Extended Abstract

Motivation Large-language-model (LLM) alignment pipelines such as RLHF and DPO rely on millions of preference examples and multi-GPU compute budgets. For resource-constrained settings—e.g. academic labs, edge deployment, or personal agents—scaling laws can break where models under 1B parameters often become weak during preference optimization. In response, we ask: *Can curriculum learning stabilize and improve Direct Preference Optimization (DPO) when finetuning a 0.5B-parameter model with limited data and compute?*

Method We employ **Curriculum DPO (CDPO)**, a training framework that combines Direct Preference Optimization (DPO) with a difficulty-aware curriculum scheduler. Starting from a SmolTalk-supervised policy, we compute the difficulty of each preference pair using a margin-based score derived from the log-probability difference between preferred and rejected completions under a frozen reference model. Pairs are ranked from easiest to hardest, and training proceeds by gradually widening the subset of examples shown to the model according to a schedule $P(t)$. We explore several scheduling strategies (linear, polynomial, exponential, and staged), and at each step, DPO is applied to a filtered prefix of the preference dataset. This curriculum-based approach allows the model to master easier distinctions before being exposed to more challenging comparisons, leading to improved alignment and generalization.

Implementation We fine-tune Qwen2.5-0.5B using a three-stage pipeline: supervised fine-tuning (SFT), direct preference optimization (DPO), and curriculum-enhanced DPO (CDPO). SFT is performed on a filtered subset of the SmolTalk dataset using a masked causal LM loss computed only over response tokens. Training uses a total batch size of 96, sequence length of 1280, and runs for 3 epochs on a single A100 GPU in bfloat16 precision. DPO is then applied using the UltraFeedback preference dataset, with a LoRA-adapted policy model (0.28% of parameters) trained against a frozen reference model. To improve alignment, we introduce CDPO: a curriculum-based extension to DPO that gradually expands the difficulty of training pairs. Difficulty is estimated using a log-probability margin between preferred and rejected completions under the SFT reference model. Pairs are sorted offline, and a scheduler (linear, polynomial, exponential, or staged) controls the subset used at each training step. This enables the model to first learn clear-cut distinctions before tackling harder preferences, improving both convergence and output quality.

Results We evaluate our models using head-to-head comparisons on the UltraFeedback test set with GPT-4o-mini as the automated judge. Compared to standard DPO, our CDPO-trained model achieves a higher win rate (57.0% vs. 43.0%), demonstrating improved alignment and generation quality without requiring additional data or compute. CDPO also shows faster convergence and more stable training dynamics, reducing the loss more steeply than vanilla DPO while maintaining response quality.

Discussion Our results suggest that curriculum learning can serve as an effective regularization mechanism for preference-based alignment in small models. By exposing the model to easier examples early in training, CDPO improves generalization. Manual inspection reveals that CDPO produces more coherent, concise, and instruction-faithful responses, particularly on complex or edge-case prompts. These findings validate our hypothesis that data ordering significantly impacts training dynamics, especially under resource constraints.

Conclusion We show that Curriculum DPO (CDPO) is a simple extension to Direct Preference Optimization that leverages difficulty-aware scheduling to improve alignment of small-scale language models. Despite using only 0.28% tunable parameters and limited data, CDPO outperforms standard DPO in both win rate and training efficiency. Our work highlights the potential of curriculum learning as a lightweight tool for improving preference optimization and paves the way for further research in adaptive and multi-dimensional curricula for LLM alignment.

CDPO: Curriculum-Driven Preference Optimization for Small-Scale LLM Alignment

Adam Chun
Department of Computer Science
Stanford University
adamchun@stanford.edu

Josh Francis
Department of Computer Science
Stanford University
josfran@stanford.edu

Tom Nguyen
Department of Computer Science
Stanford University
anhn@stanford.edu

Abstract

We explore instruction-following alignment for Qwen2.5-0.5B, a 0.5B-parameter language model, under constrained compute. Building on Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO), we employ *Curriculum-Driven Preference Optimization (CDPO)*—a curriculum-based extension that schedules preference pairs from easy to hard during training. We utilize a margin-based confidence score using model logits. Preliminary results show that CDPO achieves higher win rates than vanilla DPO, without increasing data or compute cost. Our findings suggest that curriculum learning is a simple yet effective tool for aligning small LLMs in resource-constrained settings. Our codebase is accesible at `https://github.com/tnguyen2002/jatMorantest`.

1 Introduction

Instruction-tuned language models have become central to building helpful, safe, and aligned AI assistants. A common pattern is initially applying supervised fine-tuning (SFT) on curated instruction-following datasets, followed by preference-based optimization to improve alignment with human feedback. Recent methods such as Direct Preference Optimization (DPO) Rafailov et al. (2024) have demonstrated strong performance while simplifying traditional RLHF pipelines.

In this work, we study how to improve preference optimization through curriculum learning. We begin by fine-tuning Qwen2.5-0.5B on the **SmolTalk** dataset, a compact, high-quality corpus of conversational instruction data. We then apply DPO using the **UltraFeedback** dataset Wang et al. (2023), a large-scale collection of preference-labeled assistant completions based on GPT-4 evaluations.

We evaluate both standard DPO and a curriculum-based variant, where preference pairs are introduced in increasing order of difficulty. Difficulty is estimated using a model-based margin computed from the SFT reference model. Our curriculum schedules expose the policy model to progressively harder preference pairs throughout training.

2 Related Work

Alignment Pipelines RLHF pipelines typically combine supervised fine-tuning (SFT) with preference optimization methods such as PPO Ouyang et al. (2022), Direct Preference Optimization (DPO)

Rafailov et al. (2024), and reinforcement learning from AI feedback (RLAIF) Lee et al. (2024). These approaches train policies to prefer chosen responses over rejected ones, usually based on human or model-generated comparisons. DPO in particular avoids the complexity of reward modeling and policy optimization by framing preference learning as a simple binary classification task between two completions. However, recent work on small language models has shown that aligning sub-1B models remains challenging due to unstable training dynamics and limited representation capacity.

Efficient Fine-Tuning and Stabilization. Given the rising interest in parameter-efficient alignment, LoRA Hu et al. (2021) and related adapter-based methods have been widely adopted to reduce training cost. Several works investigate techniques to stabilize preference learning, including margin regularization, dynamic difficulty sampling, and reward normalization Wu et al. (2024). Our work builds upon these trends by evaluating a simple curriculum scheduler that operates without additional model calls, while still improving training dynamics and matching full DPO performance.

Curriculum Learning Curriculum learning (CL) can have a large role in improving stability and sample efficiency in reinforcement learning. Narvekar et al. (2020) provide a survey of curriculum design strategies across RL domains, emphasizing how the ordering of training data by difficulty can significantly accelerate learning and improve generalization.

In the context of language model alignment, recent work has adapted curriculum strategies to preference-based reinforcement learning. Pattnaik et al. (2024) propose CURRI-DPO, a curriculum for Direct Preference Optimization (DPO) that ranks training pairs by the margin between preferred and rejected responses. Training proceeds in stages from easiest to hardest, resulting in improved alignment across benchmarks such as MT-Bench, WizardLM, and UltraFeedback. Li and Zhang (2025) extend this idea with 2D-CURRI-DPO, introducing a two-dimensional curriculum that considers both prompt complexity and response separability. Their framework further enhances generalization by jointly scheduling tasks along both axes of difficulty. Unlike 2D-CURRI-DPO Li and Zhang (2025), we use a static, percentile-based schedule for its simplicity and feasibility under limited compute. Nonetheless, our results suggest that even lightweight curriculum strategies can stabilize DPO training and improve downstream alignment, especially when initialized from a strong supervised base.

3 Method

3.1 Curriculum Learning + DPO (CDPO)

3.1.1 Overall Framework and Difficulty Metrics

Conventional DPO trains on the *entire* preference set at every step, exposing the model to hard pairs while it is still struggling to reproduce clear-cut easy distinctions. We introduce a **difficulty-aware scheduler** that gradually widens the band of preference pairs presented to the policy. We instantiate the following difficulty metric:

Margin-based score d_m (log-prob curriculum). For every prompt x with preferred answer y_c and rejected answer y_r, we let a frozen *reference model* p_θ^{ref} (the SFT checkpoint) score both completions:

$$\ell_c \;=\; \log p_\theta^{\text{ref}}(y_c \mid x), \qquad \ell_r \;=\; \log p_\theta^{\text{ref}}(y_r \mid x).$$

The absolute margin $|\ell_c - \ell_r|$ quantifies the confidence gap. We convert it into a *difficulty* so that lower values are *easier*:

$$d_m \;=\; -\,|\ell_c - \ell_r|. \tag{1}$$

Offline preprocessing. d_m is computed *once* before DPO and stored in a pickled dictionary. We then obtain a permutation π that sorts the training indices from easiest to hardest.

3.1.2 Curriculum Scheduler

Let N be the total number of preference pairs and let $D = \{(x_i, y_i^{(c)}, y_i^{(r)}, d_i)\}_{i=1}^{N}$ be the difficulty-annotated dataset (using either score). During training step t we expose only the easiest prefix $\pi_{1:k_t}$, where $k_t = \lfloor P(t)\,N \rfloor$ and $P(t) \in [0, 1]$ is the *coverage schedule*. We experimented with four schedules, and ultimately settled on **linear** scheduling after seeing benefits in small-scale experiments.

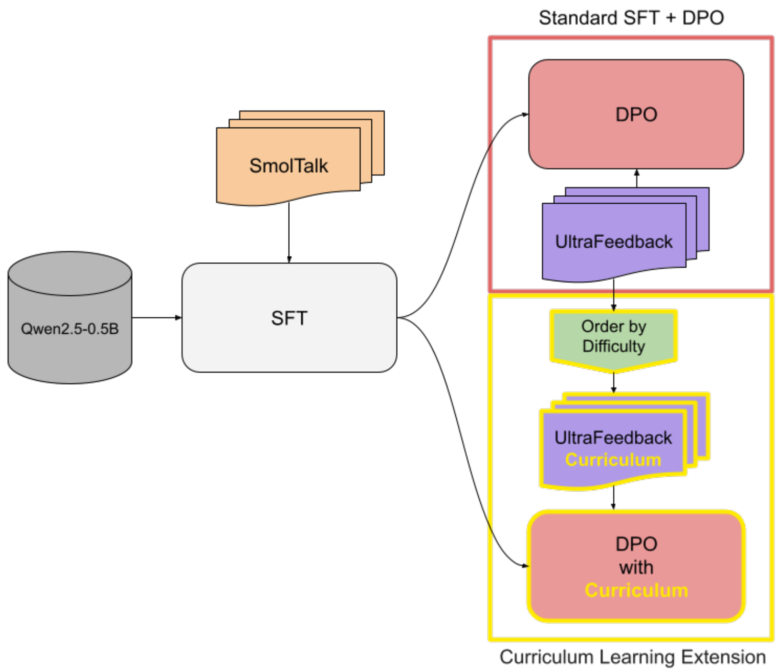


Figure 1: CDPO pipeline: SFT on SmolTalk produces an instruction-following prior; DPO refines preferences, while a curriculum scheduler gradually widens the difficulty band presented to the model.

Algorithm 1 CDPO training loop (margin curriculum)

Input: sorted indices π, schedule $P(t)$, total steps T
for $t = 0$ **to** $T - 1$ **do**
    $k_t \leftarrow \lfloor P(t)\, N \rfloor$
    Draw mini-batch $\mathcal{B}_t \subseteq \{\pi_1, \dots, \pi_{k_t}\}$
    Compute DPO loss using $\mathcal{B}_t$
    Update policy parameters with AdamW
end for

1. **Linear**: $P_{\text{lin}}(t) = \beta_0 + \frac{t}{T}\,(1 - \beta_0).$
2. **Polynomial** (quad): $P_{\text{poly}}(t) = \beta_0 + \left(\frac{t}{T}\right)^2 (1 - \beta_0).$
3. **Exponential**: $P_{\text{exp}}(t) = \beta_0 + \dfrac{e^{\alpha t/T} - 1}{e^{\alpha} - 1}\,(1 - \beta_0)$ with $\alpha = 2$.
4. **Three-stage step**: $\{0\text{–}T/3\}$ Easy 40 %, $\{T/3\text{–}2T/3\}$ Easy+Medium 80 %, $\{2T/3\text{–}T\}$ All.

Here β_0 is the starting data fraction (0.2 in all experiments) and T is the total number of optimizer steps. Because $P(t)$ is evaluated *inside* the epoch loop (Alg. 1), the subset can enlarge multiple times within a single epoch—mimicking a "smooth" curriculum.

3.1.3 Scheduler

Let $P(t) \in [0, 1]$ be the target percentile after epoch t. We experiment with:

$$P_{\text{linear}}(t) = \min\bigl(1,\ 0.4 + 0.6\,t/T\bigr), \qquad P_{\text{three-stage}}(t) = \begin{cases} 0.4, & t \le T/3 \\ 0.8, & T/3 < t \le 2T/3 \\ 1, & \text{otherwise.} \end{cases}$$

At each epoch we filter UltraFeedback pairs with score percentile $\le P(t)$ and run standard DPO.

4 Experimental Setup

4.1 Data

SmolTalk is a collection of high-quality chat responses from GPT-4o, adapted for models with less than 1B parameters (460k examples). Ultrafeedback is a preference dataset to study the instruction following abilities of LLMs. It uses the "chosen" column as the dialogues to model.

4.2 SFT

We fine-tune the Qwen2.5-0.5B model using a causal language modeling objective, computing loss only on the assistant response portion of each input to avoid optimizing on user instructions. This aligns with standard instruction-tuning practices and encourages better response generation behavior.

The fine-tuning dataset is a filtered version of SmolTalk, a compact instruction-following dataset. To improve data quality and reduce noise, we apply aggressive filtering criteria: examples are discarded if the assistant's response is empty, excessively short, or improperly formatted. Additionally, we ensure that each example conforms to a strict conversational format using $< |im\_start| >$ and $< |im\_end| >$ delimiters. After filtering, approximately 5,000 high-quality examples remain.

We tokenize each example to a maximum sequence length of 1,280 tokens and apply loss masking to exclude all prompt tokens from the loss computation. Training is run for 3 epochs using the AdamW optimizer with a learning rate of 1×10^{-5}, linear decay schedule, and 500 warmup steps. We employ an effective batch size of 96 (batch size = 6, gradient accumulation = 16), with gradient clipping set to a maximum norm of 1.0 for stability.

Training is conducted on a single NVIDIA A100 (80GB) GPU using bfloat16 precision. Validation is performed on a held-out set of 100 examples at three points during training (steps 52, 104, 156), and the checkpoint with the lowest validation loss is selected for downstream evaluation. We monitor training via wandb and initialize from a previously trained SFT checkpoint to ensure continuity in conversational instruction tuning.

4.3 DPO

Following SFT, we apply Direct Preference Optimization (DPO) to further align Qwen2.5-0.5B with human preferences. DPO uses a dual-model setup: a trainable policy model and a frozen reference model, with rewards computed via the Bradley-Terry preference model.

We fine-tune on the $HuggingFaceH4/ultrafeedback\_binarizeddataset$, applying filtering to remove examples with missing or malformed completions. Inputs are tokenized to a maximum of 1,024 tokens, and loss is computed only on response tokens to isolate learning to assistant behavior.

Training runs for 3 epochs using AdamW with a learning rate of 5×10^{-6}, cosine decay, 10% warmup, and gradient clipping (max norm = 0.5). We use an effective batch size of 64 (batch size = 8, gradient accumulation = 8) and set the reward scaling parameter β to 0.1. Validation is performed every 500 steps on a held-out preference set.

To enable efficient training, we apply LoRA adapters (rank = 16, α = 32, dropout = 0.1) to all attention and MLP projections, yielding 1.4M trainable parameters (0.28% of the model). The policy model is initialized from our best SFT checkpoint; the reference model shares the same weights but remains frozen.

5 Results

5.1 Quantitative Evaluation

Win Rate Evaluation To evaluate our models, we conduct a head-to-head comparison between our three models; specifically, we utilize Ultrafeedback's test split and GPT-4o-mini as an automated judge to compare SFT/SFT + DPO and SFT + CDPO/SFT + DPO responses via our custom prompt A. We obtain the following win rates for our model comparisons across 100 samples:

Table 1: Win rates from head-to-head comparisons on the UltraFeedback test set, judged by GPT-4o-mini. Bold indicates the higher win rate.

Comparison	SFT Wins	SFT+DPO Wins	Ties
SFT vs. SFT+DPO	42.0%	**56.0%**	2.0%
	SFT+CDPO Wins	**SFT+DPO Wins**	**Ties**
SFT+CDPO vs. SFT+DPO	**57.0%**	43.0%	0.0%

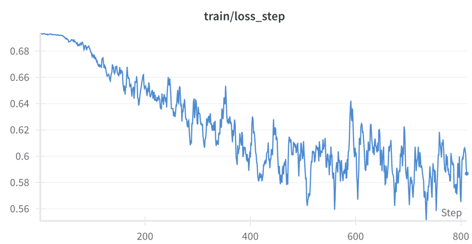
Table 1 summarizes the results of our pairwise model comparisons on the UltraFeedback test set. Each pair was evaluated in a head-to-head setting, where responses were rated on quality using our evaluation prompt (Appendix A).

The comparison between **SFT and SFT+DPO** shows a clear improvement from incorporating preference optimization: SFT+DPO achieves a win rate of **56.0%** compared to only 42.0% for SFT, with 2.0% of cases resulting in a tie. This suggests that even a single stage of preference tuning yields meaningful gains in response quality, likely due to alignment with desirable generation attributes (e.g., relevance, clarity, and completeness).

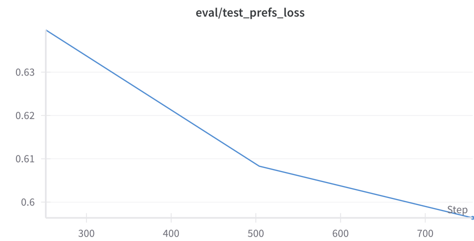
More notably, in the **SFT+CDPO vs. SFT+DPO** comparison, **SFT+CDPO** achieves a win rate of **57.0%** compared to 43.0% for SFT+DPO, with no ties. This indicates that CDPO further improves generation quality beyond standard DPO, despite being trained on the same underlying preference data. These results support our hypothesis that CDPO can help models more effectively learn distinctions in preference signals.

5.2 Training Dynamics

Figure 2 shows loss curves across our train/eval stages:

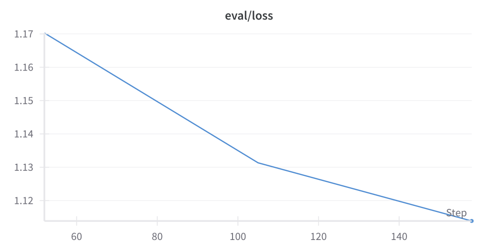


(a) SFT+DPO Train Loss

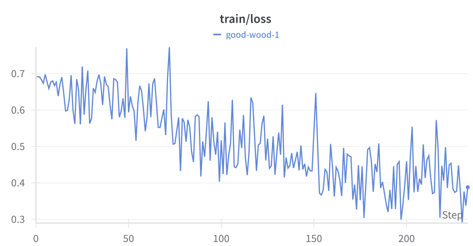


(b) SFT+DPO Eval Loss

(c) SFT Train Loss



(d) SFT Eval Loss



(e) CDPO Train Loss

Figure 2: Training and evaluation loss curves for SFT, DPO, and CDPO.

Loss-curve analysis. Figure 2 shows the training and evaluation loss curves for our three fine-tuning strategies: vanilla SFT (Figs. c–d), SFT + DPO (Figs. a–b), and CDPO (Fig. e).

SFT baseline (Figs. c–d). On the training set, pure SFT slowly reduces its loss from approximately 1.25 to 1.10 over 150 steps, and its evaluation loss falls from 1.17 to 1.12. The relatively shallow slope and high final loss establish our starting point.

SFT + DPO (Figs. a–b). Incorporating direct preference optimization lowers the initial loss (≈ 0.68) and yields a steeper downward trend: training loss reaches ≈ 0.58 after 800 steps, despite higher per-step variance. Its evaluation loss likewise drops from ≈ 0.635 to ≈ 0.598, substantially outperforming the SFT baseline.

CDPO (Fig. e). Curriculum DPO drives the fastest convergence: within just 200 steps the training loss plunges from ≈ 0.70 to ≈ 0.35, with reduced noise toward the end.

In sum, both DPO variants dramatically accelerate and deepen loss minimization compared to standard SFT, with CDPO achieving the steepest descent in far fewer steps. These optimization gains mirror the higher head-to-head win rates we observe in Table1.

5.3 Qualitative Analysis

SFT SFT-only outputs were often verbose or vague, particularly for ambiguous queries. In many cases, responses lacked precision or contained unnecessary filler text.

SFT + DPO DPO fine-tuning made the model more concise and opinionated. It began prioritizing clearer arguments and direct responses. However, in complex prompts, hallucinations and overconfidence in incorrect answers occasionally appeared, likely due to noisy preference signals in the training set.

SFT + Curriculum Learning (CDPO) CDPO yielded responses that were not only concise but also better aligned with user intent. Manual inspection shows fewer hallucinations and more coherent answers.

Comparison: SFT vs SFT+DPO Below is a representative example from the UltraFeedback test set where **SFT+DPO outperforms SFT**, as judged by GPT-4o-mini.

Prompt:

```
If "A man is reading a book about ancient history." does that mean that "A man is
    studying modern science."?
Options:
- yes
- it is not possible to tell
- no
Step by step answer:
```

— SFT Response —

```
A man can't be studying science and ancient history at the same time. The
    information in both sentences contradicts each other. Therefore, the answer is
    no.
```

— DPO Response (Winner) —

```
A man can read books about ancient history without studying modern science. The
    context of the sentences doesn't contradict or provide enough information to
    infer they are different subjects.

So, the correct answer is: it is not possible to tell.
```

Comparison: SFT+CDPO vs SFT+DPO Below is an example from the UltraFeedback test set where **SFT+CDPO outperforms SFT+DPO**, as judged by GPT-4o-mini.

Prompt:

```
How can virtual reality technology be integrated with psychological therapies to
    create accessible, self-managed interventions for individuals experiencing
    chronic stress or anxiety?
```

— SFT+DPO Response —

```
Virtual Reality (VR) technology has the potential to revolutionize how we approach
    and treat chronic stress and anxiety. By creating immersive environments that
    simulate real-world scenarios, VR can help individuals manage their symptoms
    in a safe and controlled setting.

One way VR can be integrated is through gamification, which can make therapy more
    engaging and enjoyable. For example, therapists could use VR games to teach
    mindfulness techniques, encourage relaxation exercises, or provide
    opportunities for social interaction. This would allow individuals to practice
    these skills in an environment they can control, reducing feelings of
    isolation and overwhelm.

Another approach is to use VR as a therapeutic tool for cognitive-behavioral
    therapy (CBT), where individuals can explore and confront their negative
    thoughts and behaviors through immersive experiences. For instance, therapists
    might use VR to simulate different situations, such as a high-stress work
```

```
    scenario, and then guide patients through CBT exercises designed to challenge
    and change those thought patterns.

Additionally, VR could be used to promote physical activity, helping individuals
    overcome mental fatigue and improve overall well-being. Patients could wear VR
    headsets during physical activities to enhance engagement and motivation. This
    could include workouts, yoga, or even adventure-based VR experiences.

Furthermore, VR could be integrated into treatment plans by providing access to
    personalized coaching sessions, allowing individuals to receive guidance from
    trained therapists who can adjust treatment strategies based on individual
    needs and progress.

To ensure accessibility and usability, VR platforms should prioritize inclusivity
    and ease of use for all users. This includes features like voice commands,
    natural language input, and easy navigation across multiple devices.

Overall, integrating VR technology into psychological therapies offers numerous
    benefits for individuals struggling with chronic stress and anxiety. It
    creates a safe and controlled space for exploration and growth, while also
    promoting practical applications for mental health treatment. As research
    continues to inform VR development, it will likely continue to evolve and
    expand its reach in the realm of psychological therapies.
```

— SFT+CDPO Response (Winner) —

```
Virtual reality (VR) technology has the potential to revolutionize how we approach
    and treat chronic stress and anxiety by providing a safe and immersive
    environment where individuals can practice mindfulness, relaxation techniques,
    and coping strategies in a controlled setting. Here's how VR can be integrated
    into psychological therapies:

1. **Mindfulness-based interventions**: VR meditation sessions can simulate
    mindfulness practices, allowing clients to focus on their breath, thoughts,
    and sensations without distractions.
2. **Relaxation training**: Clients can experience guided relaxation exercises,
    such as progressive muscle relaxation or visualization, using VR headsets that
    immerse them in calming environments like forests or oceans.
3. **Cognitive-behavioral therapy (CBT)**: VR-based CBT can help individuals
    identify and challenge negative thought patterns related to stress and
    anxiety, fostering cognitive restructuring and emotional regulation.
4. **Sleep disturbances**: VR sleep therapy simulates realistic sleep scenarios,
    helping individuals develop better sleep habits and reduce insomnia symptoms
    through personalized guidance.
5. **Anxiety reduction**: VR exposure therapy can provide individuals with repeated
    opportunities to confront feared situations, reducing anxiety levels over time.
6. **Self-management**: By practicing these interventions in a controlled
    environment, clients can become more confident and adaptable in managing their
    own mental health challenges.

To achieve this, therapists can employ various VR technologies, including:

* **Augmented Reality (AR) and Virtual Reality (VR)**: These technologies allow
    clients to interact with digital environments, creating immersive experiences
    that mimic real-world interactions.
* **Real-time feedback and analytics**: VR platforms can provide real-time feedback
    on client progress, enabling therapists to refine treatment plans and adjust
    interventions as needed.
* **Customizable modules**: Therapists can design customized modules, tailored to
    individual needs and preferences, incorporating specific VR elements, such as
    music or ambient sounds, to enhance therapeutic outcomes.

Some examples of successful VR-based interventions include:
```

```
* A study published in the Journal of Anxiety Disorders found that VR-assisted
     exposure therapy was effective in reducing symptoms of generalized anxiety
     disorder in adults.
* Another study demonstrated that VR-induced relaxation helped alleviate symptoms
     of post-traumatic stress disorder (PTSD).
* A research paper published in the International Journal of Behavioral Medicine
     highlights the benefits of VR-based mindfulness meditation programs for
     improving attention and well-being.

By combining VR technology with traditional psychological therapies, clinicians can
     create more accessible, evidence-based interventions that promote long-term
     healing and resilience in individuals struggling with chronic stress and
     anxiety. However, it's essential to consider factors like accessibility, cost,
     and scalability when implementing VR-based interventions into clinical
     settings. Nevertheless, the potential benefits far outweigh any potential
     drawbacks.
```

6 Discussion

Limitations Our biggest limitation was time and compute. After the new leaderboard was released, we regenerated responses using our SFT and SFT+DPO checkpoints. Despite performing very well on the original leaderboard, we realized we did not meet the performance needed on the new leaderboard. We decided this was most likely due to SFT being only trained on 10% of the training data, and DPO only being trained with a 512 token max response length. After the poster we went and retrained via SFT on the entire dataset, and retrained DPO and CDPO with longer responses. However, due to a lack of time, we were only able to train CDPO on 10% of the training data.

Curriculum Learning Why does curriculum help? We infer that it helps because of (i) *Gradient signal quality:* early exposure to high-margin pairs yields clearer gradients, preventing premature over-confidence; (ii) *Implicit regularization:* gradual difficulty acts like a continuation method, smoothing the loss landscape for small-capacity models.

7 Conclusion

Curriculum Learning Curriculum scheduling is a lightweight yet powerful tool for aligning small LLMs. Our proposed method improves training stability, accelerates convergence, and achieves higher win rates than vanilla DPO—without any increase in data and only a small additional one-time cost for computing difficulties. These results demonstrate that even under tight resource constraints, strategic data ordering can give you meaningful performance gains.

General Takeaways Our findings support two key lessons: (1) **Stability matters**—doubling batch size and applying curriculum learning both reduce loss variance and lower training noise; (2) **Data ordering matters**—gradually introducing difficult examples enhances learning dynamics.

Future Directions CDPO opens several promising future directions. First, testing adaptive schedules that dynamically re-rank data as the policy evolves. Second, combining the log probabilities with heuristics like edit distance and response length to develop a more comprehensive formula for difficulty. Finally, exploring multi-dimensional curricula that combine prompt complexity with response separability, building on recent advances such as 2D-Curri-DPO.

8 Team Contributions

- **Adam Chun:** Implemented SFT and data and evaluation pipelines, did most of the training because he had the best GPU, always managed to get our loss to go down when things looked dire
- **Tom Nguyen:** Led the implementation of DPO, co-implemented CDPO, ran model evaluation and analysis

- **Josh Francis:** Led the implementation of CDPO, co-implemented DPO, handled leaderboard submissions

Changes from Proposal We originally proposed synthetic-data augmentation for RLHF. We pivoted to curriculum learning after the milestone, because we thought it might offer better performance benefits at a lower compute cost.

References

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:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. arXiv:2309.00267 [cs.CL] https://arxiv.org/abs/2309.00267

Mengyang Li and Zhong Zhang. 2025. 2D-Curri-DPO: Two-Dimensional Curriculum Learning for Direct Preference Optimization. *arXiv preprint arXiv:2504.07856* (2025).

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. 2020. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research* 21, 181 (2020), 1–50.

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

Pulkit Pattnaik, Rishabh Maheshwary, Kelechi Ogueji, Vikas Yadav, and Sathwik Tejaswi Madhusudhan. 2024. Curry-DPO: Enhancing Alignment using Curriculum Learning & Ranked Preferences. In *Findings of EMNLP*.

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

Yuntao Wang, Yuzhang Yao, Bill Lin, Xuefei Li, Shangbin Liu, Ziyang Wang, Yilun Xu, Yizhong Ge, Yi Zhu, Yanwei Zhang, Xiang Lin, Xiangru Li, Yuxuan Zhou, Tian Zhou, Hanjie Ou, Uri Alon, Zihao Liu, Cheng-Yang Hsieh, Jiannan Li, Junxian He, Keli Zhang, Tushar Khot, Rui Wang, Qian Liu, Yuwei Zhang, Caiming Xiong, Graham Neubig, Noah A. Smith, Xiao Lin, and Daniel Khashabi. 2023. UltraFeedback: Boosting Language Models with High-Quality Feedback. https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized. Accessed 2025-06-09.

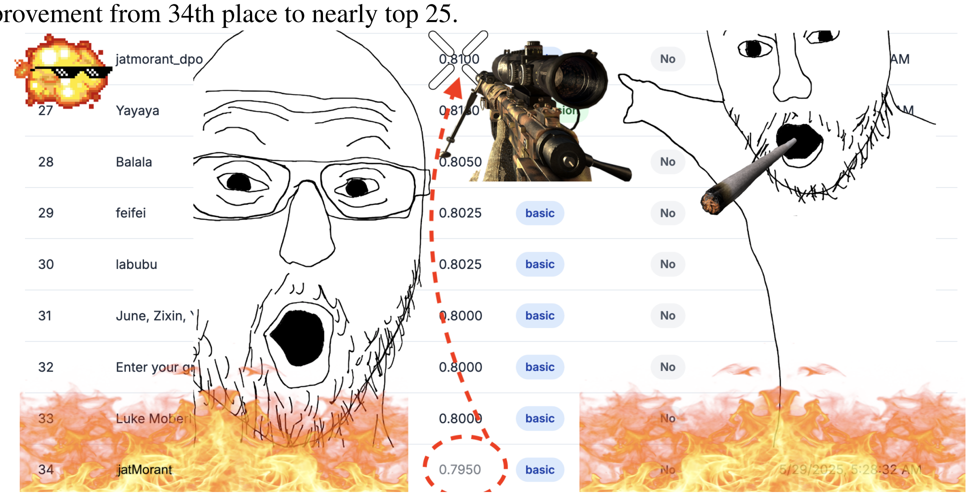
Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2024. α-DPO: Adaptive Reward Margin is What Direct Preference Optimization Needs. arXiv:2410.10148 [cs.LG] https://arxiv.org/abs/2410.10148

A GPT-4o-mini Prompt for Win Rate Evaluation

```
evaluation_prompt = f"""
You are an expert evaluator. Compare the following two responses to the given prompt and determin

**Prompt:** {prompt}

**Response A:** {response_a}

**Response B:** {response_b}

**Evaluation Criteria:**
- Helpfulness and relevance to the prompt
- Accuracy and factual correctness
- Clarity and coherence
- Completeness of the answer
- Overall quality

**Instructions:**
- Respond with exactly one word: "A" if Response A is better, "B" if Response B is better, or "ti
- Consider the evaluation criteria above
- Be objective and fair in your assessment

**Your evaluation:**"""
```

B Model Performance on Initial Leaderboard

Table 2: Instruction following milestone leaderboard results. Implementing DPO led to a significant improvement from 34th place to nearly top 25.



	jatmorant_dpo	0.8100		No	AM
27	Yayaya				
28	Balala	0.8050		No	
29	feifei	0.8025	basic	No	
30	labubu	0.8025	basic	No	
31	June, Zixin,	0.8000	basic	No	
32	Enter your g	0.8000	basic	No	
33	Luke Mober	0.8000	basic	No	
34	jatMorant	0.7950	basic	No	5/29/2025, 5:28:32 AM