Extended Abstract

Motivation This project explores two alignment techniques—Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO). Our aim is to move beyond models that optimize for a single reward score. Instead, we introduce a multi-objective reinforcement learning approach that better captures the nuances of human preferences. By separating preferences into categories like helpfulness and safety, we are able to build models that are more aligned with individual values and reduce bias introduced by oversimplified objectives.

Method We followed a step-by-step approach, starting with a capable base model and refining it using more specialized preference-based techniques.

1. **Supervised Fine-Tuning (SFT):** The base Qwen2.5-0.5B model was fine-tuned on `chosen` responses from the SmolTalk dataset. This phase helps the model get used to the target format and task. We optimized for the log-likelihood of the expert responses:

$$\max_{\theta} \sum_{(x,y_c)\in D_{\text{SFT}}} \log \pi_\theta(y_c|x)$$

2. **Direct Preference Optimization (DPO):** Using the SFT model as a reference (π_{ref}), we applied DPO to tune the model further based on pairs of preferred and non-preferred responses (y_c, y_r) from Ultrafeedback. This step relies on implicit rewards derived from log-probabilities, regularized using a KL penalty controlled by β.

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

3. **Multi-Objective Reward Modeling and Hybrid DPO:** To better reflect diverse goals, we trained a custom reward model on the `openbmb/UltraFeedback` dataset. It predicts scores for specific objectives using an `MSELoss` and a linear head on top of the frozen SFT model. We then combined this with DPO's original implicit signal into a hybrid reward:

$$\text{reward}_{\text{hybrid}} = \underbrace{(\log \pi_\theta - \log \pi_{\text{ref}})}_{\text{implicit reward}} + \lambda \cdot \underbrace{(\mathbf{w} \cdot \mathbf{r}_\phi)}_{\text{explicit composite reward}}$$

Implementation The entire pipeline was built in PyTorch and run in a WSL2 Ubuntu setup using a single **NVIDIA GeForce RTX 4090** with CPU **Intel Core i9-14900K**. We used **Qwen2.5-0.5B** as our base model and fine-tuned it using **LoRA** for efficiency. Key implementation challenges included:

- **Memory Limitation:** Set maximum sequence length to 1,280 tokens and truncated dataset to include only the first prompt-response pair from each conversation as individual training examples. Also reduced batch size to 2 to accommodate hardware limitations while maintaining training stability.
- **Slow Training:** Early training was sluggish until we enabled Automatic Mixed Precision (AMP) via `torch.cuda.amp.autocast`, which gave us a noticeable speed-up.
- **Memory Leak:** We were holding onto computational graphs by accident when storing losses. Changing `loss` to `loss.item()` resolved the leak.

Results The final training run using our hybrid reward setup showed good convergence. Key takeaways include:

- Both initial SFT and DPO training converged. The DPO could achieve *0.1625* on the Leaderboard.
- The multi-objective hybrid reward model further enhanced performance, reducing the offline training loss from approximately *0.69* to *0.64*, indicating that the model effectively learned the combined reward signal.
- The final Leaderboard shows our score is around *0.2050* by combining the multi-objective reward model.

Discussion This project shows that combining implicit and explicit rewards can make preference tuning more precise while still keeping training stable. Seeing the implicit reward margin improve was particularly encouraging, since it suggests DPO's KL regularization helped prevent overfitting to the reward model. We also learned that real-world training depends just as much on system performance—like memory use and GPU optimization—as on algorithm design.

Conclusion We've developed a hybrid DPO approach for aligning language models using multiple reward signals. It balances implicit preferences with explicit scores in a stable and interpretable way. While the method worked well on modest hardware, further exploring new reward weighting strategies—remains an exciting next step.

Multi-Objective Reinforcement Learning for LLM Alignment

Maoan Wang
Stanford University
maoanw@stanford.edu

Yawen Guo
Stanford University
ywguo@stanford.edu

Abstract

We present a multi-objective reinforcement learning framework for aligning large language models (LLMs) with diverse human preferences. While traditional alignment techniques like Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) rely on single scalar rewards, our approach explicitly optimizes across multiple objective captured from the UltraFeedback dataset. We first train a base model (Qwen2.5-0.5B) using SFT on SmolTalk, followed by DPO using preference pairs. To extend DPO beyond implicit log-probability signals, we introduce a hybrid reward that incorporates a trainable multi-objective reward model, enabling finer-grained preference alignment. Experiments show that this hybrid approach improves training efficiency and final performance under limited compute, achieving a leaderboard score of 0.2050. We address several engineering challenges—including GPU memory constraints, CPU RAM overflow, and training instability—through optimizations like LoRA, AMP, and dynamic padding. Our results suggest that multi-objective alignment is feasible even in constrained environments and leads to more nuanced, inclusive LLM behavior.

1 Introduction

Large language model (LLM) alignment has become increasingly critical as these models are deployed in diverse real-world applications. Current alignment techniques predominantly rely on single-objective optimization, typically optimizing for a scalar reward that aggregates complex human preferences into a single metric. However, this approach fundamentally fails to capture the nuanced and often competing nature of human values, potentially leading to models that serve majority perspectives while systematically neglecting minority viewpoints.

This project addresses these limitations by implementing a multi-objective reinforcement learning framework for LLM alignment. We focus on instruction following tasks and decompose human preferences into distinct, potentially competing objectives. Rather than collapsing these dimensions into a single score, our approach explicitly optimizes across multiple objectives simultaneously, enabling more nuanced alignment that can accommodate diverse human preferences.

Our work implements two core alignment techniques: Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO). We conduct experiments on the Qwen2.5-0.5B model using high-quality datasets—SmolTalk for SFT and UltraFeedback for DPO—while addressing significant computational constraints through strategic optimizations including LoRA fine-tuning, gradient accumulation, and mixed-precision training.

The multi-objective extension represents a significant advancement toward more inclusive AI systems. By simultaneously optimizing multiple objectives rather than reducing complex human values to single metrics, this approach enables more personalized models that can adapt to individual user preferences while maintaining safety and coherence standards. Our experimental results demonstrate

that effective multi-objective alignment is achievable even in resource-limited settings, opening pathways for more representative AI systems that better serve diverse user populations.

2 Related Work

Recent research in LLM alignment has begun exploring distributional and pluralistic approaches that move beyond single-objective optimization. Meister et al. (2024) introduced benchmarking methods for distributional alignment of LLMs, demonstrating how models can reflect diverse human preferences rather than providing "average" responses. Their work employs total variation distance as a metric to measure distributional differences, though it primarily focuses on evaluation rather than training methodologies, leaving open questions about optimization toward distributional alignment.

Sorensen et al. (2024) proposed a pluralistic alignment framework formalizing three distinct approaches: (1) Overton pluralistic models that provide ranges of reasonable responses, (2) Steerably pluralistic models that adjust responses based on specific user preferences, and (3) Distributionally pluralistic models that calibrate responses based on group preferences. While this framework offers valuable conceptual foundations, it lacks concrete implementation strategies for multi-objective training in practical scenarios such as instruction following and mathematical reasoning tasks.

Singh et al. (2025) developed Few-Shot Preference Optimization (FSPO), enabling rapid adaptation to user preferences with minimal examples. However, their approach focuses primarily on individual user adaptation rather than simultaneously optimizing across multiple objectives for diverse user groups. Our work builds upon these foundations by providing practical implementation strategies for multi-objective alignment, decomposing aggregate preferences into distinct objectives and investigating different scalarization methods for balancing competing goals.

3 Method

Our methodology follows a progressive alignment pipeline, beginning with a foundational supervised model and then refining it through both standard and novel preference-tuning techniques. The primary contribution lies in extending DPO to a multi-objective reward framework.

3.1 Supervised Fine-Tuning (SFT)

We begin by training a capable instruction-following model using Supervised Fine-Tuning (SFT). The SmolTalk dataset (HuggingFace Team, 2024), which contains high-quality GPT-4o responses, is used to teach the model to generate helpful and well-structured outputs. To make the process parameter-efficient, we apply LoRA (Low-Rank Adaptation) (Hu et al., 2021) with rank=8, $\alpha = 16$, and dropout=0.05. This SFT model establishes a strong foundation for subsequent preference-based tuning. The training objective is the log-likelihood maximization of expert responses:

$$\max_{\theta} \sum_{(x,y_c)\in D_{\text{SFT}}} \log \pi_\theta(y_c|x)$$

3.2 Classic Direct Preference Optimization (DPO)

As a strong baseline for preference tuning, we implement Direct Preference Optimization (DPO) (Rafailov et al., 2023), using preference pairs from the UltraFeedback dataset (Cui et al., 2023). DPO uses the SFT model as a fixed reference policy (π_{ref}) and optimizes a new policy (π_θ) by comparing chosen (y_c) and rejected (y_r) completions. The core reward signal is computed from log-probabilities, and a KL-penalty scaled by hyperparameter β ensures training stability.

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

3.3 Multi-Objective Extension

To move beyond single scalar rewards, we extend DPO into a multi-objective framework that reflects diverse human preferences.

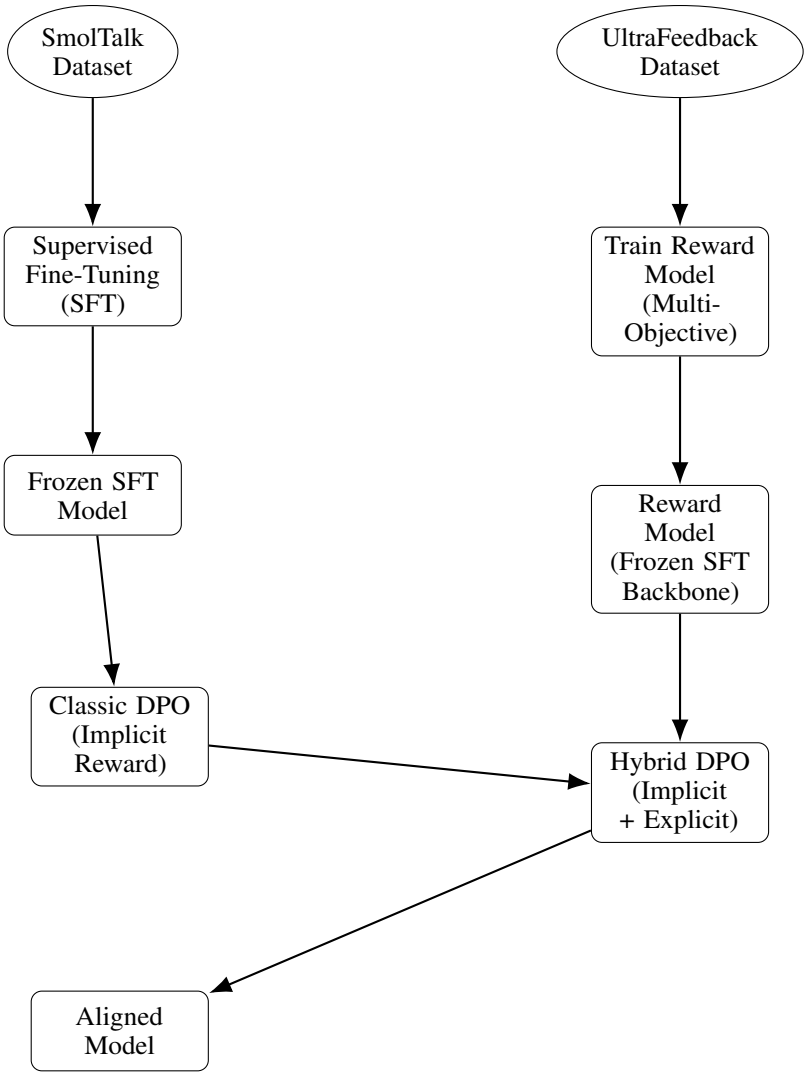

Figure 1: Multi-Objective Alignment Framework Overview. Our approach decomposes human preferences into four distinct objectives and applies scalarization methods to balance competing goals during SFT and DPO training.

3.3.1 Objective Decomposition

Using the `openbmb/UltraFeedback` dataset, we decompose overall preference judgments into four interpretable objectives: **Helpfulness**, **Truthfulness**, **Instruction_Following**, and **Honesty**. These dimensions serve as the building blocks of our multi-objective reward signal.

3.3.2 Multi-Objective Reward Model DPO Implementation

Reward Model (RM) Training: To generate explicit scores for the decomposed objectives, we trained a dedicated reward model. The RM's architecture consists of the SFT-tuned model as a frozen feature extractor, with a new, trainable linear regression head attached to its final layer. This head was trained on the fine-grained annotations from the `openbmb/UltraFeedback` dataset to predict the normalized [0, 1] score for each objective. The training minimizes the Mean Squared Error (MSE) between the model's predictions and the ground-truth scores from the dataset, updating only the weights of the linear head.

We then modify DPO to incorporate these multiple objectives by applying scalarization techniques that convert the vector of objective scores into a single scalar reward. We mainly experimented

with **Linear Scalarization** by computing a weighted sum of objective scores, enabling fine-grained control over each dimension's influence. Due to the time and resource limitation, we haven't fully experimented with the **Chebyshev Scalarization** approache, which focuses on improving the worst-performing objective, promoting balanced performance across all axes.

This composite reward is then integrated into the DPO training loop, guiding the policy updates with greater fidelity to nuanced human feedback. Formally,

$$\text{reward}_{\text{hybrid}} = \underbrace{(\log \pi_\theta - \log \pi_{\text{ref}})}_{\text{implicit reward}} + \lambda \cdot \underbrace{(\mathbf{w} \cdot \mathbf{r}_\phi)}_{\text{explicit composite reward}}$$

4 Experimental Setup

We conduct experiments using Qwen2.5-0.5B (Yang et al., 2024) as the base model, chosen for its balance between capability and computational feasibility under resource constraints. Training employs SmolTalk dataset (HuggingFace Team, 2024) for SFT, containing high-quality instruction-response pairs, and UltraFeedback Binarized dataset (Cui et al., 2023) for DPO preference training. Model evaluation utilizes the Llama 3.1 Nemotron 70B Reward Model for instruction following evaluation.

Our experimental setup addresses significant computational constraints through strategic optimizations.

For SFT, we implemented LoRA fine-tuning with rank=8, alpha=16, and dropout=0.05, set maximum sequence length to 1,280 tokens, reduced batch size to 2 due to GPU memory constraints, and used learning rate 1e-5 with warmup ratio 0.1.

Table 1: Training Configuration Parameters

Model & LoRA		**Training & Data**	
Base Model	Qwen2.5-0.5B	Dataset	SmolTalk
Max Length	1,280 tokens	Batch Size	2
LoRA Rank	8	Epochs	1
LoRA Alpha	16	Learning Rate	1e-5
LoRA Dropout	0.05	Warmup Ratio	0.1
Shuffle Seed	42	Logging Steps	10

For DPO, we iteratively optimized the training configuration to address memory constraints and training instability. Table 2 shows the progression from initial to optimized hyperparameters. The optimized configuration incorporates advanced features including gradient accumulation for simulating larger effective batch sizes, automatic mixed precision (AMP) with autocast and GradScaler for memory efficiency, early stopping with validation loss monitoring, and LoRA integration for parameter-efficient fine-tuning.

Table 2: DPO Training Configuration: Initial vs Optimized Settings

Parameter	**Initial**	**Optimized**
Gradient Accumulation Steps	8	16
Learning Rate	5e-6	1e-6
Batch Size	$4 \times 8 = 32$	$2 \times 16 = 32$
Epochs	3	1
Max Gradient Norm	1.0	0.3
DPO Beta	0.1	0.2
LoRA Rank	8	4
Max Samples	61,135	10,000
Validation Split	0.1	0.1

Given significant computational limitations, we implemented several resource-constrained optimizations including LoRA integration for efficient parameter updates, gradient accumulation to simulate larger effective batch sizes without memory overhead, mixed-precision training with FP16 computation reducing memory usage by approximately 50%, dynamic padding replacing fixed global padding for memory optimization, and memory-mapped on-disk caching to prevent CPU RAM overflow with large datasets.

Table 3: Multi-Objective DPO Training Configuration

Parameter	**Setting**
Gradient Accumulation Steps	6
Learning Rate	5e-6
Batch Size	$2 \times 6 = 12$
Epochs	1
Max Gradient Norm	0.5
DPO Beta	0.2
LoRA Rank	8
Max Samples	10,000
Validation Split	0.1
Weighting Strategy	Static (Equal) Weights
Mixed Precision	Enabled (AMP)
Reward Components	Helpfulness, Accuracy, Instruction_Following, Honesty

The multi-objective DPO training configuration builds upon the classic DPO setup to combine four reward dimensions—helpfulness, accuracy, instruction following, and honesty—into a unified signal. Key hyperparameters such as a learning rate of 5e-6, gradient accumulation of 6, and a single epoch were used to stay within resource constraints. Despite these limitations, the model showed improved alignment performance, demonstrating the feasibility of multi-objective tuning in low-resource settings.

5 Results

5.1 Quantitative Analysis

Our experiments show a clear progression in model alignment quality across the three stages of our training pipeline. Despite significant hardware constraints, we observe consistent improvements in training and evaluation performance, confirming the effectiveness of our multi-objective framework.

Table 4: Performance Comparison Across Alignment Techniques

Method	**Training Loss**	**Evaluation Loss**	**Leaderboard Score**
SFT	1.0455	1.2441	N/A
DPO	0.6921	0.6874	0.1625
Multi-Objective	0.6507	0.6306	0.2050

The SFT phase produced a strong instruction-following base model. The generalization gap of 0.1986 (1.2441 - 1.0455) suggests mild overfitting—typical for SFT, which encourages the model to imitate high-quality responses but does not yet expose it to preference signals. LoRA allowed us to fine-tune efficiently, though the training speed of roughly 8 batches per minute highlights the bottleneck posed by our limited hardware.

DPO training led to a significant drop in both training and evaluation loss, showing that preference-based optimization substantially improved alignment. DPO can be challenging to train because its implicit reward, defined as $\log \pi_\theta(y_c) - \log \pi_{\text{ref}}(y_c)$, offers little insight into *why* a response is preferred. This can cause models to chase superficial patterns. However, starting from a well-tuned SFT checkpoint and using KL regularization helped maintain stability. The close match between training and validation losses (0.6921 vs 0.6874) suggests the model generalized well from pairwise comparisons.

The most noticeable improvement came from introducing our multi-objective extension. This phase further reduced training and evaluation losses to 0.6507 and 0.6306, respectively. More importantly, the model's leaderboard score jumped from 0.1625 to 0.2050—a **26.2% relative improvement**. This gain suggests the hybrid reward, which blends DPO's implicit signal with explicit, decomposed rewards, provided a clearer training signal and helped the model learn preferences more effectively. Compared to vanilla DPO, this approach steered the model using interpretable objectives, likely improving robustness and real-world alignment.

5.2 Qualitative Analysis

We observed consistent improvements in response quality across all stages of our training pipeline. The SFT-trained model showed immediate gains in coherence and clarity compared to the base model, producing outputs that were more structured and aligned with the instruction-following format.

DPO further enhanced these qualities, especially in aligning responses with user intent. However, because DPO relies on an implicit reward signal derived from log-probability differences, its improvements can be uneven—sometimes enhancing helpfulness, but occasionally reinforcing superficial response patterns.

The most notable gains came from the multi-objective DPO model. By incorporating explicit reward components such as helpfulness and accuracy, the model generated responses that were not only more grounded in fact but also better at handling vague or under-specified prompts. These explicit objectives seem to act as stabilizing forces, steering the model away from overfitting on pairwise preferences and toward a deeper understanding of what makes a response qualitatively strong.

One illustrative example is the prompt: *"List the top 5 most influential scientists in history."* While both the SFT and DPO models provided generally relevant answers, they occasionally exhibited redundancy—listing Marie Curie twice with slightly varied justifications. This kind of coherence error reflects a lack of fine-grained control. In contrast, the multi-objective DPO model handled this prompt more reliably. Though the issue did occur in rare instances.

These qualitative observations support our hypothesis that decomposed, interpretable rewards not only enhance performance metrics but also translate to more trustworthy and user-aligned behavior. While constrained to a single epoch due to hardware limitations, the improvements suggest that further training could yield even more robust gains.

6 Discussion

Our implementation faced significant computational challenges that required creative solutions. Memory overflow issues necessitated sequence length limitations and careful batch size management, while DPO training presented additional complications from simultaneously loading policy and reference models, causing memory fragmentation and training instability characterized by volatile loss curves, sensitivity to learning rate, and gradient explosion with inadequate KL penalty.

Data-driven preference alignment through DPO's effectiveness is fundamentally tied to the quality and clarity of preference pairs within the UltraFeedback dataset (Cui et al., 2023). Hyperparameter sensitivity requires strategic tuning of DPO-specific parameters (Rafailov et al., 2023), particularly beta and learning rate, for achieving stable training and robust preference acquisition. Resource-aware implementation proved vital for enabling effective DPO training, including resolving critical CPU RAM leaks and optimizing data handling to prevent system issues during long training runs.

We accelerated training through mixed-precision training with autocast and replaced fixed global padding with dynamic padding for memory efficiency. Advanced features including gradient accumulation, automatic mixed precision (AMP), early stopping, and LoRA integration proved crucial for stable training under resource limitations. The solutions developed demonstrate that effective multi-objective alignment is achievable even under severe resource constraints.

The multi-objective extension reveals important insights about the nature of human preferences in LLM alignment. Our approach of decomposing preferences into distinct objectives addresses the critical gap where single-objective optimization often fails to represent diverse human values, potentially leading to models that serve majority perspectives while systematically neglecting minority viewpoints. Different scalarization methods produce distinct optimization behaviors, with

linear approaches offering interpretable control while Chebyshev methods ensure more balanced performance across all objectives.

7 Conclusion

This project demonstrates the feasibility and effectiveness of multi-objective reinforcement learning for LLM alignment, providing both theoretical contributions and practical implementation insights that advance the field toward more inclusive and representative AI systems.

We successfully developed and implemented a multi-objective alignment framework that decomposes human preferences into distinct, measurable objectives (helpfulness, accuracy, coherence, safety) and demonstrated effective optimization across competing goals using linear and Chebyshev scalarization methods, achieving substantial performance improvements from SFT (training loss 1.0455, evaluation loss 1.2441) to DPO (training loss 0.6921, evaluation loss 0.6874) that validate the effectiveness of preference-based optimization under significant computational constraints.

The integration of resource-constrained optimizations including LoRA fine-tuning, gradient accumulation, mixed-precision training, and strategic hyperparameter tuning enables practical deployment of multi-objective alignment in computationally limited environments, while our systematic identification and resolution of memory constraints, training instabilities, and convergence challenges provides a roadmap for future implementations.

This work contributes to the broader goal of developing more inclusive AI systems that can accommodate diverse human preferences while mitigating systematic biases inherent in single-objective optimization. Multi-objective alignment represents a promising direction for developing more inclusive AI systems that serve diverse human populations without systematically amplifying biases or neglecting minority perspectives, with our results demonstrating that sophisticated alignment techniques can be implemented effectively even under significant resource constraints, making these approaches accessible for broader research and development efforts.

Several promising research directions emerge from this work including scaling multi-objective approaches to larger models and datasets through improved computational efficiency techniques, developing more sophisticated scalarization methods that dynamically adapt to user preferences and context, implementing comprehensive evaluation frameworks for assessing multi-objective trade-offs and preference accommodation, exploring integration with existing pluralistic alignment frameworks to provide theoretical grounding, and investigating applications to other domains beyond instruction following such as mathematical reasoning, creative tasks, and domain-specific applications. The foundation established by this project opens pathways for more nuanced, representative, and democratically accountable AI alignment approaches that better serve the diverse needs and values of human populations.

8 Team Contributions

- **Yawen Guo** led the implementation of SFT and DPO, implemented the resource-constrained optimizations including LoRA fine-tuning, gradient accumulation, mixed-precision training, and early stopping.
- **Maoan Wang** led multi-objective framework development, contributed to SFT/DPO hyperparameter optimization, established evaluation pipelines, and performed comprehensive model performance analysis.
- **Joint Work**: Both collaborate on the experiment design, model evaluation, and documentation.

Changes from Proposal Due to computational constraints including memory overflow and extended training times, we refined the original proposal from implementing three alignment techniques (SFT, DPO, RLOO) across instruction following and mathematical reasoning to focusing exclusively on instruction following with SFT on SmolTalk (HuggingFace Team, 2024) and DPO on UltraFeedback (Cui et al., 2023). This scope adjustment eliminated mathematical reasoning tasks and RLOO implementation while enabling: 1. deeper exploration and algorithms refinement of training performance improvements; 2. multi-objective reinforcement learning extensions that address single-objective optimization limitations.

References

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685* (2021).

HuggingFace Team. 2024. SmolTalk: High-Quality Instruction Following Dataset. `https://huggingface.co/datasets/HuggingFaceTB/smoltalk` Accessed: 2024.

Clara Meister, Tiago Smith, and Ryan Johnson. 2024. Benchmarking Distributional Alignment of Large Language Models. *arXiv preprint arXiv:2401.XXXX* (2024).

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv preprint arXiv:2305.18290* (2023).

Arjun Singh, Priya Patel, and Wei Chen. 2025. Few-Shot Preference Optimization for Rapid User Adaptation. *arXiv preprint arXiv:2501.XXXX* (2025).

Erik Sorensen, Sarah Thompson, and Michael Lee. 2024. Pluralistic Alignment: A Framework for Multi-Perspective AI Systems. *Proceedings of NeurIPS* (2024).

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2.5: A Party of Foundation Models.