YFPO: Yoked Feature Preference Optimization with Neuron-Guided Rewards

Anonymous ACL submission

Abstract

Preference optimization has become a widely used post-training paradigm for improving the reasoning abilities of large language models. Existing methods typically learn from preferred and dispreferred responses as external behavioral supervision, while largely ignoring capability-related signals encoded in the model's internal representations. In this work, we study whether such internal signals can provide useful auxiliary supervision for mathematical reasoning. We introduce **YFPO** (**Yoked Feature Preference Optimization**), a neuron-guided preference optimization framework that couples response-level preference learning with neuron-level rewards. YFPO first uses AttnLRP to identify math-related neurons, and then derives an auxiliary reward from the activation margin of these neurons between preferred and dispreferred responses. This reward is combined with the standard preference optimization objective, encouraging the model to align external preferences with internal math-related features. We conduct small-scale experiments on GSM8K with a compact language model. Results show that neuron-guided rewards can influence preference optimization dynamics and yield measurable improvements in several settings, suggesting that internal representations can serve as lightweight and interpretable signals for reasoning-oriented post-training.

1 Introduction

Preference optimization has become an important post-training paradigm for improving large language models (LLMs), especially on reasoning-oriented tasks. Existing methods, including PPO-style RLHF and direct preference optimization methods such as DPO, IPO, KTO, ORPO, and SimPO, typically rely on external supervision signals such as human preferences, reward models, verifier feedback, or chosen/rejected response pairs (Schulman et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023; Azar et al., 2023; Ethayarajh et al., 2024; Hong et al., 2024; Meng et al., 2024). For mathematical reasoning, such signals are often derived from final-answer correctness, process supervision, or automatic verifiers (Cobbe et al., 2021; Lightman et al., 2023; Shao et al., 2024; Yu et al., 2026). While effective, these objectives mainly tell the model which output is preferred, but rarely use internal computational signals involved in generating that output.

In this work, we ask whether internal model representations can provide useful auxiliary signals for preference optimization. Prior interpretability studies suggest that LLMs contain neuron-level structures associated with factual knowledge, language behavior, and task-specific patterns (Dai et al., 2022; Tang et al., 2024). Attribution methods such as AttnLRP further enable relevance estimation over internal components (Achtibat et al., 2024). For mathematical reasoning, neurons with high relevance on math examples may provide coarse but informative signals about whether the model engages math-related internal features.

We introduce **YFPO** (**Yoked Feature Preference Optimization**), a neuron-guided preference optimization framework that augments DPO-style training with neuron-level rewards. YFPO first uses AttnLRP to identify a fixed set of math-related neurons. During preference optimization, it extracts their activations for preferred and dispreferred responses, constructs an auxiliary reward from the activation margin, and combines this reward with the standard DPO objective. As illustrated in Figure 1, this design uses the chosen/rejected structure of preference data to align external response-level preferences with internal math-related features.

Our contributions are threefold: (1) we study how internal math-related neuron signals can complement external preference data in preference optimization; (2) we propose YFPO, which converts

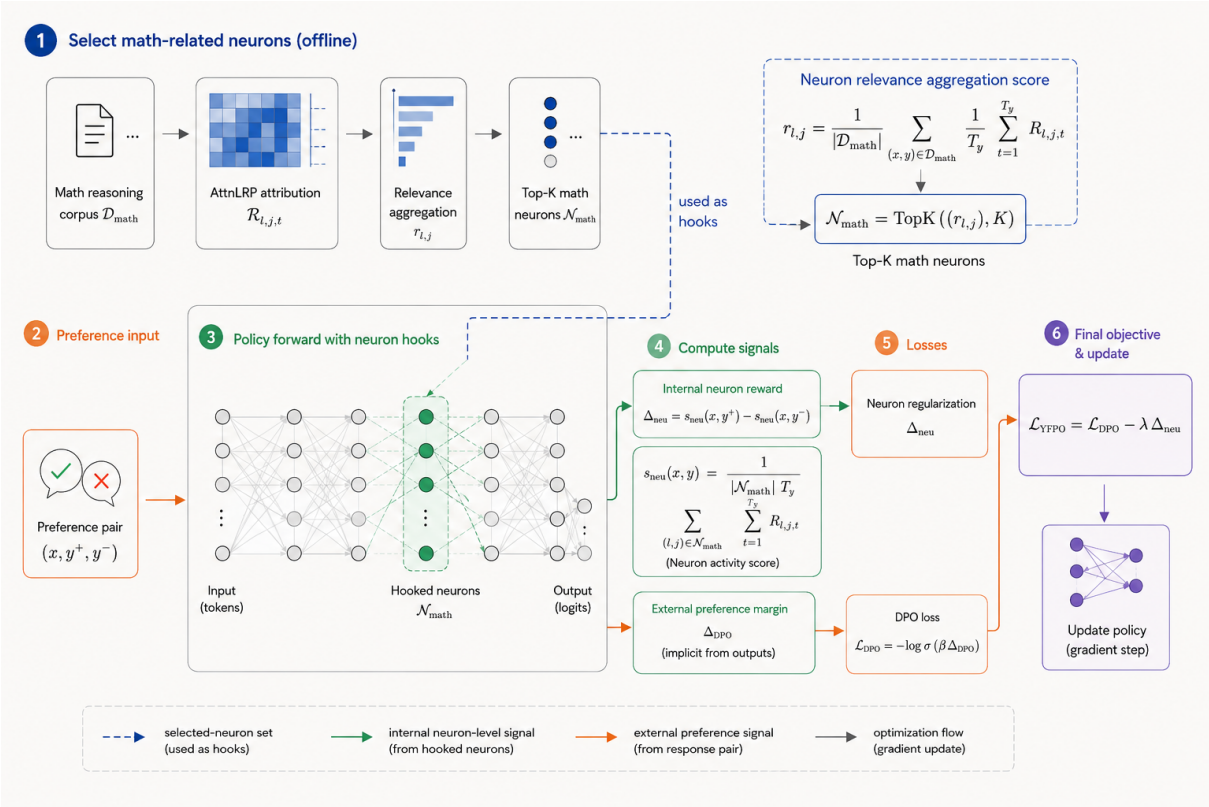

Figure 1: Overview of YFPO. In the offline stage, AttnLRP is used to identify a fixed set of top-K math-related neurons. During preference optimization, YFPO extracts activations from these neurons for preferred and dispreferred responses, constructs a neuron reward margin, and combines it with the DPO objective.

AttnLRP-selected neuron activations into an auxiliary reward for DPO-style training; and (3) we conduct small-scale experiments on Qwen2-0.5B and GSM8K, showing that neuron-guided rewards can influence preference optimization dynamics and yield measurable improvements in several settings.

2 Related Work

Preference optimization and reasoning. Preference optimization has become a central technique for LLM post-training. Early RLHF methods typically use reward models and PPO-style reinforcement learning to align model outputs with human preferences (Schulman et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022). More recent direct preference optimization methods, including DPO, IPO, KTO, ORPO, and SimPO, simplify this pipeline by optimizing on chosen/rejected responses or preference-like labels without explicitly training a separate reward model (Rafailov et al., 2023; Azar et al., 2023; Ethayarajh et al., 2024; Hong et al., 2024; Meng et al., 2024). For mathematical reasoning, preference optimization and reinforcement learning often rely on final-answer correctness, process supervision, verifiers, or preference data as external feedback signals (Cobbe et al., 2021; Lightman et al., 2023; Shao et al., 2024; Yu et al., 2026). In contrast, YFPO studies whether internal neuron-level signals associated with mathematical reasoning can provide auxiliary guidance during preference optimization.

Neuron-level interpretability. Neuron-level interpretability aims to identify internal components associated with specific knowledge, languages, or task behaviors. Prior work has used feature attribution, probing, activation statistics, and intervention to analyze neurons related to factual knowledge or language-specific behavior (Sundararajan et al., 2017; Belinkov, 2022; Dai et al., 2022; Tang et al., 2024). AttnLRP extends layer-wise relevance propagation to transformer models and enables efficient attribution over latent representations (Achtibat et al., 2024). While these methods are commonly used to explain model behavior, YFPO uses relevance-selected math-related neurons as training signals: it converts their activation margin between preferred and dispreferred re-

sponses into an auxiliary reward, connecting internal capability representations with external preference learning.

3 Method

We propose **YFPO**, a neuron-guided preference optimization method that augments DPO-style training with an auxiliary reward from math-related neurons. YFPO contains two stages: offline neuron selection and online preference optimization. In the offline stage, we identify neurons associated with mathematical reasoning using AttnLRP. In the training stage, we compute a neuron reward margin between preferred and dispreferred responses and add it to the DPO objective.

3.1 Math-Related Neuron Selection

We use AttnLRP (Achtibat et al., 2024) to identify neurons associated with mathematical reasoning Given mathematical reasoning samples. AttnLRP assigns relevance scores to intermediate MLP channels according to their contribution to output tokens.

We treat each intermediate MLP channel before `down_proj` as a neuron, indexed by layer l and channel j. For each neuron, we aggregate its signed relevance over mathematical samples and response tokens:

$$r_{l,j} = \mathbb{E}_{x,y,t}\left[R_{l,j,t}(x,y)\right], \tag{1}$$

where $R_{l,j,t}$ denotes the relevance assigned to neuron (l,j) at response token t. We select the Top-K neurons with the largest signed relevance scores as the math-related neuron set:

$$\mathcal{N}_{\text{math}} = \mathrm{TopK}_{(l,j)}\left(r_{l,j}\right). \tag{2}$$

Unless otherwise stated, we use $K = 1024$ and keep $\mathcal{N}_{\text{math}}$ fixed during preference optimization.

3.2 Neuron-Guided Reward

Given a preference example (x, y^+, y^-), where y^+ and y^- denote the preferred and dispreferred responses, we extract activations of neurons in $\mathcal{N}_{\text{math}}$ using forward hooks. For a response y, we define its math-neuron score as:

$$s_{\text{neu}}(x,y) = \mathbb{E}_{(l,j)\in\mathcal{N}_{\text{math}},\, t\in\mathcal{T}_y}\left[a_{l,j,t}(x,y)\right], \tag{3}$$

where $a_{l,j,t}$ is the activation of neuron (l,j) at response token t, and $\mathcal{T}_y$ denotes response tokens. The neuron reward margin is:

$$\Delta_{\text{neu}} = s_{\text{neu}}(x,y^+) - s_{\text{neu}}(x,y^-). \tag{4}$$

A larger Δ_{neu} indicates that the preferred response obtains a higher activation score on the selected math-related neurons than the dispreferred response.

3.3 YFPO Objective

YFPO uses DPO (Rafailov et al., 2023) as the base objective. For a preference example (x, y^+, y^-), the standard DPO margin is:

$$\begin{aligned}\Delta_{\text{DPO}} = &\left[\log \pi_\theta(y^+|x) - \log \pi_\theta(y^-|x)\right] \\ &- \left[\log \pi_{\text{ref}}(y^+|x) - \log \pi_{\text{ref}}(y^-|x)\right].\end{aligned} \tag{5}$$

YFPO adds the neuron reward margin to the DPO loss:

$$\mathcal{L}_{\text{YFPO}} = -\log \sigma(\beta\Delta_{\text{DPO}}) - \lambda\Delta_{\text{neu}}, \tag{6}$$

where β controls the DPO preference strength and λ controls the neuron-guided reward strength. When $\lambda = 0$, YFPO reduces to standard DPO. When $\lambda > 0$, the model is optimized to prefer chosen responses while increasing their math-neuron activation margin over rejected responses.

Additional implementation details are provided in Appendix A.

4 Experiments

We conduct small-scale experiments to examine whether neuron-guided rewards can complement DPO-style preference optimization for mathematical reasoning.

Setup. We use Qwen2-0.5B as the base model and `xlinla/Math-Step-DPO-10K` as the preference dataset. We evaluate both the full 10K training set and a randomly sampled 2K subset. For YFPO, math-related neurons are selected offline using AttnLRP with $K = 1024$, and the neuron reward margin Δ_{neu} is added to the DPO objective. We compare standard DPO with DPO+YFPO under different λ values. All models are evaluated on GSM8K using greedy decoding. The original Qwen2-0.5B obtains 0.3654 accuracy on GSM8K. Additional training-dynamics analysis and training details are provided in Appendices B and C.

Results. Table 1 shows results on the full 10K training set. Standard DPO provides limited improvement over the original model, reaching a best accuracy of 0.3662. Adding YFPO improves the best checkpoint accuracy under both tested λ

Metric	DPO	$\text{YFPO}_{0.05}$	$\text{YFPO}_{1.0}$
C1	0.3647	0.3738	0.3738
C2	0.3662	0.3586	0.3669
C3	0.3639	0.3707	0.3692
Avg.	0.3649	0.3677	0.3700
Best	0.3662	0.3738	0.3738
$\Delta_{\text{Avg.}}$	–	+0.0028	+0.0051
Δ_{Best}	–	+0.0076	+0.0076

Table 1: GSM8K results on Qwen2-0.5B using the full 10K training set. YFPO_{λ} denotes DPO+YFPO with coefficient λ. C1/C2/C3 denote three saved checkpoints.

Metric	DPO	$\text{YFPO}_{1.0}$	$\text{YFPO}_{0.1}$	$\text{YFPO}_{0.2}$	$\text{YFPO}_{0.3}$
C1	0.3601	0.3616	0.3730	0.3616	0.3715
C2	0.3647	0.3692	0.3609	0.3715	0.3669
C3	0.3669	0.3624	0.3669	0.3624	0.3639
Avg.	0.3639	0.3644	0.3669	0.3652	0.3674
Best	0.3669	0.3692	0.3730	0.3715	0.3715
$\Delta_{\text{Avg.}}$	–	+0.0005	+0.0030	+0.0013	+0.0035
Δ_{Best}	–	+0.0023	+0.0061	+0.0046	+0.0046

Table 2: GSM8K results on Qwen2-0.5B using a randomly sampled 2K subset. YFPO_{λ} denotes DPO+YFPO with coefficient λ. C1/C2/C3 denote three saved checkpoints.

values, with both settings reaching 0.3738. The $\lambda = 1.0$ setting also improves average accuracy from 0.3649 to 0.3700.

Table 2 shows results on the 2K subset. YFPO improves the best checkpoint accuracy under all tested λ values. The strongest best-checkpoint result is obtained with $\lambda = 0.1$, reaching 0.3730 and improving over DPO by 0.0061. In terms of average accuracy, $\lambda = 0.3$ performs best, improving from 0.3639 to 0.3674. These results suggest that math-related neuron rewards can provide useful auxiliary signals even under a lower-resource preference training setting.

Analysis. The results suggest that YFPO provides a useful auxiliary signal for preference optimization. On the full 10K set, YFPO improves the best checkpoint accuracy from 0.3662 to 0.3738, and the $\lambda = 1.0$ setting improves the average accuracy from 0.3649 to 0.3700. On the 2K subset, YFPO improves best checkpoint accuracy under all tested λ values, with the strongest result reaching 0.3730. These gains are modest but consistent across data-scale settings, indicating that math-related neuron activations can complement external chosen/rejected supervision.

However, YFPO does not follow a simple monotonic trend with respect to λ. Different reward strengths lead to different best checkpoints, showing that neuron-guided rewards are sensitive to reward scale and training stage. Appendix B further shows that YFPO preserves the external DPO preference signal while introducing a distinct neuron-level signal. This suggests that the neuron reward is not merely a post-hoc explanatory metric, but an active training signal that can influence preference optimization dynamics.

5 Conclusion

We introduced YFPO, a neuron-guided preference optimization method that augments DPO-style training with rewards derived from math-related neurons. YFPO uses AttnLRP to identify relevant neurons offline and computes a neuron activation margin between preferred and dispreferred responses during training. In this way, YFPO connects external response-level preference supervision with internal capability-related signals.

Experiments on Qwen2-0.5B and GSM8K show that YFPO brings measurable improvements over standard DPO under multiple settings. On both the full 10K training set and the 2K subset, YFPO improves best-checkpoint accuracy, and in several cases also improves average checkpoint accuracy. These results indicate that math-related neuron activations can provide useful auxiliary supervision.

Our analysis further suggests that neuron-guided rewards influence preference optimization dynamics while preserving the external DPO preference signal. Although the gains remain modest and sensitive to reward scale and checkpoint selection, YFPO provides measurable evidence that internal neuron-level signals can complement external preference data for reasoning-oriented post-training.

6 Limitations

This work has several limitations. First, our experiments are limited in scale. We mainly evaluate YFPO on Qwen2-0.5B and GSM8K, without sys-

tematic experiments on larger models, instruction-tuned models, or broader mathematical reasoning benchmarks. Therefore, the current results should be interpreted as initial evidence in a small-model setting rather than a general conclusion across model scales and tasks.

Second, the gains of YFPO are modest and checkpoint-sensitive. Although YFPO improves over DPO under several settings, its effectiveness depends on the reward coefficient λ, training stage, and data scale. Future work should conduct more systematic λ sweeps, multi-seed experiments, and larger-scale evaluations to better assess the stability of neuron-guided rewards.

Third, YFPO relies on a fixed set of AttnLRP-selected neurons and uses their activation margin as an auxiliary reward. This design is simple and interpretable, but the selected neurons may change their role during training, and activation signals do not necessarily imply true causal contribution to reasoning ability. Future work should explore dynamic neuron selection, random-neuron baselines, causal intervention, and extensions to other post-training objectives such as SimPO, ORPO, GRPO, and process reward learning.

References

Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. 2024. Attnlrp: Attention-aware layer-wise relevance propagation for transformers. In *International Conference on Machine Learning*.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *International Conference on Machine Learning*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. 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*.

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. *Advances in Neural Information Processing Systems*.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. 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*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *International Conference on Machine Learning*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2026. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244.

A Additional Implementation Details

Hook placement. We define neurons as intermediate MLP channels before `down_proj`. Therefore, during YFPO training, forward hooks are applied to the input activation of `down_proj`, rather than its output. This keeps the activation space used for neuron reward computation aligned with the neuron set selected by AttnLRP.

Response-token masking. The neuron score is computed only over response tokens. Prompt tokens are excluded because the length and content of the input problem may otherwise dominate the activation score. We average over both selected neurons and response tokens to reduce sensitivity to sequence length and neuron count.

Signed relevance selection. YFPO uses the signed relevance values produced by AttnLRP for neuron selection. Specifically, neurons are selected according to their aggregated relevance scores $r_{l,j}$, rather than the absolute values $|r_{l,j}|$. Thus, the selected neurons are those with the largest positive relevance to mathematical reasoning outputs under the attribution procedure.

Activation-based reward. The neuron reward is computed from the raw activations of the selected neurons, without taking absolute values. This reward encourages preferred responses to obtain a higher activation score on the selected math-related neurons than dispreferred responses.

Reference model. The reference model is used only for computing the DPO log-probability margin. It does not participate in the neuron reward computation. The neuron reward is computed from the current policy model and remains differentiable with respect to the policy parameters.

Fixed neuron set. The selected math-related neuron set $\mathcal{N}_{\text{math}}$ is kept fixed throughout preference optimization. This makes different training runs comparable and avoids introducing extra instability from dynamically changing neuron sets.

B Training Dynamics

Figure 2 shows the training dynamics of YFPO on the 2K subset. The external DPO preference margin follows broadly similar trends across DPO and YFPO variants, suggesting that the neuron reward does not collapse the original preference optimization process. Meanwhile, the internal neuron reward exhibits different trajectories across λ values, indicating that it acts as an additional optimization signal rather than a static explanatory metric. These dynamics support the main results: YFPO can improve best-checkpoint accuracy under multiple reward strengths, but its effect remains sensitive to reward scale and training stage.

Hyperparameter	Value
Base model	Qwen2-0.5B
Training precision	bfloat16
Random seed	42
Data seed	42
Training epochs	3
Save strategy	per epoch
Batch size / device	4
Gradient accumulation	2
Effective batch size / device	8
Learning rate	1×10^{-6}
Max sequence length	4096
DPO loss type	sigmoid
DPO β	`DPOConfig` default
Optimizer	`DPOConfig` default
Warmup schedule	`DPOConfig` default
Selected neurons K	1024
YFPO coefficient λ	reported in Tables 1 and 2
Evaluation	GSM8K, greedy decoding

Table 3: Training configuration used for DPO and DPO+YFPO experiments.

C Additional Training Details

We report results from three saved training checkpoints, denoted as C1, C2, and C3 in the main text. The average accuracy reflects overall checkpoint-level behavior, while the best accuracy reflects the strongest observed checkpoint under each setting. This reporting protocol is used because YFPO shows visible checkpoint-level variance under different reward strengths.

Training configuration. Table 3 summarizes the main training configuration. We train Qwen2-0.5B with bfloat16 precision and use the sigmoid DPO loss. The random seed and data seed are both set to 42. Checkpoints are saved at the end of each epoch, yielding C1, C2, and C3. For YFPO, we use $K = 1024$ selected math-related neurons and sweep λ over the values reported in the main tables. Unless explicitly specified, other DPO-related hyperparameters follow the default values of `DPOConfig`. We use the default values in the TRL version used in our experiments.

The 2K subset used in Table 2 is a randomly sampled subset of `xlinla/Math-Step-DPO-10K`.

It refers to a reduced training-data setting, not a random-neuron baseline.

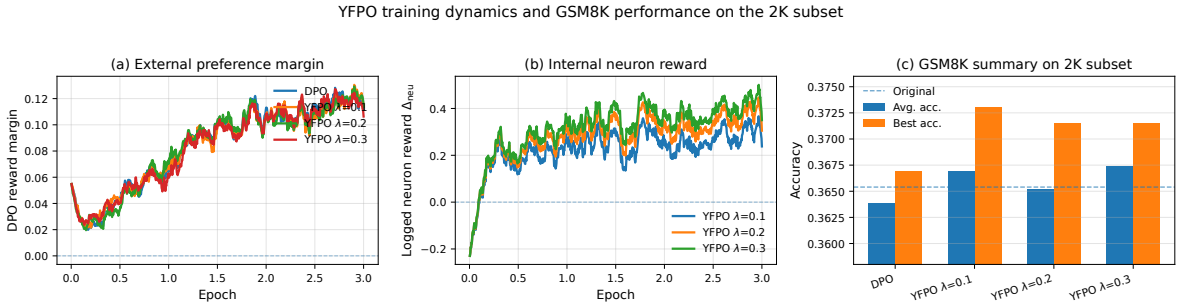

Figure 2: Training dynamics and GSM8K performance of YFPO on the 2K data subset. Left: external DPO preference margin during training. Middle: logged internal neuron reward for different λ values. Right: average and best GSM8K accuracy across three checkpoints. YFPO preserves the external preference optimization signal while introducing a distinct neuron-level optimization signal, and achieves better best-checkpoint accuracy than DPO under multiple λ settings.