NAPO: NEURON ACTIVATION-BASED PREFERENCE OPTIMIZATION FOR MULTILINGUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models exhibit uneven reasoning abilities across languages due to imbalances in multilingual training data. Recently, preference optimization has substantially improved multilingual reasoning performance. However, existing methods typically derive preference signals either from external translation models or from coarse-grained internal signals, such as hidden-state representations. These signals may entangle reasoning-relevant information with language-specific or surface-form noise, resulting in inaccurate preference estimation and ultimately limiting improvements in multilingual reasoning. To address this issue, we propose **N**euron **A**ctivation-based **P**reference **O**ptimization (**NAPO**), a framework that derives cross-lingual preference signals from internal fine-grained neuron activations. Specifically, we first identify reasoning-relevant neurons shared across languages. We then introduce a position-weighted alignment reward computed in the identified neuron space to construct pairwise preference data. Finally, we iteratively optimize the policy within the standard direct preference optimization framework. We conduct extensive experiments on three popular multilingual mathematical reasoning benchmarks across 10 languages. The results show that NAPO consistently outperforms multiple strong preference optimization-based baselines.

1 INTRODUCTION

Large language models (LLMs) have demonstrated superior performance on reasoning tasks such as mathematical problem solving Ahn et al. (2024); Ouyang et al. (2022). However, existing studies show that due to the imbalance of pre-training and fine-tuning data across languages, current LLMs predominately benefit the high-resource languages such as English, while performance on a wide range of low-resource languages remains inferior Qin et al. (2023); Shi et al. (2023). Given that reasoning ability is generally considered language-shared, this cross-lingual gap has motivated growing interest in improving the multilingual reasoning performance of LLMs Chen et al. (2024); She et al. (2024); Tran et al. (2026); Zhang et al. (2025).

To improve reasoning in low-resource languages, one line of research focuses on constructing multilingual training data for continual supervised fine-tuning. Although these approaches have achieved substantial progress Shi et al. (2023); Chen et al. (2024); Zhu et al. (2024), translated data may contain errors introduced by machine translation, particularly when translating long and structurally complex mathematical rationales. Moreover, supervised fine-tuning primarily encourages imitation of provided demonstrations and does not directly provide a mechanism for ranking competing reasoning trajectories, which may limit its generalization to out-of-domain problems.

In contrast to supervised fine-tuning, preference optimization provides a flexible way to improve multilingual reasoning by converting response quality or cross-lingual consistency into relative supervision over candidate responses. Recent studies have shown that such preference signals can improve reasoning consistency and generalization in lower-resource languages, typically by post-training the model with direct preference optimization She et al. (2024); Tran et al. (2026); Zhang et al. (2025); Yang et al. (2025a).

However, existing preference optimization methods typically rely on rewards derived from external models, such as translation model She et al. (2024) or English-centric models Yang et al. (2025a). Such methods entangle reasoning trajectories with translation artifacts or surface-form differences, lead to inaccurate preference estimation and ultimately limit improvements in multilingual reasoning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To reduce reliance on external supervision, recent methods derive internal preference signals from hidden representations Tran et al. (2026) or generation preference Yang et al. (2025b). Nevertheless, these signals are typically computed over dense representations at layer-level hidden states or over aggregate output-level scores. Consequently, both coarse-grained internal signals may mix reasoning-irrelevant information, making it difficult to isolate the components that actually encode cross-lingual reasoning consistency.

To address these limitations, we propose **N**euron **A**ctivation-based **P**reference **O**ptimization (**NAPO**), which uses the similarity of neuron activation patterns as a preference signal. Motivated by research on neuron interpretability, we hypothesize that, although responses in different languages may vary substantially in surface form and token length, reasoning-consistent responses should induce more similar internal neuron activation patterns than reasoning-inconsistent responses. Specifically, we leverage neuron activation probabilities to identify neurons shared across multiple languages. We then construct pairwise preference data using a position-weighted alignment reward computed within the identified neuron space. Finally, we iteratively optimize the policy using the standard direct preference optimization framework.

Our major contributions can be summarized as follows:

- We propose NAPO, a neuron activation-based preference optimization framework for multilingual mathematical reasoning that addresses the limitations of previous methods relying on external signals or coarse-grained internal representations.
- We develop a position-weighted alignment reward estimation method based on neuron activation patterns, which provides fine-grained alignment signals and an interpretable basis for estimating cross-lingual rewards.
- We conduct extensive experiments on three widely used multilingual mathematical reasoning benchmarks covering 10 languages. The results demonstrate that **NAPO** consistently outperforms multiple strong preference-optimization baselines.

2 Related Work

2.1 Multilingual Mathematical Reasoning

Multilingual mathematical reasoning aims to enable large language models to solve mathematical problems accurately across diverse languages. Early work in this area mainly focused on benchmark construction, introducing datasets such as MGSM, MSVAMP, and MNumGLUE to evaluate whether mathematical reasoning and chain-of-thought prompting can transfer across languages (Shi et al., 2023; Chen et al., 2024; Mishra et al., 2022). Building on these benchmarks, one line of research improves multilingual reasoning through translated training data and continual supervised fine-tuning (Chen et al., 2024; Zhu et al., 2024; Huang et al., 2024). Although effective, these methods rely on imitation of translated or annotated supervision, and thus remain vulnerable to translation noise and overfitting. Recent work moves beyond imitation-based tuning and formulates multilingual reasoning improvement as a preference optimization problem She et al. (2024). MAPO compares target-language responses with English references through translation-based rewards (She et al., 2024), while language-imbalance-driven rewarding further exploit the stronger English capability of LLMs to construct preference data (Yang et al., 2025b;a). Some studies further calibirate preference signals via latent representation similarity Tran et al. (2026) and consistency-guided reference selection Zhang et al. (2025). These studies demonstrate the value of reward-based alignment for multilingual reasoning. However, most existing rewards remain output-level, translation-mediated, English-biased, or based on coarse dense representations. In contrast, our work explores whether neuron-level activation patterns can provide a more internal and reliable signal for multilingual preference construction.

2.2 Neuron Activation Interpretability

Neuron activation interpretability studies suggest that internal states of LLMs encode meaningful linguistic, factual, and reasoning-related information rather than serving as opaque intermediate computations. Foundational work shows that transformer feed-forward layers can behave like key-value memories (Geva et al., 2021), and that specific neurons may mediate factual knowledge (Dai

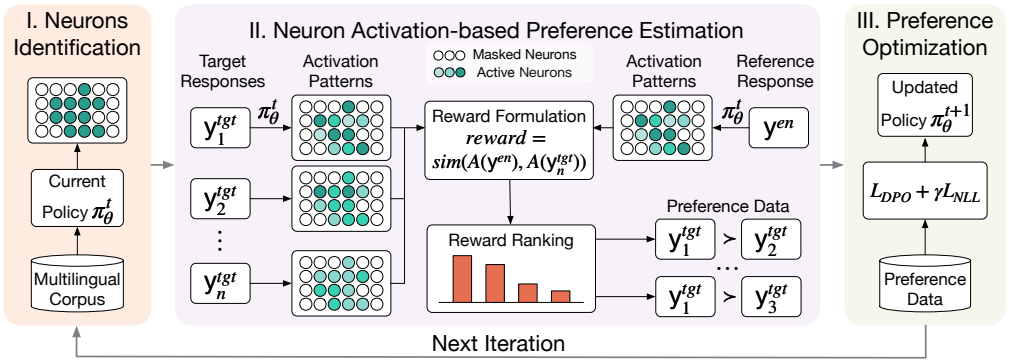


Figure 1: Overview of the proposed NAPO framework. (I) We identify neurons whose activation statistics are shared across languages, thereby defining a language-agnostic comparison space. (II) We compare target-language responses with their English references in this shared activation space to obtain an activation-based preference reward. (III) We use the resulting preference pairs to update the policy and repeatedly reconstruct the preference data as the policy evolves.

et al., 2022), positional information (Voita et al., 2024) in pretrained transformers. In multilingual settings, recent studies identify language-specific neurons as important factors behind multilingual capabilities (Tang et al., 2024), analyze how alignment changes multilingual behavior from a language-neuron perspective (Zhang et al., 2026), and evaluate cross-linguistic alignment through neuron-level correspondence (Huang et al., 2025). Beyond language identity, a growing line of work connects neuron activations with reasoning mechanisms: neuron activation has been used to explain chain-of-thought arithmetic reasoning (Rai & Yao, 2024), comparative neuron analysis has revealed arithmetic mechanisms inside LLMs (Yu & Ananiadou, 2024), and activation-differential analysis has been used to enhance CoT reasoning (Tang et al., 2025). These studies motivate our use of neuron-level activations as signals for multilingual reasoning alignment. Different from prior work that primarily analyzes or intervenes on neurons for interpretation, we use selected activation to construct preference rewards for multilingual mathematical reasoning.

3 Proposed Method

In this section, we introduce **N**euron **A**ctivation-based **P**reference **O**ptimization (**NAPO**), which consists of three stages. Figure 1 presents an overview of the framework.

Problem Formulation Let (x, y) denote query and corresponding response in different languages. At iteration t, the current policy π_θ generates a candidate set $\{y^{tgt}\}$ for each target language input x^{tgt}. Following prior work She et al. (2024); Tran et al. (2026); Zhang et al. (2025), we assume the English reference response y^{en} is available for each problem during preference construction. Our goal is to rank the preference between $\{y^{tgt}\}$ and optimize π_θ under a standard preference optimization framework.

3.1 Language-Shared Neurons Identification

To compare responses across languages, we first need a representation space whose activation statistics are sufficiently shared across languages. We therefore identify language-shared neurons from multilingual corpora before constructing activation-based preference signals. To construct activation-based preference signals, we first identify a set of language-shared neurons from multilingual corpora. Inspired by prior neuron-level interpretability Tang et al. (2024), we take neurons in the feed-forward network (FFN) as the basic units. For layer i, let $\tilde{\mathbf{h}}_i \in \mathbb{R}^d$ be the FFN input and define the neuron activation vector as

$$\mathbf{z}_i = \text{act_fn}(\tilde{\mathbf{h}}_i W_1^i),$$

where $z_{i,j}$ denotes the activation of the j-th neuron in the i-th layer. Given target low-resource language corpora, we compute for each neuron its activation probability on the target language:

$$p_{i,j}^{tgt} = \mathbb{E}\left[\mathbb{I}(z_{i,j} > 0)\right],$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. To convert $p_{i,j}^{tgt}$ into a valid probability distribution, we apply L1 normalization, yielding $\hat{p}_{i,j}^{tgt}$. We measure the cross-lingual balance of neuron (i, j) by

$$H_{i,j} = -\sum_{tgt \in L} \hat{p}_{i,j}^{tgt} \log \hat{p}_{i,j}^{tgt}.$$

We retain the top-$r\%$ neurons ranked by $H_{i,j}$ and denote the resulting set by $\mathcal{N}^{\text{ent}}$. To gurantee these neurons are both broadly shared across languages and sufficiently active, we further compute the average positive activation

$$a_{i,j}^{tgt} = \mathbb{E}\left[\max(z_{i,j}, 0)\right].$$

Let τ_b be the global top-$b\%$ cutoff computed over $\{a_{i,j}^{tgt}\}$. We keep neurons whose activations exceed this threshold in every language as N^{act}. Finally, the language-shared neuron set is defined as

$$N^{\text{sha}} = N^{\text{ent}} \cap N^{\text{act}}.$$

3.2 Position-Weighted Preference Estimation

Once the language-shared neuron subspace has been identified, we use it to compare target-language candidates with their English references. A direct token-level comparison is problematic because responses may have different lengths, and the reliability of activation patterns can vary across layers. We therefore first construct a compact activation-pattern representation and then aggregate similarity across reliable intermediate layers.

Neuron Activation Pattern Representation Given a model with L layers, where the l-th layer contains N_l^{sha} identified shared neurons, a response $y = (y_1, \ldots, y_M)$ with M generated tokens induces a corresponding activation value for each token on each shared neuron at every selected layer during generation. Therefore, the response can be represented as an *activation pattern* organized along three dimensions: token, layer, and neuron. Let $\mathbf{z}_l^m(y)$ denote the FFN activation vector at layer l for the m-th generated token. Restricting it to the shared-neuron subspace gives

$$\mathbf{a}_l^m(y) = \mathbf{z}_l^m(y)\big[N_l^{\text{sha}}\big].$$

The response y can therefore be represented as an activation pattern

$$\mathcal{A}(y) = \{\mathbf{a}_l^m(y)\}_{m=1}^{M},$$

which is organized along token, layer, and neuron dimensions.

Position-Weighted Reward To enable activation-pattern similarity computation between responses of different lengths, we first apply position-weighted pooling over the token dimension. Specifically, for each selected layer $l \in \mathcal{S}$, we aggregate token-level activations as

$$\bar{\mathbf{a}}_l(y) = \sum_{m=1}^{M} \omega_m \mathbf{a}_l^m(y),\ \omega_m = \frac{m}{\sum_{u=1}^{M} u}.$$

Since the lower and upper layers of multilingual LLMs contain more language-specific neurons, which may interfere with activation-pattern similarity computation, we restrict our comparison to intermediate layers $\mathcal{S}$ (e.g., layers 11–21 for Mistral). Then, given a target-language response y^{tgt} and its English reference response y^{en}, we compute the layer-wise cosine similarity in the shared-neuron activation space:

$$s_l(y^{tgt}, y^{en}) = \cos\Big(\bar{\mathbf{a}}_l(y^{tgt}), \bar{\mathbf{a}}_l(y^{en})\Big),\ l \in \mathcal{S}.$$

Finally, we define the preference reward as the weighted sum of layer-wise similarities, where the layer weights α_l are derived from the normalized hidden-state similarity at each layer.

$$r(y^{tgt}, y^{en}) = \sum_{l \in \mathcal{S}} \alpha_l\, s_l(y^{tgt}, y^{en}).$$

Given the candidates responses with their reward scores, we randomly sample two responses and compare their rewards. A target response with a higher reward $r(y^{tgt}, y^{en})$ is more likely to be preferred in the subsequent alignment-oriented preference optimization stage. Collecting such pairs over all queries and languages yields the multilingual preference dataset $\mathcal{D} = \{x, (y_w, y_l)\}$, which is then used for iterative preference optimization in the next stage.

3.3 Iterative Preference Optimization

Given the multilingual preference dataset $\mathcal{D} = \{x, (y_w, y_l)\}$, we update the current policy with the standard DPO objective:

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

where π_θ is the current policy, π_{ref} is the reference model, and β controls the sharpness of the preference optimization. To further stabilize training, we additionally introduce a negative log-likelihood (NLL) term on the preferred responses, encouraging the model to maintain high likelihood on the chosen outputs:

$$\mathcal{L}_{\text{NLL}} = -\mathbb{E}_{(x,y_w)\sim\mathcal{D}}\left[\frac{\log\pi_\theta(y_w \mid x)}{|y_w|}\right],$$

where $|y_w|$ denotes the token length of the preferred response. The final training objective is defined as $\mathcal{L} = \mathcal{L}_{\text{DPO}} + \gamma\mathcal{L}_{\text{NLL}}$, where γ is a hyper-parameter balancing preference optimization and likelihood regularization. After each round of optimization, the updated policy is used to generate new candidate responses, and NAPO reconstructs multilingual preference pairs using the activation-based reward in Section **??**. Repeating this process for T iterations yields the final aligned model π_{θ_T}.

4 Experiments

4.1 Experimental Setup

Datasets We utilize three challenging benchmarks: MSVAMP Chen et al. (2024), MNumGLUE-Sub Mishra et al. (2022); She et al. (2024), and MGSM Shi et al. (2023) , which involves ten languages: English (en), Chinese (zh), Spanish (es), French (fr), Russian (ru), Bengali (bn), German (de), Thai (th), Swahili (sw), and Japanese (ja). We use MSVAMP as an out-of-domain benchmark that is excluded from training, in order to evaluate model robustness and generalization. MNumGLUESub is the in-domain test set corresponding to the data used for preference estimation and optimization. MGSM is the testset corresponding to MGSM8KInstruct, on which the base models are trained.

Implementation Details We identify neurons using multilingual corpora from MGSM8KInstruct, which is constructed by translating GSM8K into nine non-English languages Chen et al. (2024). For each target-language query, we sample 20 candidate responses from the current policy and compute their activation similarity rewards with respect to the English reference responses. Our code is implemented using the TRL library von Werra et al. (2020) and DeepSpeed Rasley et al. (2020) on 4 NVIDIA H200 80GB GPUs. We optimize the model with a learning rate of 1×10^{-7} using the AdamW optimizer Loshchilov & Hutter (2019). For DPO, we set the weight of the auxiliary NLL term to $\gamma = 1.0$. We run the iterative preference optimization for 3 rounds. We use the accuracy of problem-solving to measure the reasoning ability.

Baselines We compare NAPO against multiple multilingual mathematical reasoning baselines: *MetaMath* Yu et al. (2024) is an English mathematical reasoning model trained on MetaMathQA, serving as a strong English-centric reasoning baseline. *MathOctopus* Chen et al. (2024) is fine-tuned from LLaMA2 with multilingual reasoning data from MGSM8KInstruct, and is used as a multilingual mathematical reasoning baseline. *MistralOctopus* She et al. (2024) further fine-tunes MetaMath-Mistral Yu et al. (2024) on MGSM8KInstruct, combining English mathematical reasoning with multilingual supervision. *m-RFT* Yuan et al. (2023) extends rejection sampling fine-tuning to the multilingual setting, where sampled solutions that yield correct answers are used to further fine-tune the base model. *MAPO* She et al. (2024) formulates multilingual reasoning alignment as preference optimization, using translation-based answer consistency as preference supervision for DPO or PPO.

MNumGLUESub	Bn	Th	Sw	Ja	Zh	Ru	De	Es	Fr	En	Lrl.	Hrl.	Avg.
MetaMath [2024]	13.2	19.2	12.1	50.7	53.3	52.9	54.2	56.5	56.7	63.3	14.8	55.4	43.2
MathOctopus [2024]	26.6	30.9	34.3	40.9	44.4	36.0	32.6	42.0	36.2	46.9	30.6	39.9	37.1
MistralOctopus [2024]	45.2	50.7	49.5	56.5	65.3	59.1	51.4	62.1	53.9	74.6	48.5	60.4	56.8
m-RFT [2023]	52.7	60.6	61.0	67.2	70.8	65.9	61.2	71.6	64.4	78.3	58.1	68.5	65.4
MAPO† [2024]	62.3	64.6	61.6	72.1	75.1	68.0	69.3	74.4	74.2	78.5	62.8	73.1	70.0
CM-Align† [2025]	53.7	62.3	62.1	67.6	71.0	68.2	62.5	69.9	66.3	75.7	59.4	68.7	65.9
LaCoMSA† [2026]	59.1	63.1	64.4	68.5	72.9	71.9	71.0	72.3	72.5	79.3	62.2	72.6	69.5
NAPO (Ours)	**66.1**	**69.5**	**69.7**	**74.2**	**76.6**	**75.7**	**75.7**	**78.7**	**78.0**	**80.8**	**68.4**	**77.1**	**74.5**
MSVAMP	**Bn**	**Th**	**Sw**	**Ja**	**Zh**	**Ru**	**De**	**Es**	**Fr**	**En**	**Lrl.**	**Hrl.**	**Avg.**
MetaMath [2024]	14.8	17.7	14.8	51.8	54.4	59.4	59.6	63.2	62.0	64.8	15.8	59.3	46.2
MathOctopus [2024]	27.7	35.9	39.4	41.6	42.7	44.2	44.0	45.1	45.3	46.4	34.3	44.2	41.2
MistralOctopus [2024]	47.0	52.6	54.4	58.8	60.2	59.8	62.1	60.8	60.4	74.4	51.3	62.4	59.0
m-RFT [2023]	57.4	63.5	60.8	65.2	69.8	67.4	67.2	69.1	68.0	71.3	60.6	68.3	66.0
MAPO† [2024]	60.5	67.9	67.9	70.7	75.0	76.4	75.7	77.1	75.7	**80.9**	65.4	75.9	72.8
CM-Align† [2025]	60.3	67.3	67.1	69.8	72.5	70.2	72.1	73.0	71.5	76.2	64.9	72.2	70.0
LaCoMSA† [2026]	57.6	63.5	62.9	65.4	69.0	64.8	68.2	68.1	68.0	75.7	61.3	68.5	66.3
NAPO (Ours)	**63.3**	**69.9**	**72.0**	**73.4**	**77.7**	**76.9**	**78.8**	**79.1**	**78.1**	80.6	**68.4**	**77.8**	**75.0**

Table 1: Accuracy (%) on MNumGLUESub and MSVAMP. Alongside average (Avg.) accuracy, we also report average accuracy of high-resource languages (Hrl.) and low-resource languages (Lrl.) classified by Shi et al. (2023). Bold text denotes the highest accuracy among methods of the same language.

LaCoMSA Tran et al. (2026) constructs latent-representation rewards for multilingual self-alignment and adds a language-consistency factor to reduce off-target generation. *CM-Align* Zhang et al. (2025) constructs multilingual preference data via consistency-based English reference selection and cross-lingual preference construction.

4.2 Main Results

Experimental results in Table 1, 2 demonstrate that our NAPO consistently achieves the best overall performance on all benchmarks. The dagger symbol† denotes that the results for these models from our reimplementation. Specifically, NAPO outperforms the strongest average MAPO baseline by 4.5 and 2.2 points on MNumGLUESub and MSVAMP, respectively.

The improvements are more pronounced on low-resource languages. On MSVAMP, NAPO improves the low-resource average by 3.0 points, from 65.4 to 68.4, and the high-resource average by 1.9 points, from 75.9 to 77.8. One possible explanation is that low-resource languages suffer more from unstable or language-specific reasoning trajectories, and therefore benefit more from our internal preference signal, which encourages target-language responses to align with the language-shared activation patterns of the English reference.

Notably, NAPO also improves English performance, despite using no English preference pairs during optimization. Relative to the base model MistralOctopus, English accuracy increases from 74.6 to 80.8 on MNumGLUESub and from 74.4 to 80.6 on MSVAMP. We attribute this improvement to the English-anchored design of our method: by encouraging target-language responses to align with English reference activations in a language-shared subspace, NAPO also reinforces the shared reasoning patterns that support English performance.

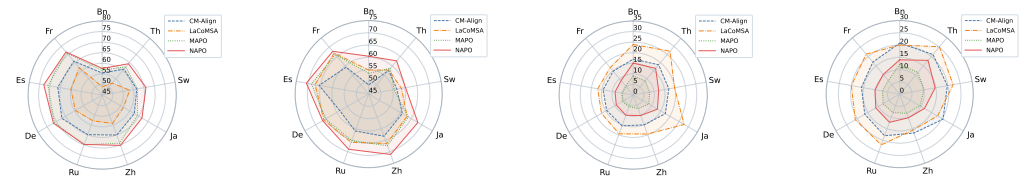
Compared with prior preference optimization baselines, NAPO yields consistently stronger results. It outperforms translation-based reward MAPO, and latent representation reward LaCoMSA. These results suggest that, for multilingual mathematical reasoning alignment, the quality of the alignment signal is a key factor beyond multilingual post-training alone.

MGSM	Bn	Th	Sw	Ja	Zh	Ru	De	Es	Fr	En	Lrl.	Hrl.	Avg.
MetaMath [2024]	6.4	4.0	3.2	39.2	38.8	47.2	56.8	58.0	52.8	63.2	4.5	50.9	37.0
MathOctopus [2024]	29.2	33.6	36.4	35.2	39.2	38.8	44.8	42.4	43.2	52.0	33.1	42.2	39.5
MistralOctopus [2024]	44.0	54.4	50.4	55.6	59.2	58.8	62.4	62.0	56.8	76.0	49.6	61.5	58.0
m-RFT [2023]	41.2	46.8	46.8	48.4	57.2	62.8	61.6	59.2	57.6	72.0	44.9	59.8	55.4
MAPO† [2024]	**52.9**	59.0	60.0	57.8	67.3	**69.2**	**68.8**	70.6	66.4	**81.9**	57.3	68.9	65.4
CM-Align† [2025]	45.2	51.2	51.2	54.8	62.8	64.8	63.2	62.0	57.6	78.4	49.2	63.4	59.1
LaCoMSA† [2026]	49.2	56.0	56.8	53.6	61.2	66.4	63.6	67.2	61.2	77.6	54	64.4	61.3
NAPO (Ours)	50.6	**59.9**	**62.6**	**58.7**	**68.9**	**69.2**	68.1	**71.8**	**67.3**	80.9	**57.7**	**69.3**	**65.8**

Table 2: Accuracy (%) on MGSM.

Model	Bn	Th	Sw	Ja	Zh	Ru	De	Es	Fr	En	Lrl.	Hrl.	Avg.
NAPO-Full	**63.3**	**69.9**	**72.0**	**73.4**	**77.7**	**79.1**	**76.9**	**78.1**	**78.8**	**80.6**	**68.4**	**77.8**	**75.0**
w/o Iteration	61.0	67.8	70.3	70.9	75.1	77.6	74.0	74.9	75.2	78.6	66.4	75.2	72.5
w/o SFT Loss	62.1	66.8	71.1	72.6	75.6	76.1	74.2	77.0	77.8	80.0	66.7	76.2	73.3
w/o Neurons Filtering	59.6	67.1	69.7	71.9	74.3	74.7	76.0	76.2	77.2	79.5	65.5	75.7	72.6

Table 3: Ablation study on MSVAMP.



(a) ACR-MSVAMP (b) ACR-MNumGLUE (c) PPL-MSVAMP (d) PPL-MNumGLUE

Figure 2: Reasoning consistency across nine languages. Higher ACR indicates higher answer consistency. Less perplexity indicates better alignment between the reasoning process.

4.3 Ablation Study

To validate the contribution of each component in NAPO, we introduce the following ablated variants. Table 3 reports the ablation results on the MSVAMP test set.

(i) *w/o Iteration*, which removes the iterative preference reconstruction process. The average accuracy dropping from 75.0 to 72.5, with the low-resource average dropping from 68.4 to 66.4, the high-resource average from 77.8 to 75.2. This suggests that iteratively refreshing activation-based preference pairs is important for progressively refining the alignment signal during training.

(ii) *w/o SFT Loss*, which completely eliminates the supervised fine-tuning loss. In this variant, the English accuracy remains nearly unchanged from 80.6 to 80.0, but the performance on non-English languages drops noticeably. This suggests that the auxiliary SFT loss mainly regularizes non-English preference optimization, helping the model preserve high-quality target-language responses during multilingual reasoning alignment.

(iii) *w/o Neurons Filtering*, which removes the entropy-based filtering step ($r = 1$) in the neuron identification stage. This variant drops from 75.0 to 72.6 in average accuracy, with the low-resource average dropping from 68.4 to 65.5. This suggests that removing entropy filtering introduces noisier neurons, weakens activation-based preference signals and reduces the stability of iterative optimization.

English Question	2 birds and 6 storks were sitting on the fence. 3 more birds came to join them. How many more storks than birds are sitting on the fence?
Chinese Question	有2只鸟和6只鹳鸟坐在篱笆上。又有3只鸟来加入它们。有多少只比鸟多的鹳鸟坐在篱笆上?
Chinese Response	**Before NAPO.** 有2 + 3 = «2+3=5»5只鸟坐在篱笆上。有6 + 3 = «6+3=9»9只鹳鸟坐在篱笆上。 有9 - 5 = «9-5=4»4只比鸟多的鹳鸟坐在篱笆上。 (There are 2 + 3 = «2+3=5» 5 birds sitting on the fence. There are 6 + 3 = «6+3=9»9 storks sitting on the fence. There are 9 - 5 = «9-5=4» 4 more storks than birds sitting on the fence.) ✗ $r_A = 0.9076$ **After NAPO.** 开始时，有2只鸟和6只鹳鸟，所以有2只鸟。又有3只鸟来加入，所以现在有2 + 3 = «2+3=5»5只鸟。还有6只鹳鸟。因此，有6 - 5 = «6-5=1»1只比鸟多的鹳鸟。 (Initially, there are 2 birds and 6 storks, so there are 2 birds. 3 more birds come to join, so now there are 2 + 3 = «2+3=5»5 birds. There are also 6 storks. Therefore, there are 6 - 5 = «6-5=1»1 more storks than birds.) ✓ $r_A = 0.9244$

Table 4: Case study from MSVAMP. The ✓(✗) mark indicates a correct (incorrect) answer. The real-valued numbers r_A denote activation similarity between the Chinese and English questions.

4.4 Analysis of Reasoning Consistency

To investigate whether NAPO improves multilingual reasoning consistency, we evaluate representative baselines on MSVAMP and MNumGLUE using two complementary metrics: Answer Consistency Ratio (ACR) and perplexity (PPL). ACR quantifies the cross-lingual overlap of correct solutions, formulated as $\mathrm{ACR}_l = |\mathcal{C}_{\mathrm{en}} \cap \mathcal{C}_l|/|\mathcal{Q}_l|$ for a target language l. A higher ACR indicates a greater degree of overlap in reasoning capabilities between non-English and English languages. As shown in Figure 2-(a, b), NAPO consistently achieves the highest ACR, demonstrating a substantive enhancement in underlying cross-lingual reasoning consistency. Meanwhile, PPL utilizes an off-the-shelf translation model[1] to measure textual consistency by evaluating how predictably a target-language response translates into its English counterpart, where lower perplexity indicates better alignment between the reasoning processes. As shown in Figure 2-(c, d), NAPO yields the second-lowest PPL behind MAPO, which explicitly optimizes translation-based signals. However, MAPO's inferior final accuracy (Table 1) implies that PPL over-indexes on superficial textual overlap, highlighting that our proposed activation-based preferences provide more robust, noise-resilient supervision.

4.5 Case Study

By analyzing cases that are answered incorrectly before NAPO but corrected after NAPO, we seek to understand why activation-based preference improves multilingual reasoning. As illustrated in Table 4, for the Chinese example, the base model incorrectly adds the newly arrived birds to both birds and storks, and thus predicts the wrong answer. In contrast, after NAPO, the model produces the correct reasoning and answer. Notably, the activation similarity between the Chinese and English questions also increases from 0.9076 to 0.9244. Consistent with this observation, when pooled over all languages, correctly solved examples exhibit higher question-level activation similarity than incorrect ones for both the base model (0.9206 vs. 0.9088) and NAPO (0.9267 vs. 0.9155). Since this similarity is computed from question activations rather than generated responses, the result suggests that NAPO improves cross-lingual alignment at the input representation level, making the Chinese question evoke an internal state that is closer to its English counterpart.

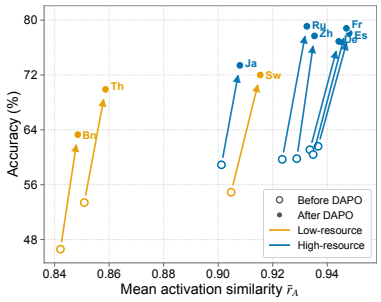
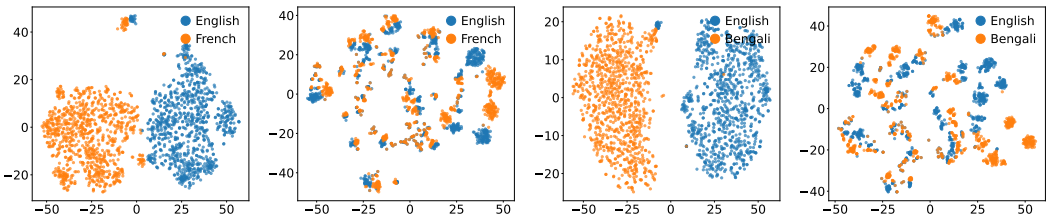


Figure 3: Relation between Activation Similarity and Accuracy.

Similar evidence can also be observed at the language level. As shown in Figure 3, languages with higher mean activation similarity generally achieve higher reasoning accuracy. We observe a strong

[1] https://huggingface.co/facebook/nllb-200-distilled-600M



(a) Embedding (Fr) (b) NAPO (Fr) (c) Embedding (Bn) (d) NAPO (Bn)

Figure 4: t-SNE visualization of multilingual alignment.

MSVAMP	Bn	Th	Sw	Ja	Zh	Ru	De	Es	Fr	En	Lrl.	Hrl.	Avg.
Qwen2.5-3B-Instruct [2024]	28.4	46.1	12.6	54.5	51.9	57.5	65.9	71.3	67.4	80.8	29.0	64.2	53.7
+NAPO (Ours)	**34.4**	**50.9**	**21.2**	**60.9**	**61.0**	**68.3**	**75.4**	**76.4**	**77.0**	**84.4**	**35.5**	**71.9**	**61.0**
Llama3.1-8B-Instruct [2024]	55.5	51.8	40.3	30.0	45.1	30.9	53.2	41.9	50.1	74.8	49.2	46.6	47.4
+NAPO (Ours)	**58.9**	**55.5**	**64.9**	**68.3**	**72.4**	**56.9**	**67.3**	**72.9**	**71.0**	**80.0**	**59.8**	**69.8**	**66.8**

Table 5: Accuracy (%) on MSVAMP across base models.

positive association between mean activation similarity and language-level accuracy. Moreover, most languages move jointly toward higher activation similarity and higher accuracy after NAPO. Taken together, these observations suggest that better cross-lingual alignment in question-level activations is closely associated with stronger multilingual reasoning performance.

4.6 Multilingual Representation Visualization

To investigate whether NAPO induces a more language-shared internal representation space, we visualize multilingual reasoning representations on the MSVAMP test set. Specifically, we extract and compare two types of representations: sentence embeddings derived from the hidden states of a middle layer (e.g., the 19th layer), and activation-pattern representations computed over the shared neurons identified in our NAPO framework (Section 3.1). We then project both representations into a two-dimensional space using t-SNE to analyze their cross-lingual clustering behaviors. As shown in Figure 4, NAPO yields a better-aligned representation space, effectively reducing language-induced separation for semantically equivalent reasoning instances across both high- and low-resource languages. These results demonstrate that the proposed activation-based reward operates within a highly language-shared internal space, mitigating the impact of surface-level textual variations.

4.7 Generalization across Base Models

To verify the generalization of NAPO across different base models, we conduct experiments with Qwen2.5-3B-Instruct Yang et al. (2024) and Llama3.1-8B-Instruct Meta Llama (2024), which differ in scale, model architecture, and training recipe. The results, presented in Table 5, show that NAPO consistently improves multilingual reasoning performance on all benchmarks. On MSVAMP, the gains are from 53.7 to 61.0 and from 47.4 to 66.8, respectively. These results suggest that our neuron activation-based preference optimization is not tied to a specific backbone model and generalizes well across model families.

5 Conclusion

In this paper, we propose a Neuron Activation-based Preference Optimization (NAPO) framework for multilingual mathematical reasoning. NAPO constructs multilingual preference signals from internal neuron activation patterns, rather than relying on external supervision such as translation models or English-centric reward models. It then measures the alignment between target-language responses and their English reference responses. This produces a more intrinsic and interpretable signal for constructing multilingual preference data. The results show that NAPO consistently outperforms multiple strong preference optimization-based baselines.

AI USE STATEMENT

In this work, large language models (LLMs) were used solely during the final stages of the writing process, specifically for tasks such as proofreading and improving linguistic clarity. Their use was limited to enhancing the readability, fluency, and grammatical correctness of the manuscript, ensuring the effective communication of our ideas. Crucially, LLMs were not involved in any core aspects of the research itself, including the formulation of methodology, experimental design, or result interpretation. We take full responsibility for the content presented in this paper.

REPRODUCIBILITY STATEMENT

We provide a comprehensive description of the proposed method in section 3, with detailed implementation and datasets specifics provided in section 4.1 and section A .

REFERENCES

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. In Neele Falk, Sara Papi, and Mike Zhang (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 225–237, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-srw.17. URL https://aclanthology.org/2024.eacl-srw.17/.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7001–7016, 2024.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Chongxuan Huang, Yongshi Ye, Biao Fu, Qifeng Su, and Xiaodong Shi. From neurons to semantics: Evaluating cross-linguistic alignment capabilities of large language models via neurons alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28956–28974, 2025.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. Mindmerger: Efficiently boosting llm reasoning in non-english languages. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 34161–34187. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3bf80b34f731313b8292f4578e820c90-Paper-Conference.pdf.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Meta Llama. Meta-llama-3.1-8b-instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, 2024. Hugging Face model card.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3505–3523, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.246. URL https://aclanthology.org/2022.acl-long.246/.

Long Ouyang, Jeffrey 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 F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2695–2709, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.163. URL https://aclanthology.org/2023.emnlp-main.163/.

Daking Rai and Ziyu Yao. An investigation of neuron activation as a unified lens to explain chain-of-thought eliciting arithmetic reasoning of LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7174–7193, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.387. URL https://aclanthology.org/2024.acl-long.387/.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 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*, KDD '20, pp. 3505–3506, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3406703. URL https://doi.org/10.1145/3394486.3406703.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. Mapo: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10015–10027, 2024.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=fR3wGCk-IXp.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. 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 (Volume 1: Long Papers)*, pp. 5701–5715, 2024.

Yiru Tang, Kun Zhou, Yingqian Min, Wayne Xin Zhao, Jing Sha, Zhichao Sheng, and Shijin Wang. Enhancing chain-of-thought reasoning via neuron activation differential analysis. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16151–16159, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.817. URL https://aclanthology.org/2025.emnlp-main.817/.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D Nguyen. Lacomsa: Language-consistency multilingual self-alignment with latent representation rewarding. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4839–4853, 2026.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. Neurons in large language models: Dead, n-gram, positional. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1288–1301, 2024.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning, 2020. URL https://github.com/huggingface/trl.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. Implicit cross-lingual rewarding for efficient multilingual preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21125–21147, 2025a.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. Language imbalance driven rewarding for multilingual self-improving. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=Kak2ZH5Itp.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=N8N0hgNDRt.

Zeping Yu and Sophia Ananiadou. Interpreting arithmetic mechanism in large language models through comparative neuron analysis. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3293–3306, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.193. URL https://aclanthology.org/2024.emnlp-main.193/.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. Scaling relationship on learning mathematical reasoning with large language models. *CoRR*, abs/2308.01825, 2023. doi: 10.48550/ARXIV.2308.01825. URL https://doi.org/10.48550/arXiv.2308.01825.

Shimao Zhang, Zhejian Lai, Xiang Liu, Shuaijie She, Xiao Liu, Yeyun Gong, Shujian Huang, and Jiajun Chen. How does alignment enhance llms' multilingual capabilities? a language neurons perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34799–34807, 2026.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. Cm-align: Consistency-based multilingual alignment for large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 25689–25702. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-emnlp.1401/.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. Question translation training for better multilingual reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8411–8423, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.498. URL https://aclanthology.org/2024.findings-acl.498/.

Iteration	Bn	Th	Sw	Ja	Zh	Ru	De	Es	Fr	En	Lrl.	Hrl.	Avg.
Iter1	60.8	69.2	70.5	71.3	75.0	74.7	74.1	76.8	75.6	78.6	66.8	75.1	72.6
Iter2	64.0	68.6	72.6	73.1	77.0	76.6	78.0	78.2	78.2	80.7	68.4	77.4	74.7
Iter3	63.3	69.9	72.0	73.4	77.7	76.9	78.8	79.1	78.1	80.6	68.4	77.8	75.0

Table 6: Performance of NAPO across three iterations on MSVAMP.

Iteration	Bn	Th	Sw	Ja	Zh	Ru	De	Es	Fr	En	Lrl.	Hrl.	Avg.
Iter1	52.9	57.6	60.6	63.5	70.6	63.3	59.9	69.5	63.5	75.1	57.0	66.5	63.7
Iter2	64.0	66.5	65.5	71.6	75.5	73.6	74.6	76.6	77.2	79.7	65.3	75.5	72.5
Iter3	66.1	69.5	69.7	74.2	76.6	75.7	75.7	78.7	78.0	80.8	68.4	77.1	74.5

Table 7: Performance of NAPO across three iterations on MNumGLUESub.

MNumGLUESub	Bn	Th	Sw	Ja	Zh	Ru	De	Es	Fr	En	Lrl.	Hrl.	Avg.
Qwen2.5-3B-Instruct [2024]	25.2	38.2	14.9	41.8	49.2	46.5	36.9	61.8	43.3	65.5	26.1	49.3	42.3
+NAPO (Ours)	**27.1**	**43.5**	**18.8**	**52.4**	**57.8**	**57.6**	**55.9**	**68.9**	**56.1**	**75.3**	**29.8**	**60.6**	**51.4**
Llama3.1-8B-Instruct [2024]	52.0	48.4	38.8	31.5	48.4	22.8	35.4	39.5	35.2	58.6	46.4	38.8	41.1
+NAPO (Ours)	**56.9**	**50.8**	**62.1**	**61.8**	**71.0**	**52.2**	**51.2**	**70.6**	**57.4**	**73.6**	**56.6**	**62.5**	**60.8**
MGSM	**Bn**	**Th**	**Sw**	**Ja**	**Zh**	**Ru**	**De**	**Es**	**Fr**	**En**	**Lrl.**	**Hrl.**	**Avg.**
Qwen2.5-3B-Instruct [2024]	9.2	26.4	**4.0**	34.4	41.2	44.8	46.4	60.4	47.2	68.8	13.2	49.0	38.3
+NAPO (Ours)	**18.0**	**34.4**	3.6	**42.0**	**49.2**	**59.2**	**59.6**	**66.8**	**59.2**	**77.2**	**18.7**	**59.0**	**46.9**
Llama3.1-8B-Instruct [2024]	47.6	**41.2**	34.8	18.0	34.4	18.0	42.8	40.4	35.6	68.4	41.2	36.8	38.1
+NAPO (Ours)	**56.8**	**41.2**	**57.2**	**53.6**	**66.4**	**43.6**	**56.4**	**73.2**	**64.8**	**82.8**	**51.7**	**63.0**	**59.6**

Table 8: Accuracy (%) on MNumGLUESub and MGSM across base models.

A Training Details

Sampling Based on the VLLM framework, we set the sampling temperature to 0.3 and the number of samples to 20.

DPO For MistralMathOctopus-7B, we employ a learning rate of 1e-7, with β set to 0.1, and a warmup step count of 100. The batch size is configured to 128. The optimization process is capped at a maximum of 1000 steps, and we save the checkpoint that corresponds to the lowest loss on the validation set. For Qwen and LLaMA 3.1, we increase the learning rate to 1e-6,as these models have not undergone supervised fine-tuning prior to preference optimization.

Neuron Filtering Hyperparameters Based on GSM8KInstruct, we searched over multiple hyperparameter configurations. We find that when the value of r is relatively small, specifically $r \leq 0.4$, we achieve relatively stable performance. An excessively large r causes performance collapse in later training, whereas a smaller r preferentially selects neurons that capture general-purpose patterns beyond the GSM8KInstruct dataset.

Intermediate Layer Selection We pre-compute the cosine similarity of hidden states between each non-English language and English on parallel sentence pairs. After excluding the first and last layers, we calculate the average cosine similarity, and select the layers whose similarity is greater than or equal to this average as the intermediate layers.

B Supplementary Results

In Tables 6 and 7, we report the three-round results of NAPO on MSVAMP and MNumGLUESub, respectively. Table 8 presents the results of NAPO using Qwen2.5-3B-Instruct and Llama3.1-8B-Instruct.

C Different Neuron Pooling Methods

We evaluate four pooling methods on the MGSM test set: position-weighted pooling, max pooling, last-token pooling, and mean pooling. As shown in Figure 5, our proposed position-weighted pooling yields the strongest representation of cross-lingual similarity structure. Specifically, it achieves higher similarity between English and target-language representations than the other methods in the intermediate layers, while producing the lowest similarity in the early and final layers (excluding the last layer). This pattern aligns well with the expected behavior of general-purpose versus language-specific neurons across layers.

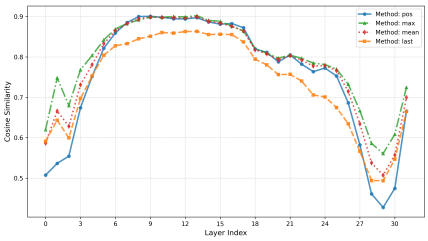


Figure 5: Comparison of four different neuron pooling methods.