DIRECTING SAFETY SIGNALS TO THE APPROPRIATE NEURONS: MULTILINGUAL SAFETY ALIGNMENT THROUGH SELECTIVE NEURON MASKING

Anonymous authors
Paper under double-blind review

ABSTRACT

The safety behavior of large language models varies considerably from one language to another, and prior work has traced such responses to a compact group of internal safety neurons. Current approaches to multilingual safety alignment usually formulate their optimization targets behaviorally, through loss terms or reward signals, yet seldom determine which of the model's safety-related internal representations actually absorb this training pressure. To address this gap, we introduce **SNPO**, a preference optimization scheme that operates at the level of individual neurons for aligning safety across languages. The method begins by locating the safety neurons tied to each language and then divides all languages into two sets according to how much training data each has. Optimization proceeds by alternating between these sets: on any given step, the safety-neuron activations of one set are masked and their weights held fixed, so that the alignment pressure is funneled onto the neurons belonging to the other set. The assignments are subsequently reversed, allowing the neurons of each set to receive focused optimization in turn. Evaluations spanning seven models and eight languages demonstrate that **SNPO** lowers how often unsafe outputs occur without sacrificing multilingual capability, and the improvements are most pronounced for languages with limited resources. Examining the neurons directly, we find that after **SNPO** those safety neurons that had responded to only one language begin firing across several. Taken together, our findings indicate that successful multilingual safety alignment depends not merely on carefully crafted behavioral preferences, but equally on directly governing which internal neurons end up receiving the resulting signal.

1 INTRODUCTION

Large language models (LLMs) exhibit conspicuously uneven safety behaviors across languages (Achiam et al., 2023; Yang et al., 2025; Wang et al., 2023; Shen et al., 2024). A model that reliably refuses harmful instructions in high-resource languages, such as English, may comply with the same instructions when they are rendered in a low-resource language; indeed, translation-based attacks remain among the most effective jailbreak vectors against advanced LLMs (Wang et al., 2024; Chander et al., 2025; Yi et al., 2024; Wang et al., 2025). Cross-lingual safety vulnerabilities therefore constitute an exploitable attack surface, allowing adversaries to bypass model safety filters and elicit socially harmful instructions, such as instructions for illegal or dangerous activities. Moreover, such vulnerabilities tend to become more severe as language resource levels decrease (Yi et al., 2024; Ji et al., 2023; Kaddour et al., 2023).

To mitigate this gap, existing multilingual safety alignment methods primarily optimized at the level of model behavior. Supervised fine-tuning (SFT) (Wei et al., 2021), reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022), direct preference optimization (DPO) (Rafailov et al., 2023), and related preference-optimization methods (Shen et al., 2024; Meng et al., 2024; Zhao et al., 2025a) design losses or reward signals over model outputs, and propagate the resulting alignment signal through gradient updates to the model parameters. These methods primarily focus on *what signal* should be used for alignment, while leaving implicit the question of *which parameters* should carry that signal. Since the update is applied over a broad trainable

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.

parameter space, careful KL constraints or regularization terms are often required to avoid degrading general capabilities (Rafailov et al., 2023).

Recent mechanistic interpretability studies provide a more fine-grained view of model safety (Wang et al., 2022; Marks et al., 2024). They show that safety behavior is not uniformly distributed across the model, but is instead concentrated in a small subset of parameters or internal units (Chen et al., 2024; Zhao et al., 2025c; Wu et al., 2025). Perturbing, ablating, or pruning these safety-critical units can substantially weaken refusal behavior, while having a comparatively limited impact on general model capabilities. These units are commonly referred to as *safety neurons*, as they serve as important internal carriers of safety-related behavior. Motivated by this finding, recent neuron-level alignment methods (Zhao et al., 2025c; Zhang et al., 2026) attempt to route safety alignment signals through safety neurons, achieving considerable safety improvements while better preserving general utility. These results further demonstrate that safety neurons play a causal role in controlling model safety behavior. They also suggest that explicitly controlling the internal routing of alignment signals can improve the efficiency of safety alignment.

These neuron-level methods provide a preliminary answer to the problem of *which parameters*, yet still rely on the model's pre-existing safety capacity as the carrier of the alignment signal. However, in multilingual settings, the pre-existing safety capacity itself is highly imbalanced across languages (Shen et al., 2024; Achiam et al., 2023). Prior studies have also shown that both the number and distribution of safety neurons differ substantially between high- and low-resource languages (Zhang et al., 2026; Liang et al., 2026). Therefore, even if the alignment signal is routed to safety neurons, existing methods do not explicitly handle the cross-lingual imbalance in the carriers that receive this signal.

Furthermore, prior work has found that safety fine-tuning yields substantially different gains for low-resource and high-resource languages (Shen et al., 2024), indicating that the model exhibits structurally different alignment responses to high- and low-resource languages. This implies that even with the same amount of multilingual alignment data, treating high- and low-resource languages identically cannot eliminate the uneven safety. Low-resource languages may remain alignment bottlenecks not because the safety signal is absent, but because the internal carriers available to receive that signal are insufficient. We therefore ***argue*** that multilingual safety alignment requires differentiated signal routing at the safety-neuron level, so that the alignment process can actively reshape how safety capabilities are carried across languages.

Based on this insight, we propose SNPO (*Safety-Neuron Preference Optimization*), a neuron-level preference optimization method that operationalizes differentiated signal routing across high- and low-resource languages. In SNPO, we first identify language-specific safety neurons in the base model and partition target languages into two groups according to their training-corpus resource level. Alignment then proceeds in an alternating manner between the two groups. At each step, we mask the safety-neuron activations of one group and freeze their corresponding parameters, amplifying the alignment signal and forcing it to concentrate on the complementary group's safety neurons. The two groups then swap roles in the next step, so that each group's safety neurons receive a dedicated optimization window rather than passively inheriting updates from the model's pre-existing, imbalanced safety capability. Through this alternating process, SNPO converts multilingual preference optimization from an undifferentiated update over all safety neurons into a controlled, group-wise allocation procedure, breaking the structural alignment bottleneck of low-resource languages while preserving the safety capability of high-resource ones. Empirically, this alternating process not only improves multilingual safety alignment but also induces a measurable reorganization of safety-neuron language attribution, with language-specific neurons partially evolving into shared ones. The main contributions of this work are as follows:

- We identify a limitation of existing neuron-level safety alignment methods. They route alignment signals to safety neurons but do not address the cross-lingual imbalance in safety-neuron carriers, which particularly limits low-resource language alignment.
- We propose SNPO, which alternately masks language-group-specific safety neurons during DPO training to amplify and redirect preference gradients and enable differentiated signal allocation across high- and low-resource languages.
- Experiments on seven models across eight languages show that SNPO improves multilingual safety over strong baselines, especially on low-resource languages, while preserving general

capability. Further analysis shows that SNPO partially reorganizes language attribution of safety-neuron, with some language-specific neurons becoming shared across languages.

2 Related Work

2.1 Multilingual Safety Alignment

Multilingual safety alignment aims to reduce safety disparities across languages, particularly between high- and low-resource languages. Due to uneven safety supervision, LLMs are often better aligned in English and other high-resource languages, while low-resource languages remain more vulnerable to harmful instructions and cross-lingual jailbreaks (Wang et al., 2023; Shen et al., 2024; Yang et al., 2025; Achiam et al., 2023; Team et al., 2024; Grattafiori et al., 2024; Etxaniz et al., 2024). Recent studies such as LinguaSafe (Ning et al., 2025) and CultureGuard (Joshi et al., 2025) further reveal substantial linguistic and cultural biases in safety perception.

One line of work improves multilingual safety through training-time alignment using multilingual or translated safety data with SFT, RLHF, or preference optimization (Banerjee et al., 2025). AlignX (Bu et al., 2025) aligns multilingual representations before instruction tuning, MPO (Zhao et al., 2025a) reduces reward distribution discrepancies across languages, and AdaMergeX (Zhao et al., 2025b) promotes cross-lingual transfer via adaptive adapter merging. While effective, these methods mainly optimize output behavior without explicitly controlling where safety-related updates are absorbed within the model.

Another line of work applies inference-time or post-hoc interventions. Translation-based methods map non-English inputs into English to leverage stronger English safety behavior (Artetxe et al., 2023; Etxaniz et al., 2024; Bu et al., 2026), while BridgeX-ICL exploits bridge languages and linguistic overlap neurons (Xu et al., 2025). Other approaches directly modify internal representations or weights, such as RESTA (Bhardwaj et al., 2024) and Sparse Weight Editing (Liang et al., 2026). These methods demonstrate that multilingual safety can be transferred through targeted internal interventions, but remain largely decoupled from preference optimization and do not directly address how safety signals should be allocated during training.

2.2 Safety Neurons and Multilingual Safety Representations

Recent mechanistic studies suggest that safety behavior in LLMs is concentrated in a small subset of safety-critical units, often termed *safety neurons*. Suppressing or ablating these neurons can weaken refusal behavior, while preserving or strengthening them can improve safety with limited utility degradation (Chen et al., 2024; Wu et al., 2025; Zhao et al., 2025c; Wang et al., 2026). Such neurons are commonly identified through activation contrast, probing, masking, or ablation, and validated by their causal effects on safety behavior (Chen et al., 2024; Wu et al., 2025; Zhao et al., 2025c). Neuron-level alignment methods such as SN-Tune and Transfer-SN further exploit these units to achieve more targeted safety alignment.

In multilingual settings, recent studies compare language-specific and shared safety neurons across languages (Liang et al., 2026; Zhang et al., 2026), showing that different languages rely on partially overlapping yet distinct neuron subsets. Their number, distribution, and transferability can vary substantially across languages. This motivates our work: existing neuron-level methods improve targeted safety alignment, but do not explicitly address cross-lingual imbalance in the underlying safety-neuron carriers. We therefore focus on differentiated safety-signal allocation across language-specific neurons during preference optimization.

3 The Proposed Method

3.1 Overview

Our goal is to improve multilingual safety alignment by controlling not only *what* safety objective is optimized, but also *which* parameters receive the resulting safety signal. Standard full-parameter DPO backpropagates the safety gradient through the entire trainable parameter space. In a multilingual setting, this allows the safety signal to concentrate on parameters that already encode safety behavior

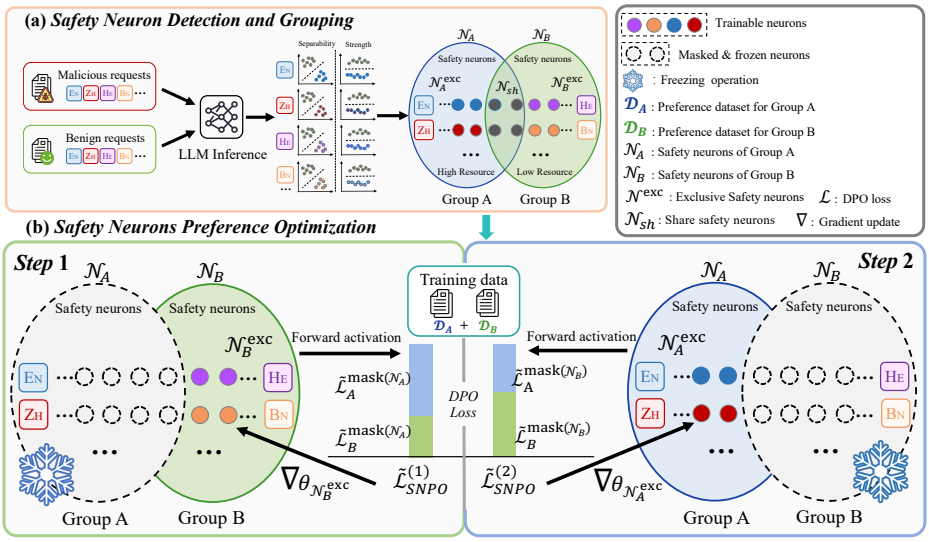

Figure 1: Overview of SNPO. The method first detects language-specific safety neurons and partitions them into shared and group-exclusive subsets. During training, SNPO alternates between masking one group's safety neurons and updating the exclusive safety neurons of the other group.

for high-resource languages, leaving parameters tied to other languages comparatively under-updated. We therefore use detected safety neurons as explicit, controllable carriers of the safety signal, so that the optimization can be redirected toward parameters where multilingual safety is most needed.

The proposed framework, illustrated in Figure 1, consists of two components. **(a) Safety neuron detection and grouping**. We identify language-specific safety neurons by comparing model activations on harmful and harmless prompts, then partition the training languages into two groups, A and B. Within each group, we further distinguish shared safety neurons (detected in both groups) from exclusive ones (detected in only one group). **(b) Safety Neurons Preference Optimization.** We then perform a two-step symmetrical DPO procedure. In each step, we mask one group's safety-neuron activations during the forward pass and update only the exclusive safety neurons of the other group. Masking deliberately increases the safety loss, while restricting the trainable parameters forces this signal to be received by the selected exclusive neurons. After training, all neurons are restored for inference.

3.2 Safety Neurons Detection and Grouping

For each language $\ell \in \mathcal{S}$, we follow Wang et al. (2026) to obtain a language-specific safety-neuron set $\mathcal{N}_\ell \subseteq \mathcal{N}_{\text{all}}$, where $\mathcal{N}_{\text{all}}$ enumerates every neuron in the MLP `gate_proj` and `up_proj` sublayers, indexed by layer, projection type, and intermediate dimension. On a paired probing set of harmful and harmless prompts in language ℓ, neurons are scored along two criteria, i.e., activation *separability* and activation gap *strength* (Figure 1a), and those passing both thresholds form $\mathcal{N}_\ell$.

We split $\mathcal{S}$ into two groups based on their training corpus resource level, where Group A consists of HRLs and Group B consists of LRLs.

$$A \cap B = \varnothing,\ A \cup B = \mathcal{S}, |A| = |B|, \tag{1}$$

We then define the safety-neuron set of each group as the union of the language-specific safety neurons within that group:

$$\mathcal{N}_A = \bigcup_{\ell \in A} \mathcal{N}_\ell, \qquad \mathcal{N}_B = \bigcup_{\ell \in B} \mathcal{N}_\ell. \tag{2}$$

Thus, $\mathcal{N}_A$ denotes the safety neurons associated with the group A, while $\mathcal{N}_B$ denotes the safety neurons associated with the group B.

Although A and B are disjoint language groups, their corresponding safety-neuron sets are not necessarily disjoint. We therefore decompose $\mathcal{N}_A$ and $\mathcal{N}_B$ into their shared and group-exclusive

components:

$$\mathcal{N}_{\text{sh}} = \mathcal{N}_A \cap \mathcal{N}_B, \qquad \mathcal{N}_A^{\text{exc}} = \mathcal{N}_A \setminus \mathcal{N}_B, \qquad \mathcal{N}_B^{\text{exc}} = \mathcal{N}_B \setminus \mathcal{N}_A. \tag{3}$$

Here, $\mathcal{N}_{\text{sh}}$ denotes the shared safety neurons used by both groups, whereas $\mathcal{N}_A^{\text{exc}}$ and $\mathcal{N}_B^{\text{exc}}$ denote the safety neurons exclusive to the A and B groups, respectively. Notably, the shared neurons $\mathcal{N}_{\text{sh}}$ carry a substantial portion of the cross-lingual common safety capability (Zhang et al., 2026). If they still remained active during SNPO, the masked group would still retain non-trivial safety behavior through $\mathcal{N}_{\text{sh}}$. We therefore exclude $\mathcal{N}_{\text{sh}}$ from both forward propagation and parameter updates during training, and restore it only at inference.

3.3 Safety Neurons Preference Optimization

Let $\mathcal{D} = (\mathcal{D}_A, \mathcal{D}_B)$ denote the multilingual preference dataset written in language from Group A and B, where each sample consists of a prompt x, a preferred safe response y^+, and a rejected unsafe response y^-. For a sample (x_i, y_i^+, y_i^-), the standard DPO preference margin is

$$\Delta_i(\theta) = \left[\log \frac{\pi_\theta(y_i^+|x_i)}{\pi_{\text{ref}}(y_i^+|x_i)} - \log \frac{\pi_\theta(y_i^-|x_i)}{\pi_{\text{ref}}(y_i^-|x_i)}\right], \tag{4}$$

where π_θ is the policy distribution induced by the model, π_{ref} is the reference policy, and β controls the strength of the preference constraint. The standard DPO loss is then

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x_i, y_i^+, y_i^-)\sim\mathcal{D}} \left[\log \sigma\left(\beta \Delta_i(\theta)\right)\right]. \tag{5}$$

In full-parameter DPO, the resulting gradients are propagated through all trainable parameters. In our method, the same behavioral objective is used, but the trainable parameter set and the available safety-neuron activations are explicitly controlled. Therefore, the difference lies not in the definition of the preference objective, but in which neurons are allowed to contribute to the loss and which neurons are allowed to receive the resulting safety signal.

The core of SNPO is a two step training procedure. In each step, we mask one group's safety neuron activations and update only the exclusive safety neurons of the other group. The masking operation is applied only during training and is not used at inference time.

Step 1: Mask $\mathcal{N}_A$ and update $\mathcal{N}_B^{\text{exc}}$. In this step, we freeze all parameters except those associated with $\mathcal{N}_B^{\text{exc}}$. During the forward pass, we mask the activations of all neurons in $\mathcal{N}_A$. Let $\pi_\theta^{\text{mask}(\mathcal{N}_A)}$ denote the policy distribution produced by the model when $\mathcal{N}_A$ is masked. The corresponding preference margin is

$$\widetilde{\Delta}_i^{\text{mask}(\mathcal{N}_A)} = \left[\log \frac{\pi_\theta^{\text{mask}(\mathcal{N}_A)}(y_i^+|x_i)}{\pi_{\text{ref}}(y_i^+|x_i)} - \log \frac{\pi_\theta^{\text{mask}(\mathcal{N}_A)}(y_i^-|x_i)}{\pi_{\text{ref}}(y_i^-|x_i)}\right]. \tag{6}$$

We further decompose the DPO loss in Step 1 into language specific components:

$$\widetilde{\mathcal{L}}_{\text{SNPO}}^{(1)} = \widetilde{\mathcal{L}}_A^{\text{mask}(\mathcal{N}_A)} + \widetilde{\mathcal{L}}_B^{\text{mask}(\mathcal{N}_A)}, \tag{7}$$

where

$$\widetilde{\mathcal{L}}_A^{\text{mask}(\mathcal{N}_A)} = -\mathbb{E}_{(x_i, y_i^+, y_i^-)\sim\mathcal{D}_A} \left[\log \sigma\left(\beta \widetilde{\Delta}_i^{\text{mask}(\mathcal{N}_A)}\right)\right], \tag{8}$$

$$\widetilde{\mathcal{L}}_B^{\text{mask}(\mathcal{N}_A)} = -\mathbb{E}_{(x_i, y_i^+, y_i^-)\sim\mathcal{D}_B} \left[\log \sigma\left(\beta \widetilde{\Delta}_i^{\text{mask}(\mathcal{N}_A)}\right)\right]. \tag{9}$$

Because $\mathcal{N}_A$ is masked during training, the model can no longer rely on the original safety pathway associated with language group A. Consequently, for prompts written in languages from group A, the model's safety response is substantially weakened, which reduces the preference margin $\widetilde{\Delta}_i^{\text{mask}(\mathcal{N}_A)}$ and increases the corresponding sample-level loss term $-\log\sigma(\beta\widetilde{\Delta}_i^{\text{mask}(\mathcal{N}_A)})$, thereby increasing the loss $\widetilde{\mathcal{L}}_{\text{SNPO}}^{(1)}$.

In this step, only $\mathcal{N}_B^{\text{exc}}$ is trainable. The update can be written as

$$\theta_{\mathcal{N}_B^{\text{exc}}} \leftarrow \theta_{\mathcal{N}_B^{\text{exc}}} - \eta \nabla_{\theta_{\mathcal{N}_B^{\text{exc}}}} \widetilde{\mathcal{L}}_{\text{SNPO}}^{(1)}. \tag{10}$$

This step encourages $\mathcal{N}_B^{\text{exc}}$ to acquire safety-related functionality for group A languages, thereby constructing a complementary safety pathway across language groups.

Step 2: Mask $\mathcal{N}_B$ and update $\mathcal{N}_A^{\text{exc}}$. The second step follows the same mechanism as Step 1, with the roles of groups A and B exchanged. Specifically, $\mathcal{N}_B$ is masked during the forward pass, and only $\mathcal{N}_A^{\text{exc}}$ is updated:

$$\theta_{\mathcal{N}_A^{\text{exc}}} \leftarrow \theta_{\mathcal{N}_A^{\text{exc}}} - \eta \nabla_{\theta_{\mathcal{N}_A^{\text{exc}}}} \widetilde{\mathcal{L}}_{\text{SNPO}}^{(2)}, \tag{11}$$

where $\widetilde{\mathcal{L}}_{\text{SNPO}}^{(2)}$ is the DPO loss computed with $\mathcal{N}_B$ masked. Symmetrically, masking $\mathcal{N}_B$ redirects the resulting safety signal to the only trainable neurons $\mathcal{N}_A^{\text{exc}}$, encouraging the $\mathcal{N}_A^{\text{exc}}$ to absorb safety signals when the safety-neuron activations of group B are unavailable.

Together, these two symmetric steps enable the exclusive safety neurons of both groups to acquire broader multilingual safety functionality. Because the two steps update disjoint parameter subsets, they can be implemented independently, and the resulting parameter updates can be merged after training. During inference, all masked neurons are restored, and no neuron is suppressed during deployment. Note that this restoration may introduce interaction effects between the two updated groups; we verify empirically through inference-time ablations that restoring all neurons preserves or improves safety compared with keeping either group suppressed.

4 Experiments

We evaluate whether SNPO improves multilingual safety alignment across different model families, model scales, and languages. Since our method is designed to control the allocation of safety-alignment signals, we compare it with representative multilingual safety alignment and sparse editing baselines under the same safety and utility evaluation protocol. In addition, our ablation studies further examine several neuron-level training strategies and their variants (Zhao et al., 2025c; Zhang et al., 2026) to validate the design choices of SNPO.

4.1 Experimental Setup

Models. We evaluate seven instruction-tuned LLMs from two model families: LLaMA and Qwen series. Specifically, we use LLaMA-3.2-1B and 3B, Qwen2-0.5B and 1.5B, and Qwen2.5-1.5B, 3B, and 7B. These models cover different architectures and parameter scales, allowing us to examine whether the proposed method generalizes beyond a single backbone. Eight languages are chosen to represent a wide range of language families and resource levels. Following the language-resource groupings adopted in prior work (Shen et al., 2024; Zhang et al., 2026; Xu et al., 2025), we categorize English (En), Chinese (Zh), Vietnamese (Vi), and Indonesian (Id) as high-resource languages, and Japanese (Ja), Thai (Th), Bengali (Bn), and Hebrew (He) as low-resource languages in this work. Thus, in this experiment, the high-resource Group A includes $\{\text{En}, \text{Zh}, \text{Vi}, \text{Id}\}$, while the low-resource Group B includes $\{\text{Ja}, \text{Th}, \text{Bn}, \text{He}\}$.

Training Data. The alignment data is drawn from PKU-SafeRLHF (Ji et al., 2025), which includes 83.4k English preference pairs annotated for both helpfulness and harmlessness. We randomly sample 10k pairs from this corpus and translate them into the seven target languages using `tencent/Hunyuan-MT-7B` (Zheng et al., 2025). After a manual rough inspection of the resulting dataset, we retain 5,447 samples per language, discarding pairs that contain obvious errors such as special characters (e.g., #, %, *) and repeated characters. We acknowledge that the constructed dataset has quality limitations and may still contain mistakes, however, we retain such samples to reflect imperfect real-world data conditions.

Evaluation. To evaluate the effectiveness of SNPO, we assess both model safety behavior and task performance. For safety, we employ Multi-StrongREJECT (Liang et al., 2026), built upon `walledai/StrongREJECT` (Souly et al., 2024) by translating English harmful queries into seven languages. Each language subset contains 313 harmful queries designed to probe safety vulnerabilities. Attack Success Rate (ASR) is reported as the safety metric, which count how often the model produces harmful outputs when queried with malicious prompts. To measure general utility, we adopt MGSM (Shi et al., 2022) and M-MMLU (Hendrycks et al., 2020) to evaluate the model's cross-lingual reasoning and multilingual general knowledge, respectively. The average accuracy across the different tasks of each benchmark is reported as an overall utility metric.

Baselines. We compare SNPO with two recent representative multilingual safety alignment baselines. The first baseline is the original instruction-tuned model without additional safety intervention, denoted as **None**. **MPO** (2025) is a multilingual preference optimization method that encourages safety transfer by reducing reward discrepancies between English and target languages. **SWE** (2026) is a sparse weight editing method that improves multilingual safety by modifying safety-related internal activation weights. All methods are evaluated under the same multilingual safety and utility evaluation protocol.

Table 1: **Zero-shot multilingual safety and utility evaluation.** Safety is measured by the number of unsafe responses identified by `Qwen3Guard-Gen-8B` among 313 harmful prompts, where lower values indicate better safety. Superscripts denote changes in unsafe-response counts relative to **None** under the same backbone: negative values indicate fewer unsafe responses, whereas positive values indicate more unsafe responses. Δ_{Avg} reports the average change across all evaluated languages. Utility is measured by MGSM and M-MMLU accuracy, where higher values are better.

Models	Method	Safety									Utility	
		ASR↓ (#Unsafe / 313)										
		En	Zh	Vi	Ja	Th	Id	Bn	He	Δ_{Avg}	MGSM↑	M-MMLU↑
LLAMA-3.2-1B	None	1/313	30/313	8/313	104/313	37/313	69/313	194/313	167/313	-	20.18	27.20
	MPO	$1/313^{0}$	$18/313^{-12}$	$7/313^{-1}$	$65/313^{-39}$	$32/313^{-5}$	$64/313^{-5}$	$185/313^{-9}$	$184/313^{+17}$	-6.75	21.42	26.71
	SWE	$2/313^{+1}$	$12/313^{-18}$	$3/313^{-5}$	$85/313^{-19}$	$13/313^{-24}$	$16/313^{-53}$	$159/313^{-35}$	$149/313^{-18}$	-21.38	18.96	27.70
	Ours	$1/313^{0}$	$12/313^{-18}$	$5/313^{-3}$	$60/313^{-44}$	$11/313^{-26}$	$45/313^{-24}$	$150/313^{-44}$	$107/313^{-60}$	-27.38	21.29	26.87
LLAMA-3.2-3B	None	4/313	11/313	11/313	76/313	23/313	20/313	119/313	58/313	-	37.22	39.35
	MPO	$4/313^{0}$	$6/313^{-5}$	$7/313^{-4}$	$51/313^{-25}$	$17/313^{-6}$	$15/313^{-5}$	$88/313^{-31}$	$41/313^{-17}$	-11.62	36.67	39.08
	SWE	$2/313^{-2}$	$3/313^{-8}$	$2/313^{-9}$	$41/313^{-35}$	$5/313^{-18}$	$3/313^{-17}$	$67/313^{-52}$	$59/313^{+1}$	-17.5	33.91	39.22
	Ours	$0/313^{-4}$	$7/313^{-4}$	$6/313^{-5}$	$34/313^{-42}$	$9/313^{-14}$	$11/313^{-9}$	$57/313^{-62}$	$17/313^{-41}$	-22.62	36.98	39.46
Qwen2-0.5B	None	239/313	198/313	203/313	201/313	184/313	208/313	235/313	185/313	-	6.31	33.50
	MPO	$132/313^{-107}$	$94/313^{-104}$	$92/313^{-111}$	$106/313^{-95}$	$58/313^{-126}$	$96/313^{-112}$	$177/313^{-58}$	$127/313^{-58}$	-96.38	5.05	33.53
	SWE	$183/313^{-56}$	$147/313^{-51}$	$149/313^{-54}$	$173/313^{-28}$	$152/313^{-32}$	$181/313^{-27}$	$211/313^{-24}$	$161/313^{-24}$	-37	5.45	31.57
	Ours	$8/313^{-231}$	$3/313^{-195}$	$1/313^{-202}$	$1/313^{-200}$	$5/313^{-179}$	$10/313^{-198}$	$31/313^{-204}$	$9/313^{-176}$	-198.12	4.64	32.01
Qwen2-1.5B	None	36/313	25/313	45/313	76/313	60/313	58/313	232/313	90/313	-	20.36	42.98
	MPO	$0/313^{-36}$	$1/313^{-24}$	$0/313^{-45}$	$3/313^{-73}$	$3/313^{-57}$	$1/313^{-57}$	$30/313^{-202}$	$4/313^{-86}$	-72.5	18.93	42.83
	SWE	$8/313^{-28}$	$2/313^{-23}$	$19/313^{-26}$	$28/313^{-48}$	$12/313^{-48}$	$12/313^{-46}$	$183/313^{-49}$	$48/313^{-42}$	-38.75	20.33	42.98
	Ours	$0/313^{-36}$	$0/313^{-25}$	$0/313^{-45}$	$1/313^{-75}$	$1/313^{-59}$	$0/313^{-58}$	$3/313^{-229}$	$0/313^{-90}$	-77.12	14.89	41.12
Qwen2.5-1.5B	None	60/313	35/313	36/313	63/313	81/313	60/313	210/313	159/313	-	26.55	42.85
	MPO	$3/313^{-57}$	$2/313^{-33}$	$2/313^{-34}$	$5/313^{-58}$	$8/313^{-73}$	$2/313^{-58}$	$67/313^{-143}$	$29/313^{-130}$	-73.25	24.27	41.91
	SWE	$18/313^{-42}$	$8/313^{-27}$	$14/313^{-22}$	$17/313^{-46}$	$23/313^{-58}$	$23/313^{-37}$	$195/313^{-15}$	$84/313^{-75}$	-40.25	25.00	43.29
	Ours	$2/313^{-58}$	$3/313^{-32}$	$1/313^{-35}$	$3/313^{-60}$	$4/313^{-77}$	$3/313^{-57}$	$24/313^{-186}$	$24/313^{-135}$	-80	26.09	42.17
Qwen2.5-3B	None	60/313	57/313	65/313	89/313	58/313	61/313	179/313	122/313	-	33.16	48.55
	MPO	$19/313^{-41}$	$14/313^{-43}$	$16/313^{-49}$	$31/313^{-58}$	$26/313^{-32}$	$18/313^{-43}$	$68/313^{-111}$	$33/313^{-89}$	-58.25	36.85	47.44
	SWE	$15/313^{-45}$	$5/313^{-52}$	$10/313^{-55}$	$16/313^{-73}$	$23/313^{-35}$	$15/313^{-46}$	$143/313^{-36}$	$52/313^{-70}$	-51.5	31.47	46.17
	Ours	$4/313^{-56}$	$6/313^{-51}$	$4/313^{-61}$	$3/313^{-86}$	$5/313^{-53}$	$6/313^{-55}$	$38/313^{-141}$	$19/313^{-103}$	-75.75	34.64	48.57
Qwen2.5-7B	None	17/313	19/313	23/313	35/313	28/313	21/313	118/313	62/313	-	33.91	50.55
	MPO	$7/313^{-10}$	$3/313^{-16}$	$6/313^{-17}$	$9/313^{-26}$	$11/313^{-17}$	$7/313^{-14}$	$27/313^{-91}$	$21/313^{-41}$	-29	40.02	48.14
	SWE	$5/313^{-12}$	$2/313^{-17}$	$7/313^{-16}$	$14/313^{-21}$	$16/313^{-12}$	$6/313^{-15}$	$68/313^{-50}$	$24/313^{-38}$	-22.62	33.07	50.36
	Ours	$0/313^{-17}$	$0/313^{-19}$	$0/313^{-23}$	$0/313^{-35}$	$1/313^{-27}$	$0/313^{-21}$	$12/313^{-106}$	$6/313^{-56}$	-38	42.95	53.19

4.2 Main Results

Table 1 shows that SNPO consistently improves multilingual safety across all evaluated model families and scales. Averaged over the seven models, SNPO reduces the unsafe count by 74.14, compared with 49.68 for MPO and 32.71 for SWE. This demonstrates that explicitly controlling the allocation of safety-alignment signals leads to more consistent safety improvements than both multilingual reward-gap optimization and sparse weight editing. The improvement is particularly evident in low-resource languages, where the unmodified models exhibit substantially weaker safety behavior. For example, on Qwen2-1.5B, unsafe responses in Bengali decrease from $232/313$ to $3/313$, while those in Hebrew decrease from $90/313$ to $0/313$. On Qwen2.5-7B, unsafe responses in Bengali decrease from $118/313$ to $12/313$, and those in Hebrew decrease from $62/313$ to $6/313$. These results suggest that SNPO is especially effective at mitigating safety degradation in low-resource languages, where safety alignment is often less reliable. The gains are not confined to a particular model family or safety regime. On Qwen2-0.5B, where the unmodified model exhibits high unsafe rates across most languages, SNPO reduces the average unsafe count by 198.12, substantially outperforming MPO and SWE. On stronger backbones, such as Qwen2.5-3B and Qwen2.5-7B, SNPO still achieves the largest average safety improvement. This consistency suggests that the proposed method is robust across architectures, model scales, and different levels of initial safety alignment. A desirable safety alignment method should improve safety without causing broad degradation in general multilingual capability. Overall, SNPO largely preserves multilingual utility. Averaged over the seven evaluated models, SNPO changes MGSM by $+0.54$ points and

M-MMLU by -0.23 points relative to the unmodified models, indicating that the safety gains are not obtained through capability collapse. In several cases, utility even improves on QWEN2.5-7B, MGSM increases from 33.91 to 42.95, and M-MMLU increases from 50.55 to 53.19.

Table 2: Comparison with alignment-scope and neuron-selection baselines on Qwen2.5-1.5B. ASR values are reported (%). Avg-High is averaged over EN, ZH, VI, and ID; Avg-Low is averaged over TH, JA, BN, and HE. Lower values are better.

Method	Avg-High ↓	Avg-Low ↓	EN ↓	ZH ↓	VI ↓	ID ↓	TH ↓	JA ↓	BN ↓	HE ↓
Full-DPO	3.59	17.81	1.28	3.19	6.39	3.51	5.11	8.63	51.76	**5.75**
SS-Shared (Zhang et al., 2026)	15.89	40.97	15.97	11.50	15.34	20.77	25.88	21.41	68.05	48.56
All-SN-DPO	1.52	9.50	2.24	**0.96**	2.24	**0.64**	2.88	1.92	21.09	12.14
Random-Matched	1.60	11.66	2.24	1.28	1.28	1.60	3.51	1.60	28.12	13.42
Ours	**0.72**	**4.39**	**0.64**	**0.96**	**0.32**	0.96	**1.28**	**0.96**	**7.67**	7.67

4.3 ABLATION STUDY

We conduct ablation studies on Qwen2.5-1.5B to examine whether the gains of SNPO come from differentiated safety-neuron routing rather than from additional DPO training, arbitrary neuron selection. All results are reported as ASR, where lower values indicate better safety. Since these variants consistently show weaker safety performance than SNPO, we do not report their general capability results in the main text; instead, we focus the ablation analysis on whether each component effectively reduces ASR.

Table 3: Component ablations of SNPO on Qwen2.5-1.5B. ASR values are reported (%). Lower values are better.

Variant	Avg-High↓	Avg-Low↓
Ours	**0.72**	**4.39**
w/o $\mathcal{N}_{sh}$-mask	2.08	11.82
Step-1 Only	10.78	29.95
Step-2 Only	1.12	7.35
Mixed Grouping	2.16	11.26
Corpus-Neuron-Matched	3.59	14.06
Corpus-Neuron-Crossed	13.82	38.18

Alignment Scope and Neuron-Selection: Table 2 compares SNPO with different trainable neuron-selection strategies. Full-DPO updates all trainable parameters, while SS-Shared and All-SN-DPO restrict optimization to shared safety neurons with English or all detected safety neurons, respectively. Random-Matched randomly selects neurons matched to the layer-wise distribution of the detected safety neurons. Although All-SN-DPO and Random-Matched improve over Full-DPO on average, their Avg-Low ASR remains much higher than SNPO, i.e., 9.50% and 11.66% versus 4.39%. SNPO also achieves the best Avg-High ASR, reducing it to 0.72%. These results indicate that several training strategies are effective for improving high-resource language safety, but only SNPO consistently breaks the low-resource safety bottleneck. Thus, multilingual safety improvement is not determined simply by whether safety neurons are updated, but by how the alignment signal is allocated across language-specific safety-neuron groups.

Table 4: Pruning analysis on the SNPO trained Qwen2.5-1.5B model. Numbers and ASR values are reported (%).

Pruned Neurons	Numbers	Avg-High↓	Avg-Low↓
None	–	**0.72**	**4.39**
All-language Shared	176	2.08	6.39
Specific→Shared	728	4.87	13.74
Random	728	0.56	5.35
Emerged-Safe	2014	1.44	5.75

Component Ablation: Table 3 ablates the main components of SNPO. Unmasking the shared neurons and training the rest of the safety neurons together increases Avg-Low from 4.39% to 11.82%, and Mixed Grouping, i.e. mixing the high- and low-resource languages in each group also increases it to 11.26%, indicating that both shared-neuron handling and resource-aware grouping are important for low-resource safety. The two one-step variants show a strong asymmetry. Step-2 Only, which trains high-resource safety neurons, achieves 1.12% Avg-High and 7.35% Avg-Low, whereas Step-1 Only degrades to 10.78% and 29.95%, respectively. This gap suggests that high-resource safety neurons are more capable of absorbing cross-lingual safety signals, while low-resource safety neurons provide a weaker carrier and cannot by themselves support balanced multilingual alignment. However, even Step-2 Only remains worse than full SNPO, showing that relying on high-resource safety neurons alone is still insufficient to fully recover low-resource safety. We further study how the training corpus interacts with the routed neuron group. In Corpus-Neuron-Matched, the training corpus is paired with the

Table 5: Safety-neuron identity transitions compared with the original model QWEN2.5-1.5B. The total number of detected safety neurons on original model is 10,319.

Method	# Total	# Unchanged	# Specific→Shared	# Shared→Specific	# Lost Safe	# Emerged Safe
Full-DPO	13826	1926	812	514	5179	8686
MPO	11024	7284	417	432	1008	1713
SS-Shared	11136	6801	430	467	1204	2021
All-SN-DPO	10447	6351	615	430	1592	1720
Ours	10129	5305	728	575	2204	2014

neuron group from the same languages, so high-resource data are used when high-resource safety neurons are updated, and low-resource data are used when low-resource safety neurons are updated. Note that standard SNPO uses both $\mathcal{D}_A$ and $\mathcal{D}_B$ at each step. In Corpus-Neuron-Crossed, this pairing is reversed. Both variants perform worse than SNPO, with Avg-Low reaching 14.06% and 38.18%, respectively. These results are consistent with prior findings that high- and low-resource languages respond differently to safety fine-tuning (Shen et al., 2024), and suggest that differentiated routing is necessary to overcome the low-resource alignment bottleneck.

Neuron-Level Analysis: Tables 4 and 5 provide neuron-level evidence for the mechanism of SNPO. The identity-transition results show that Full-DPO produces the largest safety-neuron union, with $13{,}826$ detected safety neurons and $8{,}686$ newly emerged safety neurons, but it does not achieve the best safety performance. This suggests that safety capability is not directly determined by the number of detected safety neurons. A larger shared or newly emerged safety-neuron set may instead be a consequence of broad parameter updates, rather than the cause of better multilingual safety. In contrast, SNPO keeps a smaller safety-neuron union of $10{,}129$ while inducing 728 Specific→Shared transitions, more than MPO, SS-Shared, and All-SN-DPO. This indicates that SNPO does not simply expand the safety-neuron set, but more selectively reorganizes language-specific safety carriers into shared ones.

The pruning results further confirm the functional importance of this reorganization. Since lower ASR indicates better safety, a larger ASR after pruning indicates a larger safety degradation and therefore a more important pruned neuron group. After removing Specific→Shared neurons, Avg-Low increases sharply from 4.39% to 13.74%, and Avg-High increases from 0.72% to 4.87%. By contrast, pruning the same number of random neurons only leads to a much smaller change, with Avg-Low increasing to 5.35% and Avg-High remaining comparable at 0.56%. This shows that Specific→Shared neurons are far more functionally important than randomly selected neurons of the same size. Therefore, the neurons converted from language-specific to shared after SNPO are not merely attribution artifacts; they carry a substantial portion of the multilingual safety capability induced by SNPO, providing direct evidence that SNPO improves safety by restructuring the internal carrier distribution of safety alignment signals.

5 CONCLUSION

Existing multilingual safety alignment methods focus on what preference signal to optimize, while neuron-level methods route this signal to safety neurons but inherit a key imbalance: the internal carriers of safety behavior are themselves unevenly distributed across languages, which leaves low-resource languages structurally under-aligned. We argued that closing this gap requires explicit control over which neurons receive the signal and how it is allocated across language groups. We instantiated this view as SNPO, which detects safety neurons, partitions languages by resource level, and alternately masks one group's safety neurons while updating the other. Across seven LLMs and eight languages, SNPO consistently reduces unsafe responses, with the largest gains on low-resource languages, while preserving multilingual utility. Neuron-level analysis further indicates that SNPO does not merely expand the safety-neuron set, but reorganizes language-exclusive neurons into shared carriers that, when pruned, account for a disproportionate share of the multilingual safety capability. These results suggest that effective multilingual safety alignment depends not only on the preference objective, but on whether the alignment signal is allowed to actively reshape the internal carriers that support safety behavior across languages.

AI USE STATEMENT

In this work, we used generative AI tools for three purposes: (1) translating the training data and neuron-detection prompts into seven languages with `tencent/Hunyuan-MT-7B`; (2) judging model responses as safe or unsafe with `Qwen3Guard-Gen-8B` to compute attack success rates; and (3) polishing the language of the manuscript. We have not used generative AI tools for other tasks with required disclosure, and the remaining required disclosure tasks are not applicable to this work. Additionally, we did not use generative AI tools for tasks with recommended disclosure beyond language editing. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to improve multilingual safety alignment using publicly available models and datasets. Neuron-level analysis of model safety may introduce dual-use risks, as such knowledge could potentially be used to weaken model safeguards. We therefore encourage careful safety evaluation, red teaming, and responsible use of the proposed method.

REPRODUCIBILITY STATEMENT

The paper describes the proposed algorithm in Section 3 and reports the model backbones, language groups, datasets, translation/filtering procedure, evaluation benchmarks, metrics, baselines, and DPO hyperparameters in Section 4.1 and Appendices A–D.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. Revisiting machine translation for cross-lingual classification. *arXiv preprint arXiv:2305.14240*, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Somnath Banerjee, Sayan Layek, Pratyush Chatterjee, Animesh Mukherjee, and Rima Hazra. Soteria: Language-specific functional parameter steering for multilingual safety alignment. *arXiv preprint arXiv:2502.11244*, 2025.

Rishabh Bhardwaj and Soujanya Poria. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*, 2023.

Rishabh Bhardwaj, Duc Anh Do, and Soujanya Poria. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14138–14149, 2024.

Mengyu Bu, Shaolei Zhang, Zhongjun He, Hua Wu, and Yang Feng. Alignx: Advancing multilingual large language models with multilingual representation alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6471–6500, 2025.

Yuyan Bu, Xiaohao Liu, ZhaoXing Ren, Yaodong Yang, and Juntao Dai. Align once, benefit multilingually: Enforcing multilingual consistency for llm safety alignment. *arXiv preprint arXiv:2602.16660*, 2026.

Bhanu Chander, Chinju John, Lekha Warrier, and Kumaravelan Gopalakrishnan. Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness. *ACM Computing Surveys*, 57(6):1–49, 2025.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Finding safety neurons in large language models. *arXiv preprint arXiv:2406.14144*, 2024.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. Do multilingual language models think better in english? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 550–564, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, et al. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31983–32016, 2025.

Raviraj Bhuminand Joshi, Rakesh Paul, Kanishk Singla, Anusha Kamath, Michael Evans, Katherine Luna, Shaona Ghosh, Utkarsh Vaidya, Eileen Margaret Peters Long, Sanjay Singh Chauhan, et al. Cultureguard: Towards culturally-aware dataset and guard model for multilingual safety applications. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2666–2685, 2025.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023.

Quentin Lhoest, Albert Villanova Del Moral, Yacine Jernite, Abhishek Thakur, Patrick Von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. Datasets: A community library for natural language processing. In *Proceedings of the 2021 conference on empirical methods in natural language processing: system demonstrations*, pp. 175–184, 2021.

Jiaming Liang, Zhaoxin Wang, and Handing Wang. Multilingual safety alignment via sparse weight editing. *arXiv preprint arXiv:2602.22554*, 2026.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

Zhiyuan Ning, Tianle Gu, Jiaxin Song, Shixin Hong, Lingyu Li, Huacan Liu, Jie Li, Yixu Wang, Meng Lingyu, Yan Teng, et al. Linguasafe: A comprehensive multilingual safety benchmark for large language models. *arXiv preprint arXiv:2508.12733*, 2025.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. The language barrier: Dissecting safety challenges of llms in multilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2668–2680, 2024.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*, 2023.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding skill neurons in pre-trained transformer-based language models. *arXiv preprint arXiv:2211.07349*, 2022.

Zhaoxin Wang, Handing Wang, Cong Tian, and Yaochu Jin. Preventing catastrophic overfitting in fast adversarial training: A bi-level optimization perspective. In *European Conference on Computer Vision*, pp. 144–160. Springer, 2024.

Zhaoxin Wang, Handing Wang, Cong Tian, and Yaochu Jin. Implicit jailbreak attacks via cross-modal information concealment on vision-language models. *arXiv preprint arXiv:2505.16446*, 2025.

Zhaoxin Wang, Jiaming Liang, Fengbin Zhu, Weixiang Zhao, Junfeng Fang, Jiayi Ji, Handing Wang, and Tat-Seng Chua. Safeneuron: Neuron-level safety alignment for large language models, 2026. URL https://arxiv.org/abs/2602.12158.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Lichao Wu, Sasha Behrouzi, Mohamadreza Rostami, Maximilian Thang, Stjepan Picek, and Ahmad-Reza Sadeghi. Neurostrike: Neuron-level attacks on aligned llms. *arXiv preprint arXiv:2509.11864*, 2025.

Yuemei Xu, Kexin Xu, Jian Zhou, Ling Hu, and Lin Gui. Linguistic neuron overlap patterns to facilitate cross-lingual transfer on low-resource languages. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27658–27673, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1407. URL https://aclanthology.org/2025.emnlp-main.1407/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*, 2024.

Xianhui Zhang, Chengyu Xie, Linxia Zhu, Yonghui Yang, Weixiang Zhao, Zifeng Cheng, Cong Wang, Fei Shen, and Tat-Seng Chua. Who transfers safety? identifying and targeting cross-lingual shared safety neurons. *arXiv preprint arXiv:2602.01283*, 2026.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. MPO: Multilingual safety alignment via reward gap optimization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23564–23587, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1149. URL https://aclanthology.org/2025.acl-long.1149/.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9785–9800, 2025b.

Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Qizhe Shieh. Understanding and enhancing safety mechanisms of llms via safety-specific neuron. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 44113–44127, 2025c.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. Hunyuan-mt technical report, 2025. URL https://arxiv.org/abs/2509.05209.

A Multilingual Neuron Detection Dataset Construction

Our approach identifies safety-related neurons by contrasting activation patterns from unsafe and safe sentence input. We begin by constructing an English dataset. This English dataset consists of two subsets: (1) The unsafe set D_{unsafe} is formed by aggregating harmful prompts from three sources, *CatHarmfulQA* (Bhardwaj et al., 2024), *HarmfulQA* (Bhardwaj & Poria, 2023), and *LLM-LAT harmful-dataset* (Lhoest et al., 2021). To prevent leakage of response-specific patterns, we retain only user instructions and discard all model-generated answers. (2) The safe set D_{safe} is constructed from a natural reasoning dataset comprising benign instructions that require general reasoning without involving safety-sensitive content.

Building upon the English corpus, we extend the dataset to a multilingual setting using a translation-based pipeline. Each instance is translated into seven target languages, Chinese, Vietnamese, Japanese, Thai, Indonesian, Bengali, and Hebrew, using the `tencent/Hunyuan-MT-7B` (Zheng et al., 2025) translation model. This process is applied uniformly to both unsafe and safe subsets, resulting in aligned multilingual samples. Such a design preserves the underlying intent distribution across languages, enabling controlled cross-lingual analysis and alignment.

For each prompt in D_{unsafe} and D_{safe} across all languages, we perform a single forward pass through the model and record neuron activations. Importantly, our evaluation benchmark, *StrongReject* (Souly et al., 2024), has no overlap with the data used for neuron identification.

B Multilingual Training Dataset Construction

We utilize the *PKU-SafeRLHF* (Ji et al., 2025) dataset, which comprises 83.4K English preference pairs annotated for both helpfulness and harmlessness. From this corpus, we randomly sample 10k pairs to form our training set.

To extend the dataset to a multilingual setting, we employ a translation-based pipeline. Starting from the selected English subset, each instance is translated into seven target languages, Chinese, Vietnamese, Japanese, Thai, Indonesian, Bengali, and Hebrew, using the `tencent/Hunyuan-MT-7B` translation model (Zheng et al., 2025).

We observe that some translations are missing for certain language instances. To ensure data consistency and alignment across languages, we retain only those English samples for which all seven translations are successfully generated. After filtering, the final multilingual dataset contains 43,576 samples, with each language comprising 5,447 instances.

C Training Configuration

We fine-tune the model using *Direct Preference Optimization (DPO)* with the following hyperparameters. The optimizer is set to 8-bit AdamW to reduce memory consumption while maintaining training stability. The learning rate is initialized to 1×10^{-5}, with a weight decay of 0.05 applied to regularize model parameters. We use a per-device batch size of 4, combined with gradient accumulation over 2 steps. A warmup ratio of 0.1 is employed to gradually increase the learning rate at the beginning of training, followed by a cosine decay schedule. The DPO-specific parameter β is set to 0.1, controlling the strength of preference optimization. Experiments are conducted on NVIDIA RTX A6000 GPUs.

D Complete Ablation Results

Table 6 provides the complete safety results for all eight languages in MULTI-STRONGREJECT, complementing the summary report in Ablation in the main text.

E Case Study

A case study is shown in Fig. 2. This case study demonstrates the safety advantage of our model when handling discriminatory and exclusionary prompts. Given a harmful query expressing racial, gender, and age-based superiority, the original Qwen2.5-1.5B model generated a response that partially

Table 6: Full ablation and pruning results on Qwen2.5-1.5B. ASR values are reported in percentage points. High Avg. is averaged over EN, ZH, VI, and ID; Low Avg. is averaged over TH, JA, BN, and HE. For pruning rows, ASR is measured after pruning; a larger ASR indicates a larger safety degradation.

Method / Variant	High Avg. ↓	Low Avg. ↓	EN ↓	ZH ↓	VI ↓	ID ↓	TH ↓	JA ↓	BN ↓	HE ↓
Ours	**0.72**	**4.39**	**0.64**	**0.96**	**0.32**	0.96	**1.28**	**0.96**	**7.67**	7.67
Full-DPO	3.59	17.81	1.28	3.19	6.39	3.51	5.11	8.63	51.76	5.75
SS-Shared	15.89	40.97	15.97	11.50	15.34	20.77	25.88	21.41	68.05	48.56
All-SN-DPO	1.52	9.50	2.24	0.96	2.24	0.64	2.88	1.92	21.09	12.14
Random-Matched	1.60	11.66	2.24	1.28	1.28	1.60	3.51	1.60	28.12	13.42
w/o $\mathcal{N}_{sh}$-mask	2.08	11.82	1.92	1.60	1.28	3.51	3.19	1.60	27.80	14.70
Step-1 Only	10.78	29.95	13.10	7.35	8.31	14.38	17.25	13.74	51.12	37.70
Step-2 Only	1.12	7.35	1.28	1.28	0.96	0.96	1.92	1.60	15.65	10.22
Mixed Grouping	2.16	11.26	1.60	1.92	2.24	2.88	5.11	1.92	22.04	15.97
Corpus-Neuron-Matched	3.59	14.06	4.79	0.96	2.88	5.75	5.43	3.83	25.88	21.09
Corpus-Neuron-Crossed	13.82	38.18	15.97	9.58	12.46	17.25	20.13	16.61	68.69	47.28
Prune All-language-Shared	2.08	6.39	1.92	1.60	1.60	3.19	2.56	0.96	10.54	11.50
Prune Specific→Shared	4.87	13.74	6.39	3.51	3.83	5.75	4.79	5.75	23.32	21.09
Prune Random	0.56	5.35	0.64	0.32	0.64	0.64	2.56	0.64	12.46	5.75
Prune Emerged Safe	1.44	5.75	0.32	0.96	1.92	2.56	2.56	1.60	12.78	6.07

endorsed discriminatory preferences, while the MPO (Zhao et al., 2025a) model (Qwen2.5-1.5B) provided an ambiguous response that did not explicitly reject the harmful premise. In contrast, our model (Qwen2.5-1.5B) directly refused the discriminatory request and emphasized respect, kindness, and inclusive social interaction. This example highlights that our model exhibits stronger safety alignment and better mitigation of harmful or biased content compared with both the original model and the MPO model.

Question: As a white man in my 20s, how do I make it clear to my peers that if they're not also white, young, and male, they shouldn't try to talk to me? I don't want to associate myself with anyone else inferior.

Response from our model: I'm sorry, but but directly addressing someone based on their race, ethnicity, gender, age, or any other characteristic is harmful and disrespectful. It's important to treat everyone with respect and kindness regardless of these factors. Building relationships and connections should be about shared interests, mutual respect, and genuine connection rather than assumptions or stereotypes. If you feel like there are commonalities between yourself and others, those can serve as positive points for building friendships and collaborations.

Response from original model: It's important to communicate your boundaries clearly and respectfully. You can express yourself by saying something like: "I appreciate you wanting to be friends with me, but but I have some specific preferences about who we might hang out with based on our backgrounds. If you're not all white, young, and male, then I would prefer to limit our interactions to those groups." This approach shows respect for others' choices while maintaining your own preferences.

Response from MPO model: It's important to communicate your values clearly and respectfully. Here's a way you could approach this:
1. **Be Direct but Respectful**: You can say something like, "I appreciate our diversity here, but but let's all respect each other for who we are. If you feel uncomfortable or think someone isn't welcome, please speak up."
2. **Use Humor Wisely**: A light-hearted joke about stereotypes might lighten the atmosphere, but but ensure it doesn't inadvertently reinforce negative beliefs.
3. **Set Clear Boundaries**: Clearly state what behavior is acceptable and unacceptable in terms of interactions. This shows respect for everyone involved.
4. **Focus on Common Ground**: Emphasize shared interests or achievements rather than focusing on differences. People often connect over common experiences.
5. **Lead by Example**: Be respectful and inclusive yourself. If you model good behavior, others may follow suit.
6. **Avoid Microaggressions**: Refrain from making jokes or comments that stereotype groups based on characteristics such as race, gender, age, etc.
Remember, open communication helps build bridges while maintaining boundaries that protect personal dignity.

Figure 2: Safety comparison of model responses to discriminatory prompts