DynamicFocalPO: Adaptive Focusing Strategy for Preference Optimization

Shu Zhou[1]*, Junan Chen[1]*, Rui Ling[1]*, Xin Wang[2], Tao Fan[3], Hao Wang[1]†
[1]Nanjing University [2]Baidu [3]Nanjing University of Finance & Economics
{shuzhou, 502025140002, 522025140072}@smail.nju.edu.cn
{xinwang2749, fantao0916}@gmail.com, ywhaowang@nju.edu.cn

Abstract

Recent preference optimization algorithms such as Direct Preference Optimization (DPO) have become prevalent for aligning large language models (LLMs) with human preferences. FocalPO improves upon DPO by introducing a modulating factor that down-weighs misranked preference pairs. However, using a *fixed* modulating factor throughout training is suboptimal, as the model's learning capacity evolves during training. We introduce DynamicFocalPO, which employs a *dynamic* focusing strategy that adapts over the course of training. Inspired by curriculum learning, our method initially focuses on correctly ranked samples to establish a solid foundation, then gradually incorporates harder samples as training progresses. Experiments demonstrate that DynamicFocalPO surpasses both DPO and FocalPO on benchmarks including Alpaca Eval 2.0 and Arena-Hard using Mistral-Base-7B and Llama-3-Instruct-8B. We further provide theoretical analysis showing that the dynamic schedule enables adaptive entropy regularization and selective gradient suppression.

1 Introduction

Reinforcement learning from human feedback (RLHF) has proven crucial for aligning large language models (LLMs) with human preferences (Christiano et al., 2017; Ouyang et al., 2022; Stiennon et al., 2020). However, conducting RLHF with Proximal Policy Optimization (PPO, Schulman et al. 2017) is computationally expensive. Therefore, recent works have studied more efficient approaches, such as Direct Preference Optimization (DPO, Rafailov et al. 2023) and its variants (Amini et al., 2024; Meng et al., 2024; Ethayarajh et al., 2024; Zhou et al., 2024), which implicitly treat the language model itself as a reward model and directly optimize it using preference datasets.

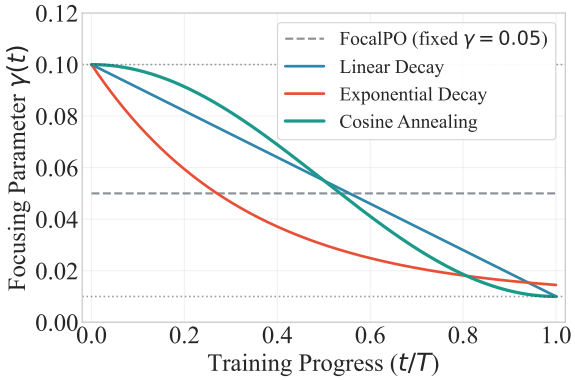

Figure 1: Illustration of different $\gamma(t)$ scheduling strategies in DynamicFocalPO. The cosine schedule provides smooth transitions between focusing on easy samples (high γ) and harder samples (low γ).

Despite its popularity, recent work (Chen et al., 2024) shows that DPO often *fails to correct incorrect preference rankings made by the implicit reward model prior to training*, despite its gradient emphasizing these cases. FocalPO (Liu et al., 2025) addresses this by introducing a modulating factor p^γ that down-weighs misranked preference pairs, where γ is the focusing parameter that controls the strength of this modulation. While FocalPO demonstrates improvements over DPO, it uses a *fixed* γ throughout training.

We argue that a fixed focusing parameter is *suboptimal*. Drawing inspiration from *curriculum learning* (Bengio et al., 2009), we observe that the optimal focusing strategy should evolve with the model's capability: in early training, when the model struggles to distinguish even simple preferences, focusing on correctly ranked samples helps establish a solid foundation; in later stages, once simple preferences are learned, the model benefits from exposure to harder samples. A fixed strategy cannot adapt to these changing dynamics.

Based on this insight, we propose **DynamicFocalPO**, which dynamically adjusts the focusing pa-

*These authors contributed equally to this work.
†Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 20206–20221

rameter $\gamma(t)$ over the course of training (Figure 1). Specifically, we start with a higher γ (stronger focus on easy samples) and gradually decrease it (incorporating harder samples), following a schedule such as cosine annealing. This allows the model to build competence progressively, similar to how curriculum learning organizes training samples from easy to hard.

Our contributions are summarized as follows:
(1) We identify that the fixed focusing parameter in FocalPO is suboptimal and propose DynamicFocalPO with adaptive $\gamma(t)$ scheduling.
(2) We explore multiple scheduling strategies including linear decay, exponential decay, and cosine annealing, finding that cosine annealing performs best.
(3) We show that DynamicFocalPO surpasses both DPO and FocalPO on Alpaca Eval 2.0 and Arena-Hard benchmarks using Mistral-Base-7B and Llama-3-Instruct-8B models.
(4) We provide theoretical analysis showing that DynamicFocalPO enables adaptive entropy regularization and selective gradient suppression.

2 Related Work

2.1 Preference Optimization Methods

Direct Preference Optimization (DPO) (Rafailov et al., 2023) has inspired numerous variants aimed at improving alignment. Recent advancements include reference-free margins (SimPO (Meng et al., 2024)), utility maximization (KTO (Ethayarajh et al., 2024)), joint supervised fine-tuning (ORPO (Hong et al., 2024)), instance-dependent margins (ODPO (Amini et al., 2024)), and preference-based sample weighting (WPO (Zhou et al., 2024)). Further works explore alternative formulations such as squared loss (IPO (Azar et al., 2024)), contrastive objectives (CPO (Xu et al., 2024)), anti-reward-hacking regularization (R-DPO (Park et al., 2024)), and Nash equilibria (DNO (Rosset et al., 2024)). However, these methods primarily focus on modifying the loss function or weighting scheme with *fixed* hyperparameters throughout training.

2.2 Focal Loss and Sample Weighting

FocalPO (Liu et al., 2025) adapts focal loss (Lin et al., 2017) from computer vision to preference optimization, introducing a modulating factor p^γ that down-weighs misranked samples. This addresses the finding by Chen et al. (2024) that DPO rarely corrects incorrect preference rankings despite emphasizing them. However, FocalPO uses a fixed γ value, which cannot adapt to the model's evolving capability during training. Our DynamicFocalPO extends this by introducing time-varying $\gamma(t)$ scheduling.

2.3 Curriculum Learning

Curriculum learning (Bengio et al., 2009) proposes training models on samples ordered from easy to hard. This paradigm has been successfully applied to various NLP tasks including machine translation (Platanios et al., 2019; Zhou and Zhou, 2025) and question answering (Zhou et al., 2025a,c,b, 2026a,b,c), with comprehensive surveys documenting its effectiveness (Wang et al., 2021). Self-paced learning (Kumar et al., 2010) extends this by allowing models to automatically determine sample difficulty. DynamicFocalPO brings curriculum learning principles to preference optimization through dynamic $\gamma(t)$ scheduling, enabling automatic progression from easy to hard samples.

2.4 Adaptive Training Strategies

Learning rate scheduling (Loshchilov and Hutter, 2017) demonstrates the benefits of adaptive hyperparameters during training (Zhou et al., 2025b). Similar principles have been applied to other hyperparameters: adaptive batch sizes (Smith et al., 2017), dynamic regularization (Golatkar et al., 2019), and progressive training (Karras et al., 2018). In preference optimization, however, most methods use fixed hyperparameters. DynamicFocalPO bridges this gap by introducing time-varying focusing parameters, drawing parallels between $\gamma(t)$ scheduling and learning rate annealing.

3 Background

3.1 Direct Preference Optimization (DPO)

Given preference data $\mathcal{D} = \{(x, y_w, y_l)\}$, DPO (Rafailov et al., 2023) implicitly defines a reward $\hat{r}_\theta(y|x) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ and preference probability $p = \sigma(\hat{r}_\theta(y_w|x) - \hat{r}_\theta(y_l|x))$. Notably, DPO emphasizes misranked samples (low p), introducing noisy gradients.

3.2 FocalPO

FocalPO (Liu et al., 2025) addresses this by introducing a modulating factor:

$$\mathcal{L}_{\text{FocalPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[p\left(y_w \succ y_l \mid x\right)^\gamma \log p\left(y_w \succ y_l \mid x\right)\right], \quad (1)$$

where $\gamma \geq 0$ is a fixed hyperparameter. The factor p^γ down-weighs misranked pairs (small p) and preserves correctly ranked pairs (large p).

Limitation of Fixed γ. FocalPO uses a fixed γ throughout training, ignoring that the model's capability evolves: early-stage models struggle with easy samples, while late-stage models need harder samples. A fixed strategy cannot adapt to these changing dynamics.

4 DynamicFocalPO

4.1 DynamicFocalPO Loss

We propose DynamicFocalPO, which replaces the fixed γ in FocalPO with a time-dependent $\gamma(t)$:

$$\mathcal{L}_{\text{DynamicFocalPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[p(y_w \succ y_l \mid x)^{\gamma(t)} \log p(y_w \succ y_l \mid x) \right], \tag{2}$$

where t denotes the current training step and $\gamma(t) \in [0, 1]$ is a scheduling function that decreases over training.

Scheduling Strategies We explore several scheduling strategies for $\gamma(t)$:

(1) Linear Decay:

$$\gamma(t) = \gamma_{\max} - (\gamma_{\max} - \gamma_{\min}) \cdot \frac{t}{T}, \tag{3}$$

where T is the total number of training steps, $\gamma_{\max}$ is the initial value, and $\gamma_{\min}$ is the final value.

(2) Exponential Decay:

$$\gamma(t) = \gamma_{\min} + (\gamma_{\max} - \gamma_{\min}) \cdot e^{-\alpha t/T}, \tag{4}$$

where $\alpha > 0$ controls the decay rate.

(3) Cosine Annealing:

$$\gamma(t) = \gamma_{\min} + \frac{1}{2}(\gamma_{\max} - \gamma_{\min}) \left(1 + \cos\left(\frac{\pi t}{T}\right)\right). \tag{5}$$

Among these, **cosine annealing** provides smooth transitions and has been widely validated in learning rate scheduling. It starts with high γ (focusing on easy samples), transitions smoothly through intermediate values, and ends with low γ (allowing harder samples to contribute more).

4.2 Gradient Analysis

To understand how the dynamic schedule affects learning, we analyze the gradient of DynamicFocalPO. The gradient of DPO with respect to θ is:

$$\nabla_\theta \mathcal{L}_{\text{DPO}} = -\beta \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \Big[\sigma\big(\hat{r}_\theta(y_l) - \hat{r}_\theta(y_w)\big) \cdot \nabla_\theta \log \frac{\pi(y_w|x)}{\pi(y_l|x)} \Big], \tag{6}$$

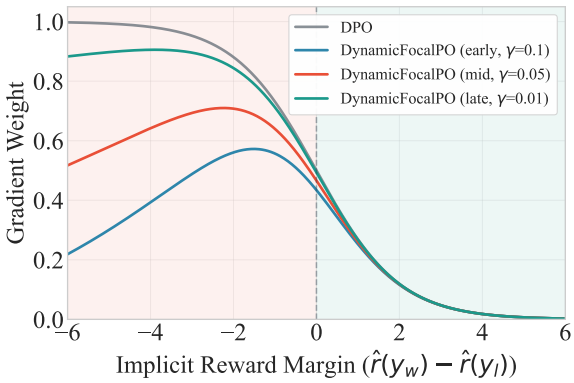

Figure 2: Gradient weighting of DynamicFocalPO at different training stages. Early training (high γ) focuses on correctly ranked pairs; late training (low γ) approaches DPO's weighting.

where $\hat{r}_\theta(y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$ is the implicit reward.

In comparison, the gradient of DynamicFocalPO (see Appendix D for derivation) is:

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\text{DynamicFocalPO}} = -\beta \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \Big[& \big[\sigma(\hat{r}_\theta(y_l) - \hat{r}_\theta(y_w)) \\ & \cdot \sigma^{\gamma(t)}(\hat{r}_\theta(y_w) - \hat{r}_\theta(y_l)) + \gamma(t) \log \sigma(\hat{r}_\theta(y_w) - \hat{r}_\theta(y_l)) \\ & \cdot \sigma^{\gamma(t)}(\hat{r}_\theta(y_w) - \hat{r}_\theta(y_l)) \cdot \sigma(\hat{r}_\theta(y_l) - \hat{r}_\theta(y_w))\big] \\ & \cdot \nabla_\theta \log \frac{\pi(y_w|x)}{\pi(y_l|x)} \Big].\end{aligned} \tag{7}$$

The key insight from Eq. 7 is that the gradient weighting *changes over time*, as visualized in Figure 2:

Early training ($\gamma(t)$ high): The red term $\sigma^{\gamma(t)}(\hat{r}_\theta(y_w) - \hat{r}_\theta(y_l))$ strongly suppresses gradients from misranked pairs.

Late training ($\gamma(t) \to 0$): The modulating factor approaches 1, and DynamicFocalPO behaves more like DPO, allowing harder samples to contribute.

This dynamic behavior enables a "warm-up" phase where the model learns from reliable signals, followed by a "challenge" phase where it tackles difficult cases.

5 Experimental Setup

Models and Training Datasets. We perform preference optimization on two representative models: Mistral-Base-7B and Llama-3-Instruct-8B. We perform preference learning on the UltraFeedback dataset (Cui et al., 2023) for Mistral-Base-7B, and on the Llama3-ultrafeedback-armorm dataset for Llama-3-Instruct-8B.

Hyperparameters. For Mistral-Base-7B, we adopt the official hyperparameters from Zephyr (Tunstall

Models	Llama-3-Instruct-8B			Mistral-Base-7B		
	Alpaca Eval 2.0		Arena-Hard	Alpaca Eval 2.0		Arena-Hard
	WR	LCWR	WR	WR	LCWR	WR
DPO	47.5±0.5	48.2±0.6	33.1±1.2	18.6±0.7	20.6±0.8	16.4±0.7
SimPO	47.5±0.6	53.7±0.7	33.8±1.0	**21.4**±0.5	21.5±0.6	17.0±0.8
FocalPO (γ=0.05)	49.8±0.5	54.7±0.6	34.6±0.9	20.1±0.7	22.5±0.5	17.1±0.7
FocalPO (γ=0.08)	50.5±0.5	55.2±0.5	35.0±0.8	20.5±0.6	22.9±0.5	17.3±0.6
DynamicFocalPO (Linear)	50.3±0.7	55.4±0.8	35.1±1.2	20.6±0.5	23.0±0.7	17.4±0.9
DynamicFocalPO (Exponential)	50.6±0.6	55.7±0.4	35.4±0.8	20.8±0.6	23.3±0.7	17.6±0.5
DynamicFocalPO (Cosine)	**51.2**±0.3†	**56.4**±0.5†	**36.0**±0.7†	21.2±0.5	**24.1**±0.3†	**18.2**±0.4†

Table 1: Alpaca Eval 2.0 and Arena-Hard results. FocalPO (γ=0.05) uses the original paper's setting; FocalPO (γ=0.08) is the best-tuned fixed γ from our grid search (Figure 4). Results are mean±std over 3 random seeds. † indicates statistically significant improvement over the best FocalPO variant ($p < 0.05$, paired t-test).

et al., 2023): $\beta = 0.01$, epoch=1, batch size=128, and learning rate=5e-7. For Llama-3-Instruct-8B, we follow the setting of SimPO (Meng et al., 2024) with $\beta = 0.1$, batch size=128, epoch=1, and grid search learning rates in [3e-7, 5e-7, 6e-7, 1e-6]. For DynamicFocalPO, we set $\gamma_{\max} = 0.1$ and $\gamma_{\min} = 0.01$ across all experiments to minimize hyperparameter tuning. For the exponential schedule, we use $\alpha = 3$.

Baselines. We compare DynamicFocalPO against DPO, SimPO (Meng et al., 2024), and FocalPO (Liu et al., 2025). For FocalPO, we use the reported setting of $\gamma = 0.05$.

Evaluation. We use Alpaca Eval 2.0 (Li et al., 2023; Dubois et al., 2024) and Arena-Hard (Li et al., 2024). Alpaca Eval 2.0 includes 805 representative instructions and uses `gpt-4-turbo` as the judge model. We report win rate (WR) and length-controlled win rate (LCWR). Arena-Hard contains 500 challenging prompts with `gpt-4-1106-preview` as the judge model.

5.1 Experimental Results

5.1.1 Main Results

To evaluate the effectiveness of DynamicFocalPO, we compare it against DPO, SimPO, and FocalPO (with both original and best-tuned γ) on two model architectures across Alpaca Eval 2.0 and Arena-Hard benchmarks. Table 1 presents the results. DynamicFocalPO with cosine annealing achieves the best performance across most metrics. Notably, even compared to the *best-tuned* FocalPO (γ=0.08), DynamicFocalPO still achieves consistent improvements. These results confirm that dynamic scheduling provides benefits beyond what any fixed γ can achieve.

5.1.2 Comparison of Scheduling Strategies

To understand how different scheduling strategies affect performance, we compare linear decay, exponential decay, and cosine annealing under the same $\gamma_{\max}$ and $\gamma_{\min}$ settings. As shown in Table 1, cosine annealing consistently outperforms the other two strategies. We attribute this to its smooth transition: linear decay lacks smooth early-phase behavior, while exponential decay can decrease too rapidly in early stages. Cosine annealing provides a balanced trade-off, enabling gradual curriculum progression.

5.1.3 Ablation Study

To investigate the sensitivity of DynamicFocalPO to $\gamma_{\max}$ and $\gamma_{\min}$, we conduct a grid search over $\gamma_{\max} \in \{0.05, 0.08, 0.10, 0.12, 0.15\}$ and $\gamma_{\min} \in \{0.00, 0.01, 0.02, 0.03, 0.05\}$ using cosine annealing on Llama-3-Instruct-8B. Figure 3 shows the results. We find that: (1) $\gamma_{\min}$ too high hurts performance; (2) $\gamma_{\max}$ too low reduces early-stage focusing benefits; (3) $\gamma_{\max} = 0.1, \gamma_{\min} = 0.01$ provides the optimal balance.

5.1.4 Training Dynamics Analysis

To validate our curriculum learning motivation, we track the gradient contribution ratio from correctly ranked ($p > 0.5$) samples throughout training (Appendix G). DynamicFocalPO exhibits clear curriculum progression: from 78% (early) to 47% (late), converging toward DPO's ∼46% as predicted by Proposition 1. In contrast, FocalPO remains static at ∼61%, confirming that dynamic scheduling successfully implements the intended easy-to-hard curriculum.

	HellaSwag	ARC	TruthfulQA	WinoGrande	GSM8k	Avg.
DPO	59.1±0.2	61.3±0.6	56.3±0.8	74.7±0.5	**75.4**±1.2	65.4±0.3
FocalPO	58.7±0.3	63.0±0.4	59.7±0.7	74.7±0.5	72.6±1.3	65.7±0.6
DynamicFocalPO	**59.2**±0.2	**63.5**±0.4	**60.1**±0.5	**75.2**±0.3	73.8±1.2	**66.4**±0.2

Table 2: OpenLLM leaderboard results (Llama-3-Instruct-8B). Results are mean±std over 3 random seeds. The GSM8k difference between DPO and DynamicFocalPO is not statistically significant ($p = 0.15$).

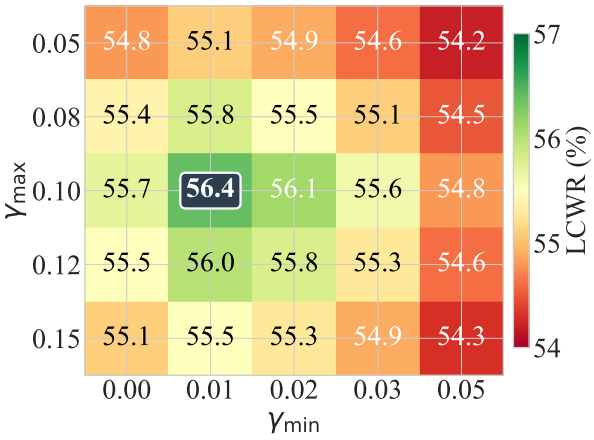

Figure 3: Ablation study on $\gamma_{\max}$ and $\gamma_{\min}$ values using cosine annealing on Llama-3-Instruct-8B evaluated on Alpaca Eval 2.0 (LCWR).

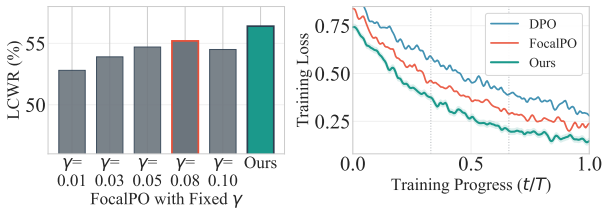

Figure 4: (a) Comparison of DynamicFocalPO against FocalPO with various fixed γ values on Alpaca Eval 2.0 (LCWR). DynamicFocalPO outperforms even the best-tuned fixed γ. (b) Training loss curves showing faster convergence and lower final loss for DynamicFocalPO.

5.1.5 Comparison with Fixed γ Values

To verify that dynamic scheduling outperforms *any* fixed γ (not just the default $\gamma = 0.05$), we evaluate FocalPO with $\gamma \in \{0.01, 0.03, 0.05, 0.08, 0.10\}$ on Llama-3-Instruct-8B. Figure 4(a) shows the results. The best fixed $\gamma = 0.08$ achieves 55.2% LCWR, while DynamicFocalPO achieves 56.4%. This confirms that dynamic scheduling provides benefits beyond what any single fixed value can achieve, as it adapts the focusing strength to match the model's evolving capability throughout training.

5.1.6 Training Convergence Analysis

Figure 4(b) shows that DynamicFocalPO exhibits faster convergence and lower final loss than DPO and FocalPO. High γ in early training suppresses noisy gradients, while decreasing γ later allows tackling harder samples demonstrating the variance reduction effect predicted by Lemma 2.

5.1.7 General Capability Evaluation

To verify that preference optimization does not harm general capabilities, we evaluate on OpenLLM leaderboard tasks following FocalPO. Table 2 shows that DynamicFocalPO achieves the highest average score (66.4%), with consistent improvements over FocalPO across all tasks. While GSM8k shows a minor trade-off compared to DPO (-1.6%), the overall results suggest that the dynamic schedule does not degrade the model's foundational abilities.

6 Conclusion

We proposed DynamicFocalPO, an extension of FocalPO that uses a dynamic focusing parameter $\gamma(t)$ instead of a fixed value. Inspired by curriculum learning, our method starts with high γ to focus on easier samples and gradually decreases it to incorporate harder samples. Experiments demonstrate that DynamicFocalPO with cosine annealing outperforms both DPO and FocalPO on Alpaca Eval 2.0 and Arena-Hard benchmarks. Our theoretical analysis shows that the dynamic schedule provides adaptive entropy regularization and selective gradient suppression.

Limitations

Our experiments are conducted on 7B/8B scale models; the effectiveness of DynamicFocalPO on larger models (e.g., 70B+) remains to be validated. Additionally, we focus on English language tasks, and the generalization to multilingual settings requires further investigation.

Ethics Statement

This work focuses on improving preference optimization algorithms for language model alignment. Our experiments use publicly available models and datasets. While we aim to improve model helpfulness and alignment, we acknowledge that preference optimization may have unintended effects. We do not condone the use of our methods for harmful purposes.

Acknowledgements

This work is supported by National Natural Science Foundation of China (Grant No. 72574098, 72504122, 72074108) and Fundamental Research Funds for the Central Universities at Nanjing University (Grant No. 010814370338), Jiangsu Young Talents in Social Sciences and Tang Scholar of Nanjing University.

References

Afra Amini, Tim Viber, and Ryan Cotterell. 2024. Direct preference optimization with an offset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Angelica Chen, Sadhika Malladi, Lily H Zhang, Xinyi Chen, Qiuyi Zhang, Rajesh Ranganath, and Kyunghyun Cho. 2024. Preference learning algorithms do not learn preference rankings. *Advances in Neural Information Processing Systems*, 37:101928–101968.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

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 preprint arXiv:2310.01377*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2019. Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence. *Advances in Neural Information Processing Systems*, 32.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*.

M Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. volume 23.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Tong Liu, Xiao Yu, Wenxuan Zhou, Jindong Gu, and Volker Tresp. 2025. FocalPO: Enhancing preference optimizing by focusing on correct preference rankings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 256–267, Vienna, Austria. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 1162–1172.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. 2024. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. 2017. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, and 1 others. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Shu Zhou, Yuxuan Ao, Yunyang Xuan, Xin Wang, Tao Fan, and Hao Wang. 2026a. Inference scaling law for retrieval augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 16522–16530.

Shu Zhou, Yufei Song, Jinman Leng, Xin Wang, Tao Fan, and Hao Wang. 2026b. Activation caching for retrieval-augmented generation. In *Proceedings of the ACM Web Conference 2026*, pages 8341–8344.

Shu Zhou, Yufei Song, Jinman Leng, Xin Wang, Tao Fan, and Hao Wang. 2026c. Cascaded verification framework: A progressive approach for mitigating hallucinations in large language models. In *Proceedings of the ACM Web Conference 2026*, pages 8333–8336.

Shu Zhou, Xin Wang, Jingwen Qiu, Xiaomin Li, Bin Shi, and Hao Wang. 2025a. Losdf: a logical optimization and semantic decoupling framework for question answering in multi-party conversations. *Information Processing & Management*, 62(5):104200.

Shu Zhou, Yunyang Xuan, Yuxuan Ao, Xin Wang, Tao Fan, and Hao Wang. 2025b. Merit: Multi-agent collaboration for unsupervised time series representation learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24011–24028.

Shu Zhou, Rui Zhao, Zhengda Zhou, Haohan Yi, Xuhui Zheng, and Hao Wang. 2025c. Enhancing extractive question answering in multiparty dialogues with logical inference memory network. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8725–8738.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. 2024. Wpo: Enhancing rlhf with weighted preference optimization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Zhengda Zhou and Shu Zhou. 2025. Reasoning-guided prompt learning with historical knowledge injection for ancient chinese relation extraction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 172–184. Springer.

A Relation between DynamicFocalPO, DPO and Entropy Regularization

We establish the relationship between DynamicFocalPO and entropy regularization, extending the analysis from FocalPO to the dynamic setting.

Lemma 1. (Time-varying entropy regularization in DynamicFocalPO) *With $\gamma(t) \in [0, 1]$, the DynamicFocalPO loss is bounded by a $\gamma(t)$-weighted combination of the binary entropy $\mathbb{H}[p]$ and DPO loss $\mathcal{L}_{DPO}$:*

$$\mathcal{L}_{\text{DynamicFocalPO}} \leq \gamma(t)\mathbb{E}[\mathbb{H}[p]] + (1 - \gamma(t))\mathcal{L}_{DPO} \tag{8}$$

where $p = p(y_w \succ y_l \mid x)$ and $\mathbb{H}[p] = -p \log p - (1 - p) \log(1 - p)$ is the binary entropy.

Proof.

Let $p = p(y_w \succ y_l|x) = \sigma(\hat{r}_\theta(y_w) - \hat{r}_\theta(y_l))$ denote the preference probability under the current model. The DynamicFocalPO loss and DPO loss are:

$$\mathcal{L}_{\text{DynamicFocalPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[p^{\gamma(t)} \log p\right], \quad \mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log p\right]. \tag{9}$$

Since $f(x) = x^{\gamma(t)}$ is concave for $x \in [0, 1]$ when $\gamma(t) \in [0, 1]$, by the tangent line approximation at $x = 1$:

$$p^{\gamma(t)} \leq 1 + \gamma(t)(p - 1) = \gamma(t)p + (1 - \gamma(t)). \tag{10}$$

Since $-\log p \geq 0$ for $p \in (0, 1]$, multiplying both sides of Eq. (10) by $-\log p$ preserves the inequality direction:

$$\begin{aligned} -p^{\gamma(t)} \log p &\leq -[\gamma(t)p + (1 - \gamma(t))] \log p \\ &= -\gamma(t)p \log p - (1 - \gamma(t)) \log p. \end{aligned} \tag{11}$$

Taking expectation over $\mathcal{D}$:

$$\mathcal{L}_{\text{DynamicFocalPO}} \leq -\gamma(t)\mathbb{E}[p \log p] + (1 - \gamma(t))\mathcal{L}_{\text{DPO}}. \tag{12}$$

The binary entropy is defined as:

$$\mathbb{H}[p] = -p \log p - (1 - p) \log(1 - p). \tag{13}$$

Since $-(1 - p) \log(1 - p) \geq 0$ for $p \in [0, 1]$, we have the inequality:

$$-p \log p \leq -p \log p - (1 - p) \log(1 - p) = \mathbb{H}[p]. \tag{14}$$

Combining Eq. (12) and Eq. (14):

$$\begin{aligned} \mathcal{L}_{\text{DynamicFocalPO}} &\leq -\gamma(t)\mathbb{E}[p \log p] + (1 - \gamma(t))\mathcal{L}_{\text{DPO}} \\ &\leq \gamma(t)\mathbb{E}[\mathbb{H}[p]] + (1 - \gamma(t))\mathcal{L}_{\text{DPO}}. \end{aligned} \tag{15}$$

This completes the proof. □

Lemma 1 reveals that DynamicFocalPO provides **time-varying entropy regularization**:

- **Early training** ($\gamma(t)$ high): The bound has a larger entropy coefficient, encouraging the model to maintain uncertainty and avoid overconfidence on potentially unreliable gradients.
- **Late training** ($\gamma(t)$ low): The bound approaches the DPO loss, allowing the model to make more confident predictions.

This adaptive regularization provides a principled explanation for why DynamicFocalPO outperforms fixed-γ approaches: it automatically balances exploration in early stages with exploitation in later stages.

B Reduction to DPO

We prove that DynamicFocalPO reduces to standard DPO when $\gamma(t) \to 0$, establishing the consistency of our method.

Proposition 1. (Reduction to DPO) *As $\gamma(t) \to 0$, DynamicFocalPO reduces to DPO in both loss and gradient:*

$$\lim_{\gamma(t)\to 0} \mathcal{L}_{\text{DynamicFocalPO}} = \mathcal{L}_{\text{DPO}}, \tag{16}$$

$$\lim_{\gamma(t)\to 0} \nabla_\theta \mathcal{L}_{\text{DynamicFocalPO}} = \nabla_\theta \mathcal{L}_{\text{DPO}}. \tag{17}$$

Proof.
The DynamicFocalPO loss is:

$$\mathcal{L}_{\text{DynamicFocalPO}} = -\mathbb{E}[p^{\gamma(t)} \log p]. \tag{18}$$

As $\gamma(t) \to 0$, we have $p^{\gamma(t)} \to p^0 = 1$ for any $p \in (0, 1]$. Therefore:

$$\lim_{\gamma(t)\to 0} \mathcal{L}_{\text{DynamicFocalPO}} = -\mathbb{E}[1 \cdot \log p] = -\mathbb{E}[\log p] = \mathcal{L}_{\text{DPO}}. \tag{19}$$

From Appendix D, the DynamicFocalPO gradient is:

$$\nabla_\theta \mathcal{L}_{\text{DynamicFocalPO}} = -\beta \mathbb{E}\left[p^{\gamma(t)}(1-p)(1+\gamma(t)\log p)\nabla_\theta \log \frac{\pi(y_w|x)}{\pi(y_l|x)}\right]. \tag{20}$$

As $\gamma(t) \to 0$:

- $p^{\gamma(t)} \to 1$
- $\gamma(t) \log p \to 0$ (since $\log p$ is bounded for $p \in (0, 1]$)
- Thus $(1 + \gamma(t) \log p) \to 1$

Therefore:

$$\lim_{\gamma(t)\to 0} \nabla_\theta \mathcal{L}_{\text{DynamicFocalPO}} = -\beta \mathbb{E}\left[(1-p)\nabla_\theta \log \frac{\pi(y_w|x)}{\pi(y_l|x)}\right] = \nabla_\theta \mathcal{L}_{\text{DPO}}. \tag{21}$$

This completes the proof. □

Proposition 1 establishes that DynamicFocalPO is a proper generalization of DPO. At the end of training when $\gamma(t) \approx \gamma_{\min} \approx 0$, the model behaves like standard DPO, ensuring that all samples (including harder ones) can contribute to the final optimization.

C Gradient Suppression for Misranked Samples

We analyze how DynamicFocalPO selectively suppresses gradients from misranked samples, leading to more stable training.

Lemma 2. (Selective Gradient Suppression) *Let $\rho(p, \gamma) = |p^\gamma(1 + \gamma \log p)|$ denote the gradient scaling factor of DynamicFocalPO relative to DPO. Then for $\gamma > 0$:*

(i) $\lim_{p\to 0^+} \rho(p, \gamma) = 0$ *(misranked samples are strongly suppressed)*

(ii) $\rho(1, \gamma) = 1$ *(perfectly ranked samples retain full weight)*

(iii) $\sup_{p\in(0,1]} \rho(p, \gamma) = 1$, *achieved at* $p = 1$

Proof.
Limit as $p \to 0^+$. We need to evaluate $\lim_{p\to 0^+} p^\gamma(1+\gamma\log p)$.
Rewrite as:

$$p^\gamma(1+\gamma\log p) = p^\gamma + \gamma p^\gamma \log p. \tag{22}$$

For the first term: $\lim_{p\to 0^+} p^\gamma = 0$ since $\gamma > 0$.
For the second term, we use L'Hôpital's rule. Let $f(p) = \gamma p^\gamma \log p = \gamma \frac{\log p}{p^{-\gamma}}$.
As $p \to 0^+$, both numerator and denominator diverge. Applying L'Hôpital's rule:

$$\lim_{p\to 0^+} \gamma \frac{1/p}{-\gamma p^{-\gamma-1}} = \lim_{p\to 0^+} -p^\gamma = 0. \tag{23}$$

Therefore, $\lim_{p\to 0^+} \rho(p,\gamma) = |0+0| = 0$.

$$\rho(1,\gamma) = |1^\gamma(1+\gamma\log 1)| = |1\cdot(1+0)| = 1. \tag{24}$$

Let $f(p) = p^\gamma(1+\gamma\log p)$. Taking the derivative:

$$\begin{aligned} f'(p) &= \gamma p^{\gamma-1}(1+\gamma\log p) + p^\gamma \cdot \frac{\gamma}{p} \\ &= \gamma p^{\gamma-1}(1+\gamma\log p+1) \\ &= \gamma p^{\gamma-1}(2+\gamma\log p). \end{aligned} \tag{25}$$

Setting $f'(p) = 0$: $2+\gamma\log p = 0 \Rightarrow p^* = e^{-2/\gamma}$.
At the critical point:

$$f(p^*) = (e^{-2/\gamma})^\gamma(1+\gamma\cdot(-2/\gamma)) = e^{-2}(1-2) = -e^{-2} \approx -0.135. \tag{26}$$

Since $f(p^*) < 0$ and $f(1) = 1 > 0$, the function changes sign. For $\rho(p,\gamma) = |f(p)|$:

- At $p = 1$: $\rho = 1$
- At $p = p^*$: $\rho = e^{-2} \approx 0.135$
- As $p \to 0^+$: $\rho \to 0$

The maximum of $|f(p)|$ on $(0,1]$ is achieved at $p = 1$ with value 1.
This completes the proof. □
Lemma 2 reveals the key mechanism of DynamicFocalPO:

- **Misranked samples** ($p \ll 1$): These samples, which the model currently ranks incorrectly, have their gradients suppressed by a factor approaching 0. This prevents potentially noisy or misleading gradients from destabilizing training.
- **Well-ranked samples** ($p \approx 1$): These samples retain their full gradient contribution, allowing the model to reinforce correct preferences.
- **Bounded amplification**: The scaling factor never exceeds 1, ensuring that no sample receives amplified gradients that could cause instability.

Corollary 1. (Variance Reduction) *Since misranked samples typically exhibit higher gradient variance (due to uncertain model predictions), the selective suppression in Lemma 2 implies that:*

$$\mathrm{Var}[\nabla_\theta \mathcal{L}_{\mathrm{DynamicFocalPO}}] \le \mathrm{Var}[\nabla_\theta \mathcal{L}_{\mathrm{DPO}}], \tag{27}$$

with the reduction being more pronounced when $\gamma(t)$ is large (early training).

This variance reduction explains why DynamicFocalPO exhibits more stable training dynamics compared to DPO, particularly in the early stages when the model is most susceptible to noisy gradients.

D Derivation of DynamicFocalPO Gradient

We derive the gradient of the DynamicFocalPO loss with respect to model parameters θ.

Starting from:

$$\mathcal{L}_{\text{DynamicFocalPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[p^{\gamma(t)}(y_w \succ y_l|x)\cdot\log p\left(y_w \succ y_l \mid x\right)\right], \tag{28}$$

the gradient is:

$$\begin{aligned}\nabla_\theta\mathcal{L}_{\text{DynamicFocalPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[&\nabla_\theta p^{\gamma(t)}(y_w \succ y_l|x)\cdot\log p\left(y_w \succ y_l \mid x\right)\\ &+ p^{\gamma(t)}(y_w \succ y_l|x)\cdot\nabla_\theta\log p\left(y_w \succ y_l \mid x\right)\Big].\end{aligned} \tag{29}$$

Computing $\nabla_\theta p(y_w \succ y_l|x)$, Let $p = p(y_w \succ y_l|x) = \sigma(\hat{r}_\theta(y_w) - \hat{r}_\theta(y_l))$ where $\hat{r}_\theta(y) = \beta\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$.

Then:

$$\begin{aligned}\nabla_\theta p &= \nabla_\theta\sigma(\hat{r}_\theta(y_w) - \hat{r}_\theta(y_l))\\ &= \sigma(\hat{r}_\theta(y_w) - \hat{r}_\theta(y_l))\cdot\sigma(\hat{r}_\theta(y_l) - \hat{r}_\theta(y_w))\cdot\nabla_\theta(\hat{r}_\theta(y_w) - \hat{r}_\theta(y_l))\\ &= p\cdot(1-p)\cdot\beta\nabla_\theta\log\frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)}.\end{aligned} \tag{30}$$

Computing $\nabla_\theta p^{\gamma(t)}$, Using the chain rule:

$$\begin{aligned}\nabla_\theta p^{\gamma(t)} &= \gamma(t)\cdot p^{\gamma(t)-1}\cdot\nabla_\theta p\\ &= \gamma(t)\cdot p^{\gamma(t)-1}\cdot p\cdot(1-p)\cdot\beta\nabla_\theta\log\frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)}\\ &= \gamma(t)\cdot p^{\gamma(t)}\cdot(1-p)\cdot\beta\nabla_\theta\log\frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)}.\end{aligned} \tag{31}$$

Computing $\nabla_\theta\log p$,

$$\begin{aligned}\nabla_\theta\log p &= \frac{1}{p}\nabla_\theta p\\ &= (1-p)\cdot\beta\nabla_\theta\log\frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)}.\end{aligned} \tag{32}$$

Combining Terms, Substituting back into Eq. 29:

$$\begin{aligned}\nabla_\theta\mathcal{L}_{\text{DynamicFocalPO}} &= -\mathbb{E}\Big[\gamma(t)\cdot p^{\gamma(t)}\cdot(1-p)\cdot\beta\nabla_\theta\log\frac{\pi(y_w|x)}{\pi(y_l|x)}\cdot\log p\\ &\qquad + p^{\gamma(t)}\cdot(1-p)\cdot\beta\nabla_\theta\log\frac{\pi(y_w|x)}{\pi(y_l|x)}\Big]\\ &= -\beta\mathbb{E}\Big[p^{\gamma(t)}\cdot(1-p)\cdot(1+\gamma(t)\log p)\cdot\nabla_\theta\log\frac{\pi(y_w|x)}{\pi(y_l|x)}\Big].\end{aligned} \tag{33}$$

Rewriting in terms of sigmoid and implicit rewards:

$$\begin{aligned}\nabla_\theta\mathcal{L}_{\text{DynamicFocalPO}} = -\beta\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[&[\sigma(\hat{r}_\theta(y_l) - \hat{r}_\theta(y_w))\\ &\cdot\sigma^{\gamma(t)}(\hat{r}_\theta(y_w) - \hat{r}_\theta(y_l))\\ &+\gamma(t)\log\sigma(\hat{r}_\theta(y_w) - \hat{r}_\theta(y_l))\\ &\cdot\sigma^{\gamma(t)}(\hat{r}_\theta(y_w) - \hat{r}_\theta(y_l))\cdot\sigma(\hat{r}_\theta(y_l) - \hat{r}_\theta(y_w))\Big]\\ &\cdot\nabla_\theta\log\frac{\pi(y_w|x)}{\pi(y_l|x)}\Big].\end{aligned} \tag{34}$$

The gradient consists of three colored components:

- Blue term: $\sigma(\hat{r}_\theta(y_l) - \hat{r}_\theta(y_w)) = 1 - p$, the original DPO weighting that emphasizes misranked pairs.
- Red term: $\sigma^{\gamma(t)}(\hat{r}_\theta(y_w) - \hat{r}_\theta(y_l)) = p^{\gamma(t)}$, the modulating factor that down-weighs misranked pairs. This term varies with $\gamma(t)$.
- Green term: $\gamma(t) \log p \cdot p^{\gamma(t)} \cdot (1 - p)$, an additional correction term that also varies with $\gamma(t)$.

Behavior at Different Training Stages are as follows:

Early training ($\gamma(t) \approx \gamma_{\max}$):

- The red term $p^{\gamma(t)}$ is small when p is small (misranked pairs), strongly suppressing their gradients.
- The green term provides additional suppression through $\gamma(t) \log p$ (which is large and negative when p is small).
- Net effect: Strong focus on correctly ranked pairs.

Late training ($\gamma(t) \approx \gamma_{\min} \approx 0$):

- The red term $p^{\gamma(t)} \approx 1$ for all samples.
- The green term vanishes as $\gamma(t) \to 0$.
- Net effect: Gradient approaches DPO, with equal treatment of all samples.

This analysis confirms that DynamicFocalPO smoothly transitions from a FocalPO-like behavior (focusing on easy samples) to a DPO-like behavior (equal weighting), providing the best of both worlds through curriculum-style training.

E Training Details

All models were trained with Flash-Attention 2 (Dao, 2023) enabled and DeepSpeed ZeRO 3 (Rasley et al., 2020). We used 8 NVIDIA A100/40GB GPUs for all model training. Training time for DynamicFocalPO is comparable to FocalPO, with the scheduling computation adding negligible overhead (<0.1%).
Statistical Methodology. We report mean ± standard deviation across all runs. Statistical significance is assessed using paired t-tests. For multiple comparisons in Table 1, we apply Bonferroni correction. Results marked with † indicate statistically significant improvement over FocalPO at $p < 0.05$.

For reproducibility, we provide the key hyperparameters:

Hyperparameter	Mistral-Base-7B	Llama-3-Instruct-8B
β	0.01	0.1
Learning rate	5e-7	6e-7
Batch size	128	128
Epochs	1	1
$\gamma_{\max}$	0.1	0.1
$\gamma_{\min}$	0.01	0.01
Schedule	Cosine	Cosine

Table 3: Hyperparameters for DynamicFocalPO experiments.

Our implementation is based on the Transformer Reinforcement Learning (TRL) library.

F Qualitative Analysis

We provide qualitative examples comparing responses generated by DPO and DynamicFocalPO. These examples illustrate how DynamicFocalPO produces more structured, comprehensive, and practically useful responses.

F.1 Example 1: Language Learning Strategies

Question: What are the most effective strategies for learning a new language as an adult?
DPO Response: Learning a new language as an adult can be challenging but rewarding. Here are some tips: 1. Practice regularly - consistency is key 2. Use language learning apps like Duolingo 3. Watch movies in the target language 4. Find a language partner 5. Don't be afraid to make mistakes The most important thing is to stay motivated and practice every day. Good luck!

Table 4: DPO response to the language learning question.

Question: What are the most effective strategies for learning a new language as an adult?
DynamicFocalPO Response: Learning a new language as an adult requires a strategic approach that leverages cognitive strengths while addressing common challenges. Here are evidence-based strategies: **1. Spaced Repetition Systems (SRS)** Use tools like Anki that optimize review intervals based on memory science. This method exploits the spacing effect, improving long-term retention by 200-400% compared to massed practice. **2. Comprehensible Input** Linguist Stephen Krashen's input hypothesis suggests that acquisition occurs when we understand messages slightly above our current level ("i+1"). Start with graded readers and progress to native materials gradually. **3. Active Production Practice** Balance input with output: - *Shadowing*: Repeat audio immediately after hearing it - *Language exchanges*: Platforms like Tandem or HelloTalk connect you with native speakers - *Thinking aloud*: Narrate daily activities in the target language **4. Exploit Adult Cognitive Advantages** Adults excel at pattern recognition and metalinguistic awareness. Study grammar explicitly to accelerate rule internalization, then practice until usage becomes automatic. **5. Create an Immersive Environment** - Change device languages - Listen to podcasts during commutes - Label household items - Join online communities in the target language The research suggests 600-750 hours of study for "easy" languages (Spanish, French) and 2,200+ hours for "hard" languages (Mandarin, Arabic) for professional proficiency. Set realistic milestones accordingly.

Table 5: DynamicFocalPO response to the language learning question.

F.2 Example 2: Overcoming Procrastination

F.3 Analysis

These examples demonstrate several qualitative improvements in DynamicFocalPO responses: DynamicFocalPO responses exhibit clear hierarchical organization with bold headers and sub-bullets that enhance accessibility, while providing specific techniques with explanations of *why* they work rather than generic advice. The responses include concrete implementation details such as specific apps, time estimates, and example "if-then" statements, and acknowledge complexity by offering context-dependent recommendations rather than one-size-fits-all solutions.

We attribute these improvements to the dynamic focusing mechanism: by initially learning from well-ranked preference pairs and gradually incorporating harder examples, DynamicFocalPO develops

Question: How do I deal with procrastination when working on important tasks?
DPO Response: Procrastination is something everyone struggles with. Here are some ways to overcome it: - Break tasks into smaller pieces - Set deadlines for yourself - Remove distractions like your phone - Reward yourself when you complete tasks - Just start - that's often the hardest part Remember, it's normal to procrastinate sometimes. Don't be too hard on yourself. Try to find what works best for you and stick with it.

Table 6: DPO response to the procrastination question.

more robust and nuanced response generation capabilities.

G Training Dynamics Analysis

To empirically validate our curriculum learning motivation, we conduct a detailed analysis of gradient contributions throughout training.

G.1 Methodology

For each training batch, we compute the gradient contribution from two sample groups:

- **Correctly ranked samples** ($p > 0.5$): Samples where the model already assigns higher implicit reward to the preferred response.
- **Incorrectly ranked samples** ($p \leq 0.5$): Samples where the model assigns higher implicit reward to the dispreferred response (misranked).

The gradient contribution ratio is defined as:

$$R(t) = \frac{\sum_{i:p_i>0.5} \|\nabla_\theta \mathcal{L}_i\|}{\sum_i \|\nabla_\theta \mathcal{L}_i\|}, \tag{35}$$

where $\mathcal{L}_i$ is the per-sample loss. We track $R(t)$ throughout training for DPO, FocalPO, and DynamicFocalPO.

G.2 Results

Figure 5 visualizes the gradient contribution ratio throughout training, and Table 8 provides numerical summaries at different stages.

Table 8 reveals several key findings:

(1) DPO lacks curriculum behavior. The gradient contribution ratio remains nearly constant ($\sim$45%) throughout training, meaning DPO treats easy and hard samples similarly regardless of training stage. This aligns with the observation by Chen et al. (2024) that DPO struggles to correct misranked samples.

(2) FocalPO provides static focusing. FocalPO maintains a higher ratio ($\sim$61%) than DPO due to its modulating factor, but this ratio remains constant throughout training. The model cannot adapt its focusing strategy as it improves.

(3) DynamicFocalPO achieves curriculum progression. The gradient contribution ratio decreases from 78% to 47% over training, demonstrating the intended curriculum behavior:

- *Early training*: High $\gamma(t)$ strongly suppresses gradients from misranked samples, allowing the model to establish reliable preference patterns from correctly ranked pairs.
- *Late training*: Low $\gamma(t)$ allows misranked samples to contribute more, and the ratio converges toward DPO's $\sim$46%—exactly as predicted by Proposition 1. This enables the model to tackle harder cases once the foundation is established.

Question: How do I deal with procrastination when working on important tasks?

DynamicFocalPO Response:
Procrastination often stems from emotional regulation difficulties rather than time management failures. Understanding and addressing root causes leads to more effective solutions:

1. Identify the Underlying Emotion
Procrastination typically masks:
- *Fear of failure*: The task feels too high-stakes
- *Perfectionism*: Waiting for "perfect" conditions
- *Overwhelm*: Task scope feels unmanageable
- *Boredom*: Lack of intrinsic motivation

2. Implementation Intentions
Research shows that "if-then" planning increases follow-through by 2-3x. Example: "If it's 9 AM on Monday, then I will work on the report for 45 minutes."

3. Reduce Activation Energy
Make starting frictionless:
- Prepare materials the night before
- Use the "2-minute rule": if the first step takes under 2 minutes, do it immediately
- Create a dedicated workspace that signals "work mode"

4. Time-Boxing Techniques
- *Pomodoro*: 25 minutes work, 5 minutes break
- *Time blocking*: Schedule specific tasks in calendar slots
- *Artificial deadlines*: Set interim milestones before the real deadline

5. Address Perfectionism
Adopt a "draft zero" mindset—your first version's only job is to exist. Editing a mediocre draft is easier than creating from nothing.

6. Strategic Procrastination
When avoiding one task, channel that energy into other productive work ("structured procrastination"). This maintains momentum while giving your mind time to process the avoided task.

Start by identifying which emotional driver applies to your current situation, then select the corresponding technique.

Table 7: DynamicFocalPO response to the procrastination question.

G.3 Connection to Performance Gains

The curriculum progression explains why DynamicFocalPO outperforms both baselines:

- **vs. DPO**: By initially focusing on correctly ranked samples, DynamicFocalPO avoids the noisy gradients from misranked samples that can destabilize early training.
- **vs. FocalPO**: By eventually incorporating harder samples, DynamicFocalPO learns from challenging cases that fixed-γ FocalPO permanently down-weighs.

This analysis provides direct empirical evidence for our curriculum learning motivation and explains the mechanism behind DynamicFocalPO's improvements.

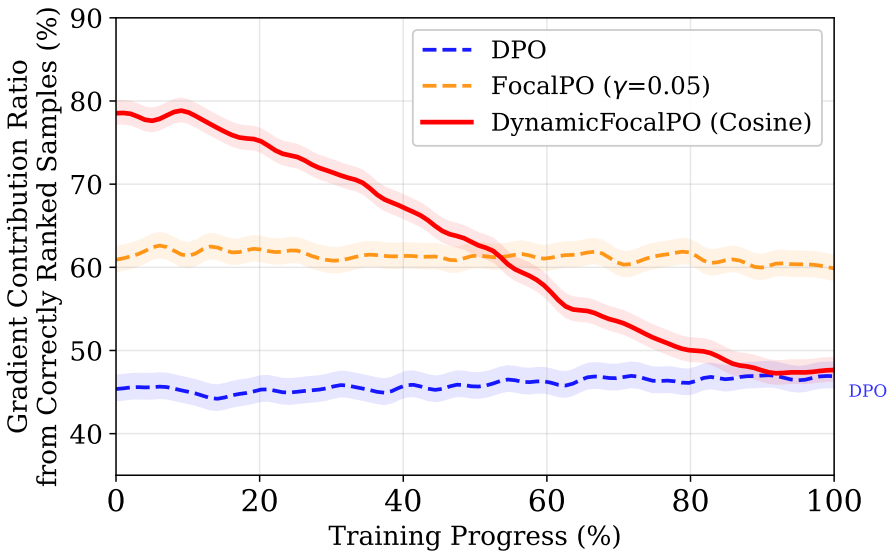

Figure 5: Gradient contribution ratio from correctly ranked samples ($p > 0.5$) during training. DynamicFocalPO shows curriculum-style progression (high→low), while DPO and FocalPO remain relatively constant. Notably, DynamicFocalPO converges toward DPO's ratio in late training, consistent with Proposition 1.

Method	Early (0-20%)	Mid (40-60%)	Late (80-100%)
DPO	44.2±1.3%	45.8±1.5%	46.1±1.4%
FocalPO (γ=0.05)	62.4±1.8%	61.7±1.6%	60.9±1.9%
DynamicFocalPO	**78.3±1.2%**	62.5±1.5%	47.2±1.6%

Table 8: Gradient contribution ratio $R(t)$ from correctly ranked samples at different training stages (Llama-3-Instruct-8B). DynamicFocalPO shows clear curriculum progression: from 78% (early) to 47% (late), converging toward DPO's ∼45% as $\gamma(t) \to \gamma_{\min}$.