikait Singh, Sheryl Hsu, Kyle Hsu, Eric Mitchell, Stefano Ermon, Tatsunori Hashimoto, Archit Sharma, and Chelsea Finn. 2025. FSPO: Few-Shot Preference Optimization of Synthetic Preference Data in LLMs Elicits Effective Personalization to Real Users. arXiv:2502.19312 [cs.LG] https://arxiv.org/abs/2502.19312

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A Roadmap to Pluralistic Alignment. arXiv:2402.05070 [cs.AI] https://arxiv.org/abs/2402.05070

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024. HelpSteer2-Preference: Complementing Ratings with Preferences. arXiv:2410.01257 [cs.LG] https://arxiv.org/abs/2410.01257

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. 2023. HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM. arXiv:2311.09528 [cs.CL]

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025. Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv:2401.11817 [cs.CL] https://arxiv.org/abs/2401.11817