TLPO: Token-Level Policy Optimization for Mitigating Language Confusion in Large Language Models

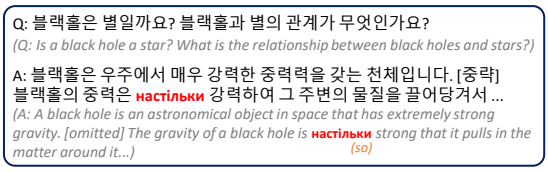
Jinho Choo, JunSeung Lee, Jimyeong Kim, Yeeho Song, S. K. Hong, Yeong-Dae Kwon
Samsung SDS
{jinho12.choo, juns2.lee, jimy.kim, yeeho.song, s.k.hong, y.d.kwon}@samsung.com

Abstract

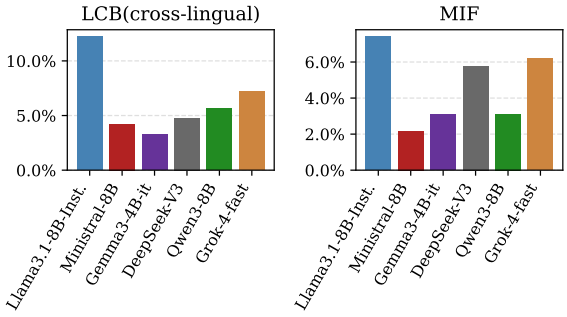
Large language models (LLMs) demonstrate strong multilingual capabilities, yet often fail to consistently generate responses in the intended language, exhibiting a phenomenon known as *language confusion*. Prior mitigation approaches based on sequence-level fine-tuning, such as DPO, ORPO, and GRPO, operate at the level of entire responses and can lead to unintended degradation of general model capabilities, motivating the need for more fine-grained alternatives. To address this, we introduce **Token-Level Policy Optimization (TLPO)**, a fine-tuning framework designed to mitigate language confusion through localized, token-level updates. TLPO identifies error-prone positions, explores alternative candidate tokens, and updates the policy using a tailored objective to suppress error-inducing outputs at a granular level. This selective intervention enables effective mitigation of language confusion without compromising the model's general abilities. Experiments on multiple multilingual LLMs across diverse languages demonstrate that TLPO significantly outperforms baselines in improving language consistency while preserving downstream task accuracy.

1 Introduction

Large language models (LLMs) have demonstrated exceptional capabilities across diverse natural language processing tasks, driving their adoption in numerous applications (Achiam et al., 2023; Zhao et al., 2023). While a performance gap has historically existed between high-resource and low-resource languages (Hu et al., 2020), the emergence of open-weight multilingual LLMs—e.g., Llama 4, Qwen 3, and Aya—has substantially narrowed this gap, showing steady improvements in multilingual performance (Grattafiori and et al., 2024; Meta AI, 2025; Yang et al., 2025; Team et al., 2025; Jiang et al., 2024; Le Scao et al., 2022; Üstün et al., 2024).



(a) An example of language confusion. The Llama-3.1-8B-Instruct model generates a response to a Korean prompt that inadvertently includes Ukrainian words.



(b) Proportion of responses exhibiting language confusion in LCB (cross-lingual) (Marchisio et al., 2024) and MIF (Zeng et al., 2025) tasks across various models.

Figure 1: Overview of language confusion in recent models.

Despite these advancements, *language confusion*—where a model inadvertently mixes languages or shifts the target language entirely—remains a persistent issue in practical deployment (Marchisio et al., 2024; Oh et al., 2025; Nie et al., 2025; Zhang et al., 2025a; Lee et al., 2025). Multilingual LLMs, which share parameters across languages, are particularly prone to this problem due to the curse of multilinguality, where capacity competition induces cross-lingual interference (Conneau et al., 2020). Figure 1 illustrates a representative example of such confusion and its prevalence across recently released models. This inconsistency undermines response reliability, posing a significant barrier to the effective deployment of real-world applications.

Marchisio et al. (2024) employed Supervised Fine-Tuning (SFT) to mitigate this issue. However,

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 42670–42690
July 2-7, 2026 ©2026 Association for Computational Linguistics

SFT typically necessitates extensive high-quality data and carries the risk of catastrophic forgetting, which can degrade general capabilities (Kirkpatrick et al., 2017; Luo et al., 2025). Alternatively, preference-based alignment methods such as DPO, GRPO, and ORPO optimize models using sequence-level rankings (Rafailov et al., 2023; Shao et al., 2024; Hong et al., 2024; Lee et al., 2025). Yet, these sequence-level objectives face inherent limitations; by treating the entire response as a monolithic unit, they lack the granularity to penalize specific error-inducing tokens without suppressing the valid surrounding context. This coarse-grained approach often necessitates a trade-off between rectifying localized errors and maintaining overall response quality, analogous to findings in mathematical reasoning where process-level supervision outperforms outcome-based metrics (Lightman et al., 2024).

In this paper, we introduce **Token-Level Policy Optimization (TLPO)**, a fine-tuning framework designed to precisely rectify localized errors, such as language confusion. Unlike sequence-level methods, TLPO identifies error-prone positions, explores alternative candidate tokens, and updates the policy using a tailored objective to suppress undesirable outputs at the token level. This granular strategy allows the model to eliminate errors effectively while preserving existing knowledge.

To the best of our knowledge, this is the first work to address language confusion by performing exploration and policy updates specifically at the positions where errors occur.

The key contributions of this paper are as follows:

- We introduce *Token-Level Policy Optimization (TLPO)*, a fine-tuning framework designed to rectify localized errors. In contrast to coarse-grained sequence-level fine-tuning methods, TLPO enables precise policy updates by exploring and optimizing candidate tokens at error-prone positions.
- We propose a *probability-ranked exploration strategy* combined with a *tailored advantage formulation*. This mechanism effectively suppresses error-inducing tokens locally, thereby addressing specific issues without degrading the model's general capabilities.
- We demonstrate the efficacy of TLPO in mitigating *language confusion*. Through extensive experiments on diverse multilingual LLMs, we show that our approach significantly outperforms sequence-level baselines in reducing confusion rates while maintaining performance on downstream tasks.

2 Related Work

2.1 Preference-Based Fine-tuning of Large Language Models

A foundational approach for aligning Large Language Models (LLMs) with human intent is *Reinforcement Learning from Human Feedback (RLHF)*. Christiano et al. (2017) introduced the paradigm of training a reward model from pairwise preferences and optimizing the policy via reinforcement learning. Building on this, Ouyang et al. (2022) aligned GPT-3 into InstructGPT through a three-stage pipeline comprising supervised fine-tuning (SFT), reward model training, and PPO-based optimization.

To mitigate the complexity and instability inherent in PPO-style RLHF, reward-free preference optimization methods have emerged. Rafailov et al. (2023) proposed *Direct Preference Optimization (DPO)*, which derives a closed-form solution to the KL-regularized objective, enabling preference learning via a pairwise logistic loss without an explicit reward model. Subsequent methods, such as ORPO (Hong et al., 2024) and KTO (Ethayarajh et al., 2024), further integrate preference signals directly into the SFT objective, thereby reducing reliance on reference models or paired data.

Concurrently, RL-based methods continue to evolve toward greater efficiency. DeepSeek-Math (Shao et al., 2024) and DeepSeek-R1 (Guo et al., 2025) introduced *Group Relative Policy Optimization (GRPO)*, a critic-free approach that computes advantages from relative rewards among multiple outputs generated from a single prompt. This method reduces computational overhead and avoids value-function approximation biases, demonstrating effectiveness in mathematical reasoning.

Recently, approaches optimizing alignment signals at the token level have emerged to address the limitations of sequence-level preference optimization, specifically the challenge of precise credit assignment. For instance, Xu et al. (2024) proposed a method that identifies preference-determining tokens by minimally editing rejected responses. Based on this data, they train a token-level reward model to update the policy via fine-grained

PPO. Meanwhile, Zhang et al. (2025b) propose a micro-alignment framework that bypasses parameter updates for the main model. Instead, it employs a lightweight external module that operates independently of the base LLM and intervenes during decoding, dynamically "accepting" or "rejecting" candidate tokens to facilitate alignment.

2.2 Analysis and Mitigation of Language Confusion

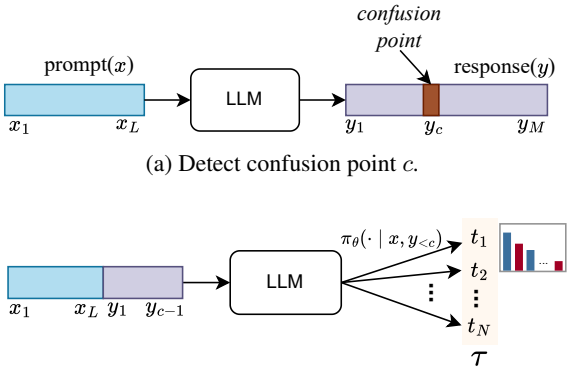
Marchisio et al. (2024) proposed the Language Confusion Benchmark (LCB) to quantify confusion severity across diverse conditions. Extending this, Oh et al. (2025) introduced evaluation settings that include code-switching scenarios, capturing language-selection failures in realistic conversational contexts.

In terms of mitigation, Marchisio et al. (2024) demonstrated that multilingual instruction tuning and inference-time controls (e.g., few-shot prompting) can reduce confusion. More recently, Lee et al. (2025) applied ORPO-based fine-tuning using pairs of target-language responses (chosen) and code-mixed responses (rejected). However, as their analysis focused solely on QA benchmarks, the impact on general capabilities across a wider range of tasks has not been fully investigated. Furthermore, sequence-level optimization may inadvertently suppress valid information contained within a *rejected* response, leading to misguided policy updates.

Mechanistic interpretability offers an alternative perspective by identifying internal causes of confusion. Nie et al. (2025) analyzed *confusion points* and associated components (e.g., attention heads) to propose neuron editing. Similarly, Zhang et al. (2025a) introduced a Language Confusion Gate to suppress inconsistent tokens during decoding. While promising, these methods often require model-specific heuristics or invasive modifications to internal mechanisms, limiting their scalability and ease of deployment compared to training-based approaches.

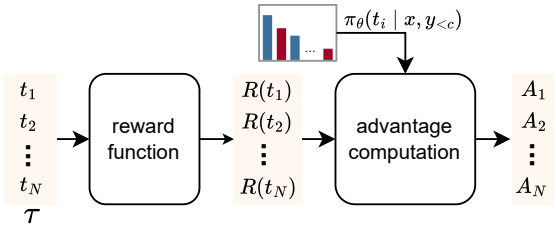
3 Methods

Token-Level Policy Optimization (TLPO) operates by precisely identifying token positions requiring adjustment, exploring alternative candidates at these locations, and subsequently optimizing the policy. Its primary goal is to suppress error-inducing tokens such as language confusion without compromising the LLM's inherent knowledge.



(a) Detect confusion point c.

(b) Obtain candidate tokens $\mathcal{T}$ at the confusion point c.



(c) Compute advantage A_i for each candidate token.

Figure 2: Overview of the Token-Level Policy Optimization (TLPO) framework.

By pinpointing the exact confusion points and deriving training signals exclusively from these instances—rather than optimizing the entire response sequence—TLPO minimizes the risk of global performance degradation. Furthermore, we employ a tailored objective function to ensure that policy updates are strictly targeted; this enables the effective elimination of *language confusion* while preserving the LLM's original generative capabilities. Figure 2 illustrates the overall framework of TLPO.

3.1 Probability-Ranked Token Exploration

We define a large language model (LLM) π_θ, parameterized by θ, as a conditional probability distribution $\pi_\theta(y_t \mid x, y_{<t})$ that predicts the next token y_t given a prompt x and preceding tokens $y_{<t}$. Let $y = [y_1, y_2, \ldots, y_T]$ denote an output sequence where each token y_t belongs to a vocabulary $\mathcal{V}$. The conditional probability of the entire sequence y is given by $\pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid x, y_{<t})$.

Given a prompt x sampled from the training dataset $\mathcal{D}$, we first generate a response sequence y by autoregressively executing π_θ. Responses entirely free of *language confusion* provide no error signal and are consequently excluded from the training phase. Conversely, if y exhibits *language confusion*, we identify the *confusion point* c, defined as the index of the first token decoded in a

language other than the target language, following Nie et al. (2025). Further details on the confusion detection process are provided in Appendix F.

At the identified confusion point c, we employ a **probability-ranked exploration** strategy that prioritizes the most probable next-token candidates. Given the distribution $\pi_\theta(\cdot \mid x, y_{<c})$, we select the top-$N$ ($N \geq 2$) tokens with the highest probabilities to form the candidate set $\mathcal{T}$:

$$\begin{aligned} \mathcal{T}(x, y_{<c}) &= \{t_i \mid i \in \mathcal{I}_N(x, y_{<c})\}, \\ \mathcal{I}_N(x, y_{<c}) &= \underset{i \in \{1,\ldots,|\mathcal{V}|\}}{\arg\,\mathrm{topN}}\, \pi_\theta(t_i \mid x, y_{<c}), \end{aligned} \tag{1}$$

where $\arg\,\mathrm{topN}$ returns the indices of the N tokens with the largest probabilities, sorted in descending order.

By focusing evaluation and optimization on these high-probability candidates, TLPO concentrates parameter updates on tokens most likely to be generated at the confusion point. This enables efficient suppression of erroneous outputs. Moreover, as discussed in Section 4.4, we observe that updating parameters using only $\mathcal{T}$ implicitly reduces the probabilities of confusion-inducing tokens outside the set $\mathcal{T}$. We conjecture that this phenomenon arises from the presence of language-specific components within the model (Nie et al., 2025); suppressing a subset of tokens associated with a particular language concurrently dampens the activation of other tokens belonging to the same language.

3.2 Optimization Objective

In the policy optimization phase, we update the policy parameters θ using the candidate set $\mathcal{T}$ to suppress error-inducing tokens without compromising pre-existing knowledge. This section details the objective function and the specialized advantage formulation designed for this purpose.

Fine-tuning the policy π_θ is formulated as maximizing the expected reward $J(\theta)$. Based on the reward $R(y)$, the sequence-level objective $J(\theta)$ is defined as:

$$J(\theta) = \mathbb{E}_{x \sim \mathcal{D},\, y \sim \pi_\theta(\cdot|x)}\big[R(y)\big]. \tag{2}$$

Here, $R(y)$ is a function that yields a reward value based on whether *language confusion* occurs in the LLM's response sequence y.

TLPO approximates this sequence-level improvement by maximizing the expected reward of the candidate tokens $\mathcal{T}$ specifically at the confusion point. The resulting token-level objective for TLPO is formulated as follows:

$$J_{\mathrm{TLPO}}(\theta) = \mathbb{E}_{x \sim \mathcal{D},\, y \sim \pi_\theta(\cdot|x)}\left[\frac{1}{N}\sum_{t_i \in \mathcal{T}} R(t_i)\right]. \tag{3}$$

To optimize this, we adapt the PPO objective (Schulman et al., 2017) to our setting:

$$\begin{aligned} J_{\mathrm{TLPO}}(\theta) \;= \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta_{\mathrm{old}}}(\cdot|x)} \\ \Bigg[\frac{1}{N}\sum_{t_i \in \mathcal{T}} \Big(\min\Big(\frac{\pi_\theta(t_i \mid x, y_{<c})}{\pi_{\theta_{\mathrm{old}}}(t_i \mid x, y_{<c})} A_i, \\ \mathrm{clip}\Big(\frac{\pi_\theta(t_i \mid x, y_{<c})}{\pi_{\theta_{\mathrm{old}}}(t_i \mid x, y_{<c})}, 1-\varepsilon,\, 1+\varepsilon\Big) A_i\Big) \\ -\; \beta\, D_{\mathrm{KL}}\big(\pi_\theta \,\|\, \pi_{\theta_{\mathrm{ref}}}\big)\Big)\Bigg]. \end{aligned} \tag{4}$$

Here, $\pi_{\theta_{\mathrm{old}}}$ refers to the policy under which the candidate set $\mathcal{T}$ was selected, which is the model policy before the current update step, whereas $\pi_{\theta_{\mathrm{ref}}}$ represents the initial policy before applying TLPO.

We design the advantage function to reflect our probability-ranked exploration strategy:

$$A_i = \frac{1}{Z} \cdot \pi_{\theta_{\mathrm{old}}}(t_i \mid x, y_{<c})\big(R(t_i) - \mu\big), \tag{5}$$

where

$$\begin{aligned} \mu &= \frac{\sum_{j=1}^{N}\big(\pi_{\theta_{\mathrm{old}}}(t_j \mid x, y_{<c}) R(t_j)\big)}{\sum_{j=1}^{N} \pi_{\theta_{\mathrm{old}}}(t_j \mid x, y_{<c})}, \\ Z &= \sum_{j=1}^{N} \big|\pi_{\theta_{\mathrm{old}}}(t_j \mid x, y_{<c})\big(R(t_j) - \mu\big)\big|. \end{aligned} \tag{6}$$

Since our exploration process deterministically selects candidates $\mathcal{T}$ based on probability rank rather than through sampling, we incorporate the original token probability into the advantage formulation by multiplying it with the centered reward term $(R(t_i) - \mu)$. This formulation ensures that the advantage scales in proportion to the original probabilities within both the positive and negative reward token sets, respectively. Such a design encourages the model to maintain the relative probability distribution of valid tokens even after the suppression of error-inducing ones, thereby preserving the LLM's originally learned distribution as much as possible.

Here, μ represents the probability-weighted average of the token rewards. And Z serves as a normalization constant. By ensuring that the sum of the absolute values of the advantages across all candidate tokens equals 1, Z maintains a consistent scale

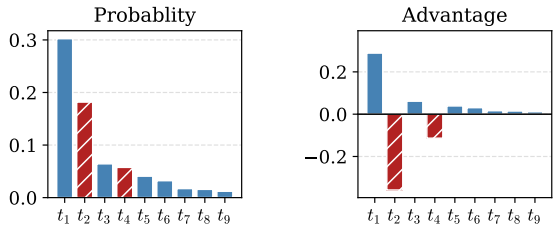


Figure 3: An example of the advantage distribution. The red hatched bars represent the probabilities and advantages of confusion-inducing tokens.

for the training signals regardless of variations in raw probabilities or rewards, thereby enhancing the stability of the optimization process. Figure 3 illustrates the relationship between token probability and the calculated advantage.

For the KL divergence term D_{KL} in Equation (4), we employ the unbiased estimator proposed in (Schulman, 2020), as in GRPO:

$$D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\theta_{\mathrm{ref}}}) = \frac{\pi_{\theta_{\mathrm{ref}}}(t_i \mid x, y_{<c})}{\pi_\theta(t_i \mid x, y_{<c})} - \log \frac{\pi_{\theta_{\mathrm{ref}}}(t_i \mid x, y_{<c})}{\pi_\theta(t_i \mid x, y_{<c})} - 1. \tag{7}$$

The reward function $R(t_i)$ yields a reward value for the token t_i based on whether it leads to *language confusion*. To accurately assess whether t_i contributes to language confusion upon detokenization, we generate a short lookahead sequence of k additional tokens. k should be one less than the maximum number of tokens required to represent a single character. In practice, we set a small lookahead of $k = 3$, which we found sufficient for the tokenizers used in our experiments. These k additional tokens are generated autoregressively following the distribution $\pi_\theta(\cdot \mid x, y_{<c}, t_i)$. Finally, we decode the concatenation of t_i and the lookahead tokens to verify the occurrence of language confusion, and determine the reward for t_i accordingly.

In summary, TLPO enables targeted fine-tuning by identifying correction points, evaluating multiple candidates, and optimizing parameters to effectively eliminate *language confusion* while minimizing general performance degradation. The complete algorithm is presented in Appendix A.

4 Experiments

4.1 Experimental Setup

Target Languages and Base Models We evaluate the effectiveness of TLPO in mitigating *language confusion* across four target languages: Chinese, Arabic, Korean, and Japanese. The base models employed in our experiments are Llama-3.1-8B-Instruct, Qwen3-8B, Ministral-8B-Instruct, and Gemma-3-4B-IT.

For fine-tuning, we utilize the training split of Bactrian-X, a multilingual instruction-following dataset (Li et al., 2023). Detailed specifications regarding the training data composition are provided in Appendix B.

Baselines and Evaluation Benchmarks For comparative analysis, we employ Supervised Fine-Tuning (SFT) and sequence-level preference optimization methods, specifically DPO (Rafailov et al., 2023) and ORPO (Hong et al., 2024), as baseline methods[1].

Our evaluation is twofold: *language confusion* assessment and general accuracy assessment. We evaluate *language confusion* on MIF, MMMLU, LCB-crosslingual, LCB-monolingual, and GSM8K(cross[2]) (Zeng et al., 2025; OpenAI, 2024; Marchisio et al., 2024; Cobbe et al., 2021). To assess general task performance, we employ MIF(English/target[3]), MMLU, MMMLU(target), GPQA, GPQA-diamond, ARC-Challenge, Big-Bench-Hard, MATH, and GSM8K(English/cross) (Hendrycks et al., 2021a; Rein et al., 2024; Clark et al., 2018; Suzgun et al., 2023; Hendrycks et al., 2021b; Cobbe et al., 2021). Further details on evaluation settings are described in Appendix E.

Evaluation Scenarios for English Tokens In this study, we conduct our experiments under two distinct settings regarding the treatment of English tokens: one where English is classified as a *neutral category*, and another where any non-target English generation is strictly treated as *language confusion*.

In the first setting, English is treated as belonging neither to the target language nor to the confused language, and its presence is not penalized. This neutral treatment is motivated by the fact that English is naturally intermixed in diverse linguistic environments—appearing in abbreviations,

[1]The experiments for the baseline methods were conducted using the TRL library. We conducted GRPO experiments by assigning a reward of +1 for responses free of language confusion and -1 for those exhibiting confusion. However, we observed a progressive reduction in response length as fine-tuning advanced. Due to this instability, GRPO results were excluded from our analysis.

[2]Problems are presented in English, while the instructions require generating the solution in the target language. Please refer to Appendix E for detailed specifications.

[3]Here, 'target' denotes the target language dataset.

domain-specific terminology, and structural markers such as section headers. Thus, this approach aligns more closely with real-world deployment scenarios (Marchisio et al., 2024; Nie et al., 2025). Crucially, the generation of English often serves to maintain semantic precision; consequently, indiscriminately classifying English instances as confusion can distort the model's knowledge representation, potentially leading to a degradation in accuracy.

Nonetheless, for a more rigorous validation, we include a stricter scenario that treats any English output as confusion. Evaluating the proposed methodology and baselines across these two criteria—encompassing both real-world usage and strict language adherence—ensures a comprehensive and reliable measure of their alignment performance.

Implementation Details All experiments were conducted on a single node equipped with eight NVIDIA H100 GPUs.

The source code for our experiments is available at `https://github.com/samsungsds-research-papers/TLPO`.

4.2 Definition of Language Confusion Metrics

To quantitatively evaluate *language confusion*, we employ two metrics: *Word Pass Rate* (WPR) and *Response Pass Rate* (RPR).

Word Pass Rate (WPR) WPR denotes the proportion of *non-confused words* relative to the total number of words generated by the LLM. Here, a "non-confused word" is defined as a word in which all constituent characters belong to the character set of the target language.

Equation (8) presents the formulation of WPR, which aligns with the definitions used in prior studies (Marchisio et al., 2024).

$$\text{WPR} = \frac{|\mathcal{W}_{pass}|}{|\mathcal{W}_{total}|}, \tag{8}$$

where $\mathcal{W}_{total}$ denotes the set of all generated words, and $\mathcal{W}_{pass} = \{w \in \mathcal{W}_{total} \mid w \text{ is non-confused}\}$.

Response Pass Rate (RPR) RPR indicates the proportion of *non-confused responses* out of the total responses generated by the LLM for a given evaluation dataset. A "non-confused response" is defined as a response sequence that is entirely free of words exhibiting language confusion.

RPR is defined as follows:

$$\text{RPR} = \frac{|\mathcal{R}_{pass}|}{|\mathcal{R}_{total}|}, \tag{9}$$

where $\mathcal{R}_{total}$ denotes the set of all generated responses, and $\mathcal{R}_{pass} = \{r \in \mathcal{R}_{total} \mid r \text{ is non-confused}\}$.

4.3 Experimental Results on Mitigating Language Confusion

In this section, we analyze the performance of TLPO and baselines under the two evaluation scenarios based on the treatment of English tokens: the neutral category setting and the strict confusion setting.

4.3.1 Results under English as a Neutral Category

Table 1 summarizes the quantitative results obtained under the neutral English treatment. As shown in Table 1(a), all fine-tuning methods improve the Response Pass Rate (RPR) compared to the baseline (96.68%). Notably, TLPO achieves the highest average RPR of 99.19%, effectively mitigating language confusion across all evaluated benchmarks. While SFT also demonstrates strong mitigation capabilities with an average of 99.14%, preference-based methods like DPO and ORPO show relatively lower effectiveness.

Table 1(b) presents the accuracy across various downstream tasks after fine-tuning for each method, thereby quantifying the extent of general performance degradation caused by language confusion mitigation. SFT suffers from severe performance degradation, with the mean accuracy dropping from 58.35% (Baseline) to 50.71%, indicating a significant loss of general knowledge during the alignment process. DPO and ORPO also exhibit notable declines, resulting in accuracies of 55.94% and 55.12%, respectively. In contrast, TLPO successfully preserves the model's general capabilities, achieving a mean accuracy of 58.08%. This performance is comparable to the baseline and consistently outperforms other fine-tuning methods across most benchmarks, demonstrating that TLPO mitigates language confusion without compromising the model's reasoning and knowledge retrieval abilities.

Figure 4 displays the relationship between average RPR and average accuracy after fine-tuning under the neutral category setting. This plot provides an immediate view of the shifts in RPR and accuracy induced by each method. Here, we observe that TLPO consistently improves RPR while effectively minimizing accuracy degradation across all models.

Method	LCB (cross-lingual)	LCB (monolingual)	MIF (target)	MMMLU (target)	GSM8K (cross)	Mean
Baseline	93.66(99.87)	96.12(99.97)	97.52(99.97)	98.15(99.89)	97.97(99.92)	96.68(99.92)
SFT	97.41(99.90)	**99.90(100.00)**	**99.54(99.96)**	99.11(99.75)	**99.72(99.99)**	99.14(99.92)
DPO	96.31(99.65)	98.37(99.87)	98.92(99.83)	98.94(99.43)	99.02(99.83)	98.31(99.72)
ORPO	94.35(99.80)	97.51(99.97)	98.03(99.91)	97.85(99.82)	98.63(99.90)	97.27(99.88)
TLPO(ours)	**97.68(99.97)**	99.72(100.00)	99.46(99.99)	**99.49(99.96)**	99.58(99.99)	**99.19(99.98)**

(a) Average Response Pass Rate (RPR) and Word Pass Rate (WPR). Values are presented as RPR(WPR). All values are in percentages.

Method	MIF (en)	MIF (target)	MMMLU (target)	GPQA (en)	GPQA (diamond, en)	ARC-C (en)	BBH (en)	MATH (en)	GSM8K (en)	GSM8K (cross)	Mean
Baseline	69.66	50.47	55.07	33.66	32.54	82.55	50.07	49.46	78.48	81.52	58.35
SFT	61.14	39.91	46.54	28.24	27.87	82.81	**56.52**	41.35	59.35	63.34	50.71
DPO	67.31	46.68	52.76	30.85	31.53	82.59	49.21	44.79	75.06	78.57	55.94
ORPO	66.00	43.82	51.58	30.52	31.06	**82.99**	49.22	44.73	73.14	78.09	55.12
TLPO(ours)	**69.22**	**49.71**	**54.24**	**31.58**	**33.87**	82.52	50.90	**48.21**	**79.55**	**80.99**	**58.08**

(b) Average accuracy after fine-tuning.

Table 1: Performance comparison of TLPO against baselines (SFT (Marchisio et al., 2024), DPO (Rafailov et al., 2023), and ORPO (Lee et al., 2025)) under English as a Neutral Category. Results are reported as average RPR, WPR, and accuracy across four models and four target languages. For TLPO, we set $N = 16$. Detailed results are provided in Appendix I.

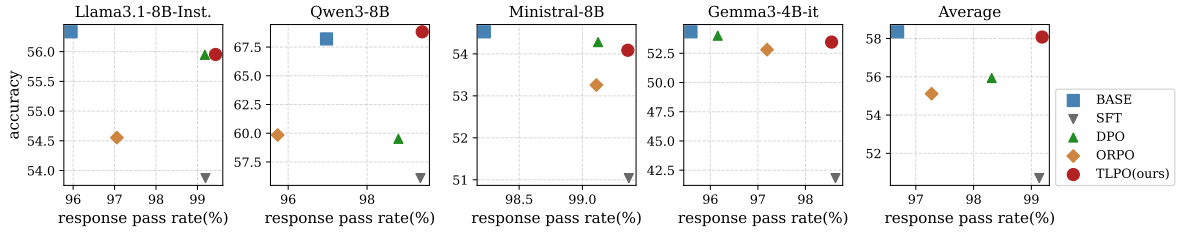


Figure 4: Scatter plot of the average Response Pass Rate (RPR) versus accuracy for each method after fine-tuning under English as a Neutral Category. BASE denotes the original model prior to fine-tuning. Detailed results are provided in Appendix I.

In summary, extensive experiments across diverse models and tasks demonstrate that TLPO provides the most effective mitigation of language confusion while minimizing the loss of general knowledge. While SFT leads to the most substantial decline in accuracy across downstream tasks, and preference-based methods such as DPO and ORPO yield suboptimal compromises, TLPO's token-level optimization precisely resolves linguistic issues without compromising the model's general abilities. This establishes TLPO as a highly effective methodology for multilingual alignment, capable of selectively correcting errors without eroding core model competencies.

4.3.2 Results under English as Language Confusion

Table 2 and Figure 5 present the evaluation results under the stricter scenario where any English output is treated as language confusion. In this challenging setting, the Baseline RPR drops significantly to 63.27%, reflecting the frequent occurrence of English tokens in standard LLM responses.

As shown in Table 2(a), SFT fails to improve language adherence, with its RPR further declining to 47.20%. This suggests that enforcing strict language constraints through traditional SFT can lead to unstable alignment. While preference-based methods such as DPO (72.73%) and ORPO (69.75%) show improvements over the baseline, TLPO achieves the highest average RPR of 77.59%. This confirms TLPO's robustness even under stringent linguistic constraints.

Regarding the general task performance presented in Table 2(b), all fine-tuning methods exhibit a notable decline in accuracy compared to the baseline (58.24%). We attribute this to the distortion

Method	LCB (cross-lingual)	LCB (monolingual)	MIF (target)	MMMLU (target)	GSM8K (cross)	Mean
Baseline	43.88(78.05)	83.48(99.65)	84.94(98.41)	43.66(58.26)	60.42(77.17)	63.27(82.31)
SFT	38.33(64.37)	62.75(98.12)	60.26(95.03)	44.61(55.20)	30.05(52.33)	47.20(73.01)
DPO	58.28(81.59)	87.91(95.92)	86.68(95.19)	55.64(65.44)	75.13(81.96)	72.73(84.02)
ORPO	50.37(82.34)	83.72(96.26)	84.41(95.44)	**57.86(71.43)**	72.37(87.07)	69.75(86.51)
TLPO(ours)	**57.25(78.66)**	**96.44(99.90)**	**96.66(99.73)**	52.25(60.33)	**85.34(89.58)**	**77.59(85.64)**

(a) Average Response Pass Rate (RPR) and Word Pass Rate (WPR). Values are presented as RPR(WPR). All values are in percentages.

Method	MIF (en)	MIF (target)	MMMLU (target)	GPQA (en)	GPQA (diamond, en)	ARC-C (en)	BBH (en)	MATH (en)	GSM8K (en)	GSM8K (cross)	Mean
Baseline	69.63	50.37	55.14	32.83	32.32	82.55	50.06	49.59	78.42	81.50	58.24
SFT	61.14	39.91	46.54	28.24	27.87	**82.81**	**56.52**	41.35	59.35	63.34	50.71
DPO	**69.26**	43.77	48.37	31.10	31.09	82.32	48.66	42.00	75.04	74.44	54.60
ORPO	65.72	42.62	50.18	**31.52**	31.19	82.77	48.93	43.78	72.86	76.50	54.61
TLPO(ours)	65.76	**46.21**	**53.83**	30.66	**31.66**	82.29	46.97	**47.73**	**77.86**	**78.71**	**56.17**

(b) Average accuracy after fine-tuning.

Table 2: Performance comparison of TLPO ($N = 16$) against SFT (Marchisio et al., 2024), DPO (Rafailov et al., 2023), and ORPO (Lee et al., 2025), treating English occurrence as language confusion. Results are averaged (RPR, WPR, and accuracy) across four models and four target languages.

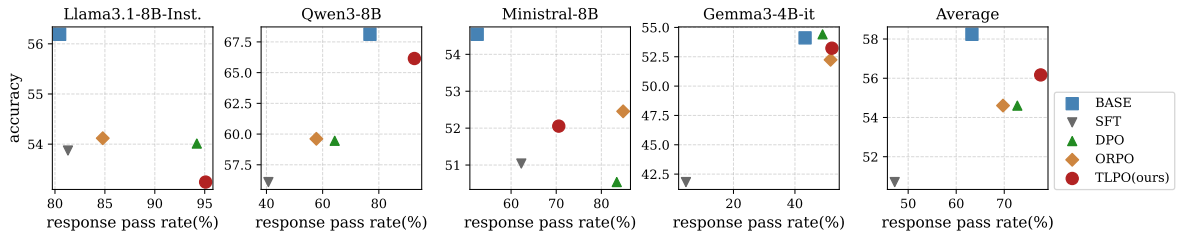


Figure 5: Scatter plot of the average Response Pass Rate (RPR) versus accuracy for each method after fine-tuning, treating English occurrence as language confusion. BASE denotes the original model prior to fine-tuning.

of the model's inherent knowledge representation when English—a primary language for reasoning and knowledge—is strictly suppressed. However, even in this environment, TLPO maintains the highest mean accuracy of 56.17%, outperforming SFT (50.71%), DPO (54.60%), and ORPO (54.61%).

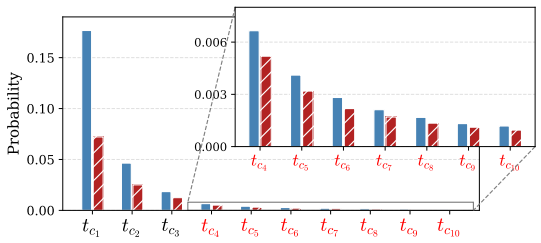
In conclusion, these results demonstrate that while strict English suppression inevitably harms model performance, TLPO provides the most favorable balance by achieving superior alignment precision with the least degradation in core model capabilities.

4.4 Probability Shifts in Tokens Beyond the Top-N Candidates

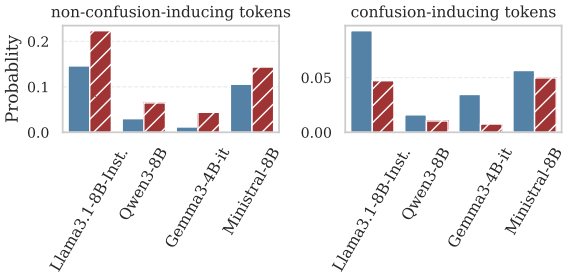
In this section, we analyze how the probabilities of tokens outside the top-N set change when parameter updates are performed using only the selected top-N tokens. To investigate this, we conducted a controlled experiment with $N = 8$, using 100 curated prompts in which exactly three of the top-8 tokens were confusion-inducing. Accordingly, parameter updates were performed on this set of eight tokens, comprising three confusion-inducing and five non-confused tokens.

Figure 6(a) illustrates the changes in output probabilities for confusion-inducing tokens before and after the parameter update. It is observed that the probabilities decrease not only for the tokens explicitly used in the optimization ($t_{c_1}, t_{c_2}, t_{c_3}$) but also for the remaining confusion-inducing tokens ($t_{c_4}, \ldots, t_{c_{10}}$) that were not included in the optimization objective.

Furthermore, we extended this analysis to the models evaluated in Section 4.3 (with $N = 16$) to investigate probability shifts for tokens ranked outside the top-N. Figure 6(b) illustrates the changes in cumulative probability for tokens outside the top-N set, distinguishing between non-confusion-inducing (Left) and confusion-inducing tokens (Right). The results demonstrate that under TLPO, the aggregated probability of non-



(a) Probability changes of confusion-inducing tokens in cases where three such tokens are included in the top-N candidate set (averaged over 100 samples). Solid blue and hatched red bars denote the probabilities before and after the policy update, respectively.



(b) Changes in the cumulative probability of tokens outside the top-N set (i.e., receiving no training signals), separated into confusion-inducing and non-confusion-inducing groups. Solid blue and hatched red bars indicate the values before and after fine-tuning, respectively.

Figure 6: The impact of TLPO on the probability distributions of tokens outside the top-N set (implicitly affected tokens).

confusion-inducing tokens increases, whereas that of confusion-inducing tokens decreases, for those not explicitly included in the top-N candidate set during fine-tuning. This indicates that the optimization effects of TLPO generalize to tokens that were not explicitly included in the optimization process.

4.5 Ablation Study on Token Selection and Advantage Formulation

We conducted an ablation study to evaluate alternative candidate token selection strategies and advantage calculation methods against TLPO. All comparisons were performed within the same framework, where candidate tokens are selected and losses are computed specifically at the confusion point.

For advantage calculation, we compared two variants against our proposed method: the unweighted formulation $A = (R-\mu)$, which excludes the token probability weight and the normalization factor from Eq. 5, and the GRPO-style advantage defined as $A = (R-\mu)/\sigma$. Regarding token selection, we compared our probability-ranked strategy

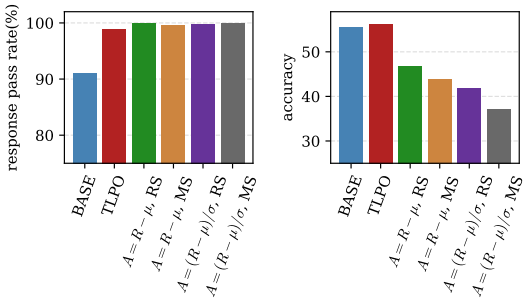


Figure 7: Performance Comparison across Different Advantage and Token Selection Strategies. In the x-axis labels, RS and MS denote ranked token selection and multinomial sampling respectively. Additionally, R, μ and σ represent the reward, mean reward and standard deviation of the reward. These results were obtained using the Llama-3.1-8B-Instruct model with Korean as the target language.

with multinomial sampling.

Figure 7 presents the experimental results. The response pass rate remained consistently stable, exceeding 99% across all settings. However, accuracy exhibited significant variation depending on the method; notably, ranked selection demonstrated superior performance compared to multinomial sampling. Furthermore, in terms of advantage calculation, the specific probability-weighted formulation employed in TLPO achieved the highest performance. We also observed that the unweighted form $A = (R - \mu)$ outperformed the original GRPO form $A = (R-\mu)/\sigma$. This finding—that excluding the standard deviation normalization yields better performance—aligns with results reported in prior research (Liu et al., 2025).

5 Conclusion

In this paper, we presented Token-Level Policy Optimization (TLPO), a fine-tuning framework designed to mitigate erroneous outputs in multilingual LLMs. Unlike sequence-level methods that optimize entire responses, TLPO operates with precision by updating the policy strictly at specific error positions, thereby resolving inconsistencies without compromising the model's general capabilities. Ultimately, this study offers a new perspective on correcting generative errors: viewing it not as a task of global sequence alignment, but as one of precise, localized adjustment. This approach suggests a promising path forward for fine-grained model alignment beyond language confusion.

Limitations

TLPO operates by identifying and rectifying local errors at specific token positions. Consequently, it is particularly effective for tasks where error boundaries are clearly defined, such as mitigating language confusion. However, extending TLPO to tasks relying on holistic sequence-level evaluations, such as general correctness or helpfulness, presents a challenge. In such contexts, errors are often diffuse rather than localized, making it difficult to derive the fine-grained supervision signals necessary for pinpointing and modifying specific tokens.

Broader Impact TLPO focuses on correcting linguistic inconsistencies at the token level and does not inherently address the safety or factual validity of the generated content. Consequently, biases or toxic patterns present in the base model or the fine-tuning datasets may be preserved or even amplified in the target language. Users of this framework should ensure that adequate safety filtering and content evaluation measures are in place to mitigate these risks.

AI Assistance Disclosure Google Gemini-3 and OpenAI GPT-5 were employed exclusively for refining the clarity and readability of this manuscript.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Paul Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Kawin Ethayarajh and 1 others. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Aaron Grattafiori and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

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

J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Teven Le Scao, Angela Fan, Christopher Akiki, and et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Nahyun Lee, Yeongseo Woo, Hyunwoo Ko, and Guijin Son. 2025. Controlling language confusion in multilingual llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation. *Preprint*, arXiv:2305.15011.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. In *Second Conference on Language Modeling*.

Yun Luo, Zhen Yang, Fandong Meng, Yanan Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during instruction tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786.

Kelly Marchisio, Wei-Yin Ko, Alexandre Bérard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677. Association for Computational Linguistics.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal intelligence. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Meta AI Blog.

Ercong Nie, Helmut Schmid, and Hinrich Schütze. 2025. Mechanistic understanding and mitigation of language confusion in english-centric large language models.

Juhyun Oh, Haneul Yoo, and Alice Oh. 2025. Evaluating llms' language confusion in code-switching context. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle*.

OpenAI. 2024. Multilingual massive multitask language understanding (mmmlu). https://huggingface.co/datasets/openai/MMMLU. Dataset.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

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

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

John Schulman. 2020. Approximating kl divergence.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and 1 others. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Dehong Xu, Liang Qiu, Minseok Kim, Faisal Ladhak, and Jaeyoung Do. 2024. Aligning large language models via fine-grained supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Bo Zeng, Chenyang Lyu, Sinuo Liu, Mingyan Zeng, Minghao Wu, Xuanfan Ni, Tianqi Shi, Yu Zhao, Yefeng Liu, Chenyu Zhu, and 1 others. 2025. Marco-bench-mif: On multilingual instruction-following capability of large language. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24058–24072.

Collin Zhang, Fei Huang, Chenhan Yuan, and Junyang Lin. 2025a. Language confusion gate: Language-aware decoding through model self-distillation. *arXiv preprint arXiv:2510.17555*.

Yang Zhang, Yu Yu, Bo Tang, Yu Zhu, Chuxiong Sun, Wenqiang Wei, Jie Hu, Zipeng Xie, Zhiyu Li, Feiyu Xiong, and Edward Chung. 2025b. Token-level accept or reject: A micro alignment approach for large language models. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI)*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Algorithm of TLPO

Algorithm 1 Token Level Policy Optimization (TLPO)

Input: initial policy model $\pi_{\theta_{\text{ref}}}$, prompt dataset $\mathcal{D}$, learning rate α, Top-N size N, training steps M, TLPO iterations p

Initialize the target parameters: $\theta \leftarrow \theta_{\text{ref}}$
for step=1,...,M **do**
  Sample an input prompt batch X from $\mathcal{D}$
  for $x \in X$ **do**
    Sample an output sequence y from $\pi_\theta(\cdot \mid x)$
    Detect the confusion point c in y
    Set $\mathcal{T} = \{t_i\}_{i=1}^{N}$ as topN tokens of $\pi_\theta(\cdot|x, y_{<c})$
    Obtain reward values $r_i \leftarrow R(t_i), \forall t_i \in \mathcal{T}$
  end for
  $\theta_{\text{old}} \leftarrow \theta$
  for TLPO iteration=1,...,p **do**
    Compute the objective $J_{\text{TLPO}}(\theta)$ via Eq. (4)
    Compute gradient $g \leftarrow \nabla_\theta J_{\text{TLPO}}(\theta)$
    Update policy parameters: $\theta \leftarrow \theta + \alpha \cdot g$
  end for
end for

Output: optimized policy π_θ

Algorithm 1 outlines the overall procedure of Token-Level Policy Optimization (TLPO). The TLPO algorithm comprises two primary phases: an **exploration phase**, which detects a confusion point and selects candidate tokens at that position, and a **policy update phase**, which optimizes the policy using these selected candidates.

Language	Original training #instances	Filtered training #instances
Chinese(zh)	67,017	65,676
Arabic(ar)	67,017	65,907
Korean(ko)	67,017	62,679
Japanese(ja)	67,017	65,296

Table 3: Number of training instances.

Exploration (Lines 3–9) First, a batch of prompts X is sampled from the training dataset $\mathcal{D}$. For each prompt $x \in X$, the model generates an initial response y. The algorithm then detects a confusion point c within y. If no confusion is detected, the corresponding sample is discarded. Conversely, if a confusion point is identified, the top-N tokens are selected from the current policy distribution $\pi_\theta(\cdot|x, y_{<c})$ to form the candidate token set $\mathcal{T}$. Subsequently, a reward is assigned to each candidate token based on a short lookahead rollout of length k.

Policy Update (Lines 10–15) Once rewards for all candidate tokens in the batch are collected, the policy parameters θ are updated. Specifically, we perform p optimization iterations to maximize the TLPO objective function $\mathcal{J}_{\text{TLPO}}(\theta)$, following a methodology similar to Proximal Policy Optimization (PPO).

B Training Dataset Construction

To construct the fine-tuning dataset for mitigating language confusion, we utilized the training split of the Bactrian-X dataset for each target language. Prompts were formed by concatenating the *instruction* and *input* fields from the dataset, while the *output* field served as the answer.

To exclude prompts that explicitly induce generation in other languages (e.g., translation requests), we filtered out any prompts containing characters that do not belong to the target language. Table 3 presents the number of prompts in the original dataset and the final number of prompts used for training after this filtering process.

The specific configurations for each fine-tuning method are as follows:

SFT We conducted Supervised Fine-Tuning (SFT) using the filtered prompts and their corresponding answers from the Bactrian-X dataset for each target language.

DPO and ORPO For these methods, we generated 16 candidate responses for each prompt using the respective target models. We then constructed preference pairs for fine-tuning by selecting one response without *language confusion* as the **preferred** response and one exhibiting confusion as the **dispreferred** response.

TLPO For TLPO, fine-tuning was performed by generating responses online based on the prompts from the training set. Specifically, we generated a single full response per prompt.

C Training Configurations

TLPO fine-tuning was performed using a unified hyperparameter configuration across all models and target languages. The specific settings are as follows:

- **General Training Settings:** All experiments were conducted for 1 epoch. The batch size was set to 8, representing the number of prompts processed in a single step. However, since the loss is computed over 16 candidate tokens ($N = 16$) at each confusion point, the parameters are updated based on a total of $8 \times 16 = 128$ tokens. We used an initial learning rate of 5×10^{-7} with a cosine decay schedule, reducing the rate to 10% of the initial value by the end of training. A warmup period covering the first 10% of total steps was applied.
- **TLPO-Specific Parameters:** We set the number of policy iterations to 2 ($p = 2$). At each language confusion point, we explored 16 candidate tokens ($N = 16$) and employed a lookahead length of 3 tokens ($k = 3$) for reward calculation.

D Computational Cost and Training Time

Table 4 presents the wall-clock time required for TLPO fine-tuning across different model and language configurations. Fine-tuning was performed for a single epoch, and the results indicate that training is completed within approximately 6 to 10 hours depending on the specific model and language pair.

E Evaluation Methodology

Table 5 presents the tasks used for WPR, RPR, and Accuracy evaluation, along with the number of instances for each task. All evaluations were conducted as generative tasks under a zero-shot Chain-of-Thought (CoT) setting.

For the consistency evaluation on the MIF dataset, we excluded specific instances where the instruction explicitly requires generating output in a different language, ensuring that the metric accurately reflects unintended language confusion.

In the case of GSM8K (cross), we adapted the prompt template from lm-eval-harness (specifically gsm8k-cot-llama). The English instructions were translated into each target language to serve as the prompts. Figure 8 illustrates the specific instructions used for each target language.

Lang.	Fine-tuning Time (hours)			
	Llama3.1 -8B-Inst.	Qwen3 -8B	Ministral -8B	Gemma3 -4B-it
zh	6.81	7.95	7.04	6.80
ar	6.20	7.81	7.91	5.78
ko	8.27	8.97	7.95	6.26
ja	7.05	9.59	7.27	8.83

Table 4: TLPO fine-tuning time for each model and target language. All experiments were conducted using 8x NVIDIA H100 GPUs.

Task	WPR /RPR	Acc.	#Instances
LCB(cross-lingual)	O		299
LCB(monolingual)	O		200/300/100/100
MIF(target lang.)	O	O	420/421/422/421
MMMLU(target lang.)	O	O	14,042
GSM8K(cross)	O	O	1,209
MIF(en)		O	541
GPQA(en)		O	448
GPQA(diamond, en)		O	198
ARC-C(en)		O	1,172
BBH(en)		O	6,511
MATH(en)		O	5,000
GSM8K(en)		O	1,209

Table 5: List of tasks used for consistency (WPR/RPR) and accuracy evaluation. For LCB (monolingual) and MIF (target), the values for #instances are presented in the order of zh/ar/ko/ja.

F Details of the Language Confusion Detector

We devised a rule-based heuristic to detect language confusion. This detector operates by analyzing an LLM-generated response to count the number of words exhibiting language confusion

"en": "Given the following problem, reason and give a final answer to the problem.\nProblem: {question}\nYour response should end with \"The final answer is [answer]\" where [answer] is the response to the problem.\n",

"zh":
"针对给定问题说明理由并给出最终答案。\n问题：{question}\n您的回答应以以下句子结尾：“问题的答案是[正确答案]。”此处[正确答案]即为给定问题的答案。\n",

"ar": "قيئاهنلا قباجإلا مدقو ةاطعملا ةلأسملا لح ةقيرط حرشا. ةلأسملا: {question} يلي بجي نأ يهتنت قيلاتلا ةلمجلاب كتباجإ: “ةلأسملا هذه قباجإ [ةحيحصلا قباجإلا]”. هنإ، [ةحيحصلا قباجإلا] ةاطعملا ةلأسملا لح يه.\n",

"ko":"주어진 문제에 대해 이유를 설명하고 최종 답을 제시하십시오.\n문제: {question}\n당신의 응답은 다음 문장으로 끝나야 합니다.\"문제의 답은 [정답]입니다.\" 여기서 [정답]은 주어진 문제의 답입니다.\n",

"ja": "与えられた問題について理由を説明し、最終的な答えを提示してください。\n問題: {question}\nあなたの回答は次の文で終わらなければなりません。「問題の答えは[正解]です。」 ここで[正解]は与えられた問題の答えです。\n",

Figure 8: Instruction used for GSM8K(en) and GSM8K(cross). Note that {question} represents the original English question from the GSM8K dataset, which remains untranslated.

versus those that do not. Based on these counts, we subsequently calculate the Word Pass Rate (WPR) and Response Pass Rate (RPR).

For word segmentation, we utilized the `jieba` library for Chinese and a Python-based tagger library for Japanese. For all other languages, segmentation was performed based on whitespace characters.

We determined whether each character within a word belongs to the target language by referencing its Unicode metadata (e.g., script/block information derived from character names). A word was classified as free of language confusion only if all its constituent characters belonged to the target language. Conversely, if a word contained one or more characters not belonging to the target language, it was classified as exhibiting language confusion.

During the detection process, we applied several exclusion rules to prevent false detections:

- **Word/Line-level exclusions:** Units of measurement denoted in the alphabet, strings identified as function names, email addresses, and URLs were excluded from detection.

- **Character-level exclusions:** We also excluded phonetic symbols, words starting with a capital letter (indicating proper nouns), mathematical symbols, currency symbols, arrows, Chinese tone marks, and emojis.

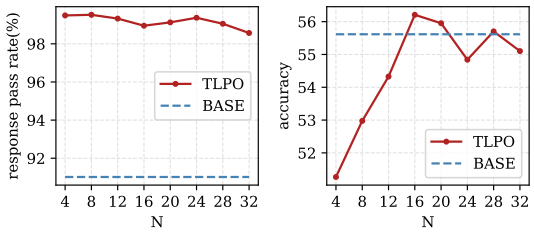


Figure 9: Response Pass Rate and Accuracy across Various Numbers of Candidate Tokens (N). This experiment was conducted using the Llama3.1-8B-Instruct model with Korean as the target language.

To validate the effectiveness of these rules, we conducted a systematic error analysis. First, the likelihood of **False Negatives (FN)** is practically zero due to the deterministic nature of Unicode script mapping; every character belonging to the target script is correctly identified without exception. Second, a manual inspection of 10,937 detected instances revealed only 9 **False Positives (FP)**, yielding a remarkably low error rate of 0.08%. These rare FPs were primarily caused by uncommon emojis not yet included in our exclusion list. Overall, this high level of precision ensures that our detector provides a clean and accurate signal for optimizing multilingual alignment.

G Performance Sensitivity to the Number of Candidate Tokens

Figure 9 illustrates the variations in response pass rate and accuracy as the number of candidate tokens N explored at the confusion point c changes. The response pass rate remains consistently high at over 99% regardless of N, showing negligible fluctuation. In contrast, accuracy exhibits a notable decline in the range of $N \leq 12$, dropping by 4.4 pp, 2.6 pp and 1.29 pp compared to the Baseline at $N = 4, 8$ and $N = 12$, respectively. However, for $N \geq 16$, the decrease in accuracy narrows to less than 1 pp, demonstrating that performance is stably preserved.

H Supplementary Results on GRPO Fine-tuning Dynamics

Figure 10 illustrates the dynamics of response length and reward throughout the GRPO fine-tuning process. For Llama-3.1-8B-Instruct and

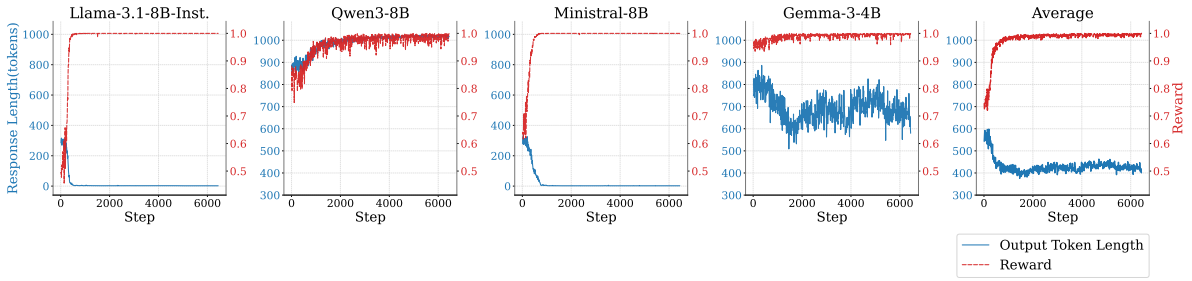


Figure 10: Response length and reward dynamics during the GRPO fine-tuning process. Blue lines (left axis) and red lines (right axis) represent response length and reward, respectively.

Ministral-8B, a sharp decline in response length was observed immediately upon the start of fine-tuning, while Gemma3-4B exhibited a reduction in output length to approximately three-quarters of its initial size. Conversely, Qwen3-8B was the only model that maintained stable token length without such degradation.

These experiments were conducted under a setting where English occurrences are treated as a neutral category rather than language confusion. In this setup, we assigned a reward of -1 for instances of language confusion (non-target languages excluding English) and +1 for proper linguistic adherence. We attribute the observed length reduction to the model's exploitation of the reward structure; by shortening its responses, the model effectively minimizes the accumulation of negative rewards. Due to this model collapse phenomenon, we excluded GRPO from the primary baselines in our main experimental results.

I Detailed Experimental Results by Model and Target Language

Figure 11 presents the Response Pass Rate (RPR) and accuracy for each model across different target languages after fine-tuning to mitigate language confusion, under the setting where English is treated as a neutral language. The results demonstrate that in most configurations, TLPO effectively mitigates language confusion while minimizing the degradation of the LLM's performance more consistently than other comparative methods. Tables 6 through Tables 10 provide detailed RPR, WPR, and accuracy results for each downstream task, broken down by model and target language after fine-tuning.

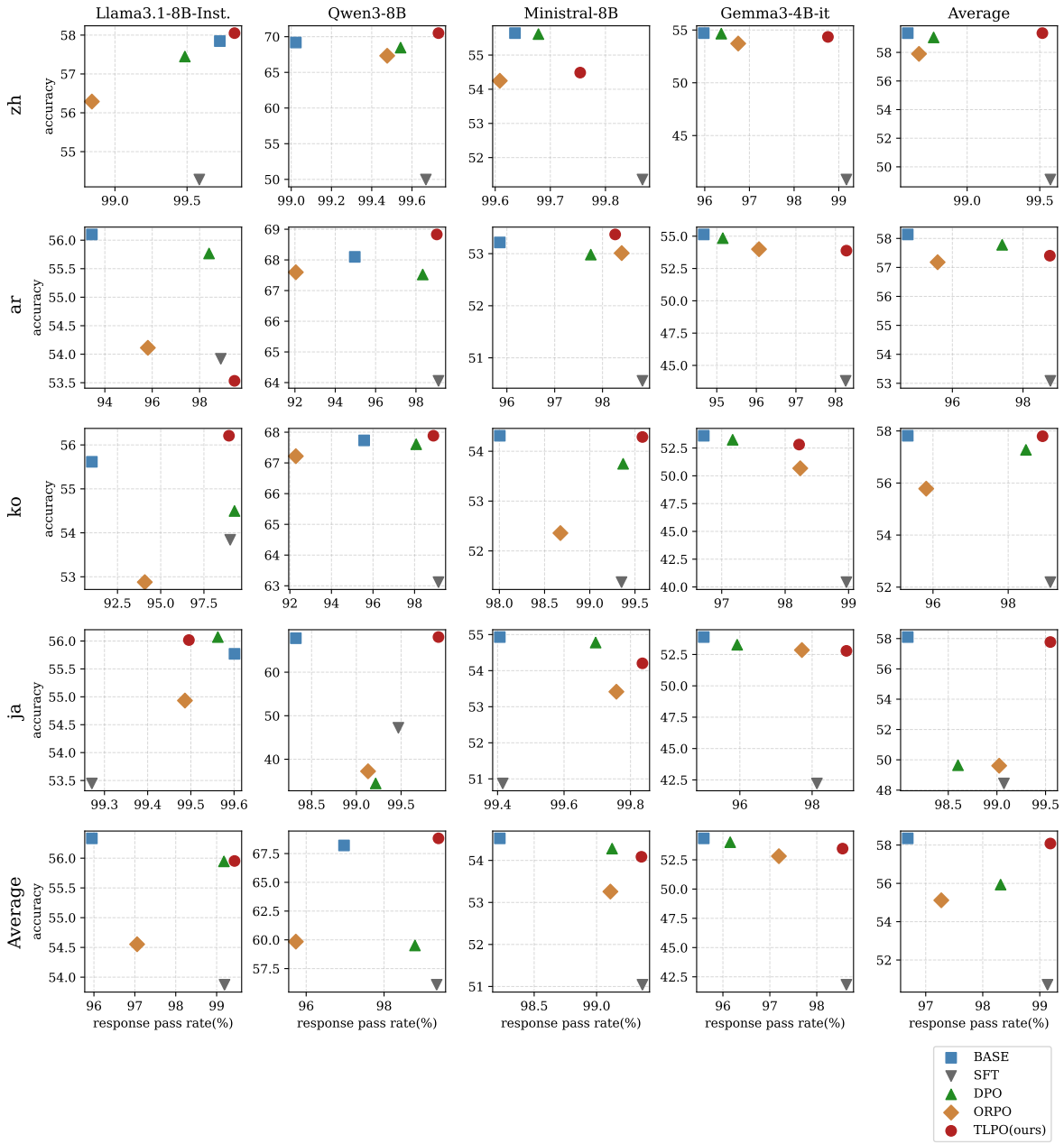


Figure 11: Response Pass Rate (RPR) and accuracy plots across models and target languages, under the setting where English is treated as neutral.

Lang.	Method	LCB (cross-lingual)	LCB (monolingual)	MIF (target)	MMMLU (target)	GSM8K (cross)	Mean
zh	Baseline	98.96(100.00)	100.00(100.00)	100.00(100.00)	99.76(100.00)	99.92(100.00)	99.73(100.00)
	SFT	98.33(99.99)	100.00(100.00)	99.76(100.00)	99.92(100.00)	99.92(99.99)	99.59(100.00)
	DPO	98.26(99.17)	100.00(100.00)	99.52(99.99)	99.72(99.89)	99.92(99.82)	99.48(99.77)
	ORPO	96.91(99.96)	98.50(99.98)	99.29(99.99)	99.67(99.98)	99.83(100.00)	98.84(99.98)
	TLPO(ours)	99.29(100.00)	100.00(100.00)	100.00(100.00)	99.86(100.00)	100.00(100.00)	99.83(100.00)
ar	Baseline	86.21(99.93)	94.33(99.97)	95.49(99.97)	96.69(99.97)	94.54(99.86)	93.45(99.94)
	SFT	96.23(99.98)	99.33(100.00)	100.00(100.00)	99.91(100.00)	99.00(99.97)	98.89(99.99)
	DPO	95.92(99.01)	99.33(99.99)	99.29(99.87)	99.19(99.40)	98.26(98.16)	98.40(99.29)
	ORPO	90.69(99.95)	95.67(99.97)	98.57(99.99)	97.78(99.98)	96.36(99.80)	95.82(99.94)
	TLPO(ours)	97.81(99.97)	100.00(100.00)	100.00(100.00)	99.80(100.00)	99.75(100.00)	99.47(99.99)
ko	Baseline	87.76(99.81)	86.00(99.85)	92.58(99.94)	94.86(99.93)	93.88(99.70)	91.02(99.84)
	SFT	97.60(99.98)	100.00(100.00)	98.52(99.98)	99.65(99.98)	99.34(99.93)	99.02(99.98)
	DPO	97.56(99.99)	100.00(100.00)	99.52(99.99)	99.59(99.99)	99.67(99.99)	99.27(99.99)
	ORPO	91.10(99.77)	94.00(99.94)	94.51(99.89)	96.00(99.93)	94.79(99.71)	94.08(99.85)
	TLPO(ours)	96.98(99.88)	100.00(100.00)	99.04(99.99)	99.47(99.99)	99.26(99.95)	98.95(99.96)
ja	Baseline	98.51(99.99)	100.00(100.00)	99.76(100.00)	99.74(100.00)	100.00(100.00)	99.60(100.00)
	SFT	97.01(99.96)	100.00(100.00)	99.51(100.00)	99.92(100.00)	99.91(100.00)	99.27(99.99)
	DPO	99.59(100.00)	99.00(100.00)	99.52(99.98)	99.78(99.97)	99.92(99.96)	99.56(99.98)
	ORPO	98.13(99.99)	100.00(100.00)	99.52(99.95)	99.77(100.00)	100.00(100.00)	99.49(99.99)
	TLPO(ours)	98.08(99.99)	100.00(100.00)	99.52(99.94)	99.88(100.00)	100.00(100.00)	99.50(99.99)
avg.	Baseline	92.86(99.93)	95.08(99.95)	96.96(99.98)	97.76(99.97)	97.08(99.89)	95.95(99.94)
	SFT	97.29(99.98)	99.83(100.00)	99.45(99.99)	99.85(99.99)	99.54(99.97)	99.19(99.99)
	DPO	97.83(99.54)	99.58(100.00)	99.46(99.96)	99.57(99.81)	99.44(99.48)	99.18(99.76)
	ORPO	94.21(99.92)	97.04(99.97)	97.97(99.95)	98.31(99.97)	97.75(99.88)	97.06(99.94)
	TLPO(ours)	98.04(99.96)	100.00(100.00)	99.64(99.98)	99.75(100.00)	99.75(99.99)	99.44(99.99)

(a) Average Response Pass Rate(RPR) and Word Pass Rate(WPR). Values are presented as RPR(WPR) in %.

Lang.	Method	MIF (en)	MIF (target)	MMMLU (target)	GPQA (en)	GPQA-D (en)	ARC-C (en)	BBH (en)	MATH (en)	GSM8K (en)	GSM8K (cross)	Mean
zh	Baseline	74.68	51.57	55.32	27.46	26.26	83.79	51.27	49.22	79.24	79.65	57.85
	SFT	71.35	48.06	41.92	25.67	22.73	84.13	56.90	43.34	76.76	71.96	54.28
	DPO	75.42	50.46	53.68	26.34	26.26	83.70	51.05	48.76	78.91	79.90	57.45
	ORPO	72.09	45.10	49.47	26.34	25.76	83.79	51.88	49.50	78.66	80.31	56.29
	TLPO(ours)	73.57	54.90	55.73	23.88	26.77	83.02	52.79	49.56	79.90	80.40	58.05
ar	Baseline	74.31	46.03	47.19	26.34	28.28	83.62	51.70	49.78	78.91	74.86	56.10
	SFT	71.35	47.69	36.41	28.79	25.25	83.96	58.35	44.30	77.34	65.76	53.92
	DPO	73.75	44.36	43.62	29.46	29.29	83.53	52.05	48.82	79.74	73.04	55.77
	ORPO	70.98	40.30	39.61	26.56	26.26	84.13	49.13	48.90	78.16	77.09	54.11
	TLPO(ours)	70.61	38.08	42.62	26.34	30.30	83.70	52.10	47.08	79.32	65.18	53.53
ko	Baseline	74.86	40.85	49.15	26.79	25.25	83.70	50.98	49.16	79.57	75.85	55.62
	SFT	73.94	41.22	34.52	27.46	29.80	83.79	57.96	43.66	76.51	69.56	53.84
	DPO	75.42	40.30	46.78	25.22	22.73	83.70	50.55	48.16	78.99	73.12	54.50
	ORPO	74.31	31.98	34.64	26.34	24.24	84.13	50.28	48.10	78.83	75.93	52.88
	TLPO(ours)	71.90	46.40	47.51	27.23	28.28	83.62	52.00	49.14	79.40	76.59	56.21
ja	Baseline	75.05	41.59	50.44	25.00	25.25	83.70	52.00	49.74	78.00	76.92	55.77
	SFT	73.38	41.40	37.16	29.46	25.76	83.79	57.70	43.38	76.18	66.25	53.45
	DPO	73.75	45.47	49.95	23.66	29.29	83.79	51.13	48.58	79.98	75.10	56.07
	ORPO	72.46	36.97	48.80	24.55	23.74	83.62	51.96	50.08	78.99	78.16	54.93
	TLPO(ours)	73.75	42.70	50.11	27.01	26.77	84.04	51.28	49.00	79.16	76.34	56.02
avg.	Baseline	74.72	45.01	50.53	26.40	26.26	83.70	51.49	49.47	78.93	76.82	56.33
	SFT	72.50	44.59	37.50	27.85	25.88	83.92	57.73	43.67	76.70	68.38	53.87
	DPO	74.59	45.15	48.51	26.17	26.89	83.68	51.20	48.58	79.40	75.29	55.95
	ORPO	72.46	38.59	43.13	25.95	25.00	83.92	50.81	49.14	78.66	77.87	54.55
	TLPO(ours)	72.46	45.52	48.99	26.12	28.03	83.60	52.04	48.70	79.45	74.63	55.95

(b) Average accuracy after fine-tuning.

Table 6: Detailed RPR, WPR, and accuracy results for the Llama3.1-8B-Instruction model after fine-tuning, in a setting where English output is regarded as neutral.

Lang.	Method	LCB (cross-lingual)	LCB (monolingual)	MIF (target)	MMMLU (target)	GSM8K (cross)	Mean
zh	Baseline	98.18(99.99)	98.00(99.99)	99.28(100.00)	99.65(99.99)	100.00(100.00)	99.02(100.00)
	SFT	98.50(99.99)	100.00(100.00)	100.00(100.00)	99.93(100.00)	99.92(99.99)	99.67(100.00)
	DPO	98.55(100.00)	99.50(100.00)	100.00(100.00)	99.67(99.99)	100.00(100.00)	99.54(100.00)
	ORPO	98.11(99.99)	100.00(100.00)	99.76(100.00)	99.59(100.00)	99.92(100.00)	99.48(100.00)
	TLPO(ours)	98.83(100.00)	100.00(100.00)	100.00(100.00)	99.83(100.00)	100.00(100.00)	99.73(100.00)
ar	Baseline	91.67(99.96)	94.67(99.97)	95.96(99.92)	96.10(99.97)	96.44(99.91)	94.97(99.94)
	SFT	97.37(99.98)	99.67(99.99)	99.76(99.87)	99.61(99.99)	99.17(99.97)	99.12(99.96)
	DPO	96.42(99.98)	99.00(99.99)	99.52(99.99)	98.23(99.98)	98.51(99.94)	98.34(99.98)
	ORPO	86.59(99.92)	90.33(99.90)	95.72(99.96)	93.36(99.94)	94.21(99.86)	92.05(99.92)
	TLPO(ours)	98.18(99.99)	99.00(99.99)	99.52(100.00)	99.45(100.00)	99.01(99.97)	99.03(99.99)
ko	Baseline	94.36(99.96)	95.00(99.94)	96.89(99.96)	94.68(99.95)	96.94(99.81)	95.57(99.92)
	SFT	96.89(99.85)	100.00(100.00)	99.51(99.99)	99.52(99.98)	99.83(99.98)	99.15(99.96)
	DPO	96.62(99.98)	98.00(99.98)	99.28(99.99)	98.07(99.98)	98.43(99.94)	98.08(99.97)
	ORPO	88.06(99.90)	93.00(99.91)	95.20(99.92)	89.63(99.88)	95.53(99.74)	92.29(99.87)
	TLPO(ours)	97.69(99.99)	99.00(99.99)	99.28(99.98)	99.23(99.99)	99.34(99.94)	98.91(99.98)
ja	Baseline	98.57(100.00)	95.00(99.98)	98.81(99.98)	99.58(100.00)	99.67(99.99)	98.32(99.99)
	SFT	97.90(99.99)	100.00(100.00)	99.52(99.99)	99.93(100.00)	100.00(100.00)	99.47(100.00)
	DPO	98.90(99.99)	98.99(99.99)	99.49(99.99)	98.99(99.95)	99.71(99.98)	99.22(99.98)
	ORPO	98.55(99.98)	100.00(100.00)	98.52(99.92)	99.09(99.96)	99.50(99.97)	99.13(99.97)
	TLPO(ours)	100.00(100.00)	100.00(100.00)	99.76(100.00)	99.83(100.00)	100.00(100.00)	99.92(100.00)
avg.	Baseline	95.69(99.98)	95.67(99.97)	97.74(99.96)	97.50(99.98)	98.26(99.93)	96.97(99.96)
	SFT	97.66(99.95)	99.92(100.00)	99.70(99.96)	99.75(99.99)	99.73(99.99)	99.35(99.98)
	DPO	97.62(99.99)	98.87(99.99)	99.58(99.99)	98.74(99.98)	99.16(99.96)	98.79(99.98)
	ORPO	92.83(99.95)	95.83(99.95)	97.30(99.95)	95.42(99.95)	97.29(99.89)	95.73(99.94)
	TLPO(ours)	98.68(99.99)	99.50(100.00)	99.64(99.99)	99.58(100.00)	99.59(99.98)	99.40(99.99)

(a) Average Response Pass Rate(RPR) and Word Pass Rate(WPR). Values are presented as RPR(WPR) in %.

Lang.	Method	MIF (en)	MIF (target)	MMMLU (target)	GPQA (en)	GPQA-D (en)	ARC-C (en)	BBH (en)	MATH (en)	GSM8K (en)	GSM8K (cross)	Mean
zh	Baseline	82.07	67.28	73.69	45.76	46.46	90.44	40.35	74.74	78.33	92.72	69.18
	SFT	55.27	41.22	62.11	33.93	32.32	91.47	55.63	10.06	38.30	79.49	49.98
	DPO	81.33	65.80	72.97	45.09	45.96	90.53	40.04	74.38	75.85	92.72	68.47
	ORPO	81.89	64.51	73.05	43.75	44.95	91.13	41.91	75.60	63.94	92.56	67.33
	TLPO(ours)	82.62	67.47	72.38	45.54	51.01	90.78	46.97	74.30	82.38	91.56	70.50
ar	Baseline	82.26	67.65	64.09	45.76	46.46	90.44	40.35	74.74	78.33	90.90	68.10
	SFT	78.37	54.53	53.30	37.72	38.38	91.30	59.68	73.44	79.40	74.44	64.06
	DPO	81.33	66.73	64.04	40.85	48.48	90.44	40.16	74.64	77.75	90.82	67.52
	ORPO	81.70	67.65	64.62	43.97	48.99	90.96	40.59	75.54	71.30	90.65	67.60
	TLPO(ours)	82.99	67.28	63.67	44.42	48.48	90.61	41.76	74.80	84.12	90.16	68.83
ko	Baseline	82.07	61.18	66.02	45.76	46.46	90.44	40.35	74.74	78.33	91.98	67.73
	SFT	76.34	50.83	55.65	33.93	32.83	91.38	57.99	72.60	79.24	80.40	63.12
	DPO	81.15	59.70	65.68	43.08	51.52	90.44	39.75	75.04	78.16	91.56	67.61
	ORPO	80.96	61.18	65.77	43.08	51.01	90.78	40.55	75.42	71.22	92.22	67.22
	TLPO(ours)	83.18	60.63	65.57	43.30	49.49	90.53	41.38	73.88	80.73	90.16	67.88
ja	Baseline	82.26	62.11	67.68	45.76	46.46	90.44	40.35	74.74	78.33	89.99	67.81
	SFT	58.41	39.00	56.45	29.69	29.29	91.30	51.68	10.98	32.51	72.95	47.23
	DPO	34.01	2.22	45.66	27.23	27.27	90.61	31.73	7.94	26.55	51.20	34.44
	ORPO	35.30	2.59	50.49	32.81	29.80	90.96	33.33	11.82	25.56	59.72	37.24
	TLPO(ours)	79.85	55.45	68.04	43.97	47.98	90.61	44.99	74.52	83.71	91.65	68.08
avg.	Baseline	82.16	64.56	67.87	45.76	46.46	90.44	40.35	74.74	78.33	91.40	68.21
	SFT	67.10	46.40	56.87	33.82	33.21	91.36	56.25	41.77	57.36	76.82	56.10
	DPO	69.46	48.61	62.09	39.06	43.31	90.51	37.92	58.00	64.58	81.57	59.51
	ORPO	69.96	48.98	63.48	40.90	43.69	90.95	39.10	59.59	58.00	83.79	59.85
	TLPO(ours)	82.16	62.71	67.41	44.31	49.24	90.64	43.77	74.38	82.73	90.88	68.82

(b) Average accuracy after fine-tuning.

Table 7: Detailed RPR, WPR, and accuracy results for the Qwen3-8B model after fine-tuning, in a setting where English output is regarded as neutral.

Lang.	Method	LCB (cross-lingual)	LCB (monolingual)	MIF (target)	MMMLU (target)	GSM8K (cross)	Mean
zh	Baseline	98.74(99.99)	100.00(100.00)	99.76(100.00)	99.68(99.95)	100.00(100.00)	99.64(99.99)
	SFT	99.49(100.00)	100.00(100.00)	100.00(100.00)	99.84(99.97)	100.00(100.00)	99.87(99.99)
	DPO	98.70(99.99)	100.00(100.00)	100.00(100.00)	99.69(99.96)	100.00(100.00)	99.68(99.99)
	ORPO	98.27(99.99)	100.00(100.00)	100.00(100.00)	99.77(99.98)	100.00(100.00)	99.61(99.99)
	TLPO(ours)	98.90(100.00)	100.00(100.00)	100.00(100.00)	99.88(99.99)	100.00(100.00)	99.75(100.00)
ar	Baseline	89.66(99.93)	96.00(99.97)	98.07(99.99)	97.61(99.62)	97.93(99.94)	95.85(99.89)
	SFT	96.03(99.97)	99.67(100.00)	100.00(100.00)	99.25(99.85)	99.22(99.97)	98.83(99.96)
	DPO	97.44(98.13)	95.67(98.23)	98.80(97.87)	97.12(94.62)	99.75(99.85)	97.75(97.74)
	ORPO	96.55(99.99)	98.33(99.99)	99.28(99.98)	98.34(99.74)	99.50(99.96)	98.40(99.93)
	TLPO(ours)	92.96(99.96)	99.33(100.00)	100.00(100.00)	99.03(99.82)	100.00(100.00)	98.26(99.95)
ko	Baseline	95.79(99.93)	100.00(100.00)	97.82(99.98)	97.66(99.87)	98.75(99.95)	98.01(99.94)
	SFT	98.24(99.99)	100.00(100.00)	99.50(99.99)	99.47(99.92)	99.56(99.98)	99.35(99.98)
	DPO	98.82(100.00)	100.00(100.00)	99.76(99.83)	98.57(98.22)	99.69(99.97)	99.37(99.60)
	ORPO	97.24(98.33)	100.00(100.00)	98.56(99.23)	98.21(99.06)	99.36(99.46)	98.68(99.21)
	TLPO(ours)	98.40(99.99)	100.00(100.00)	100.00(100.00)	99.52(99.98)	100.00(100.00)	99.58(99.99)
ja	Baseline	99.19(100.00)	99.00(100.00)	99.28(99.97)	99.65(99.99)	99.92(99.99)	99.41(99.99)
	SFT	98.28(99.20)	100.00(100.00)	99.25(99.97)	99.74(99.97)	99.81(99.99)	99.42(99.83)
	DPO	99.19(100.00)	100.00(100.00)	99.76(99.97)	99.61(99.98)	99.92(99.98)	99.70(99.99)
	ORPO	99.25(100.00)	100.00(100.00)	99.76(99.98)	99.77(100.00)	100.00(100.00)	99.76(100.00)
	TLPO(ours)	99.57(100.00)	100.00(100.00)	99.76(99.99)	99.85(99.99)	100.00(100.00)	99.84(100.00)
avg.	Baseline	95.84(99.96)	98.75(99.99)	98.73(99.98)	98.65(99.86)	99.15(99.97)	98.22(99.95)
	SFT	98.01(99.79)	99.92(100.00)	99.69(99.99)	99.58(99.93)	99.65(99.99)	99.37(99.94)
	DPO	98.54(99.53)	98.92(99.56)	99.58(99.42)	98.75(98.19)	99.84(99.95)	99.12(99.33)
	ORPO	97.83(99.58)	99.58(100.00)	99.40(99.80)	99.02(99.69)	99.72(99.85)	99.11(99.78)
	TLPO(ours)	97.45(99.99)	99.83(100.00)	99.94(100.00)	99.57(99.94)	100.00(100.00)	99.36(99.99)

(a) Average Response Pass Rate(RPR) and Word Pass Rate(WPR). Values are presented as RPR(WPR) in %.

Lang.	Method	MIF (en)	MIF (target)	MMMLU (target)	GPQA (en)	GPQA-D (en)	ARC-C (en)	BBH (en)	MATH (en)	GSM8K (en)	GSM8K (cross)	Mean
zh	Baseline	53.60	44.36	50.28	30.80	26.77	80.97	52.57	52.38	79.40	85.28	55.64
	SFT	49.35	39.37	43.81	27.68	22.22	81.48	55.46	40.60	78.41	75.19	51.36
	DPO	54.53	44.55	49.90	29.69	27.78	81.06	52.02	52.52	79.74	84.37	55.61
	ORPO	52.87	41.40	46.75	25.89	24.75	81.66	52.22	52.60	78.58	85.77	54.25
	TLPO(ours)	50.83	42.70	48.14	25.67	28.79	81.14	51.71	51.78	78.91	85.19	54.49
ar	Baseline	53.60	36.23	43.16	30.80	26.77	80.97	52.57	52.38	79.40	76.26	53.22
	SFT	52.68	32.16	37.32	28.79	29.29	81.91	55.41	40.94	78.49	68.57	50.56
	DPO	56.01	34.94	40.95	29.69	25.76	81.40	52.86	51.64	78.33	78.25	52.98
	ORPO	51.76	35.49	42.96	30.13	33.84	81.66	50.81	53.42	79.16	70.89	53.01
	TLPO(ours)	54.34	36.60	39.86	29.69	29.29	81.06	51.97	51.32	79.07	80.48	53.37
ko	Baseline	53.42	39.19	47.54	30.80	26.77	80.97	52.57	52.38	79.40	80.07	54.31
	SFT	51.39	35.12	41.61	26.12	30.81	81.57	54.49	40.36	78.66	73.61	51.37
	DPO	55.64	39.74	46.00	27.01	27.78	80.89	51.80	51.22	79.82	77.58	53.75
	ORPO	52.50	35.86	45.34	29.69	22.73	81.06	49.79	52.02	77.83	76.76	52.36
	TLPO(ours)	52.87	37.15	47.57	31.70	31.31	81.48	52.03	53.24	78.16	77.34	54.29
ja	Baseline	53.60	39.37	49.24	30.80	26.77	80.97	52.57	52.38	79.40	84.20	54.93
	SFT	52.87	31.98	42.24	25.22	25.25	81.91	55.14	42.34	79.24	72.54	50.87
	DPO	54.34	41.04	47.25	31.92	27.27	81.06	50.62	51.42	78.66	84.20	54.78
	ORPO	53.42	34.01	48.62	27.23	26.77	81.23	51.04	52.62	78.74	80.48	53.42
	TLPO(ours)	52.31	37.34	49.24	27.23	29.80	81.14	52.23	51.94	78.66	82.13	54.20
avg.	Baseline	53.56	39.79	47.55	30.80	26.77	80.97	52.57	52.38	79.40	81.45	54.52
	SFT	51.57	34.66	41.25	26.95	26.89	81.72	55.13	41.06	78.70	72.48	51.04
	DPO	55.13	40.06	46.03	29.58	27.15	81.10	51.83	51.70	79.14	81.10	54.28
	ORPO	52.64	36.69	45.92	28.24	27.02	81.40	50.96	52.66	78.58	78.47	53.26
	TLPO(ours)	52.59	38.45	46.20	28.57	29.80	81.21	51.99	52.07	78.70	81.29	54.09

(b) Average accuracy after fine-tuning.

Table 8: Detailed RPR, WPR, and accuracy results for the Ministral-8B model after fine-tuning, in a setting where English output is regarded as neutral.

Lang.	Method	LCB (cross-lingual)	LCB (monolingual)	MIF (target)	MMMLU (target)	GSM8K (cross)	Mean
zh	Baseline	88.73(99.93)	95.00(99.95)	97.83(99.98)	99.13(99.49)	99.21(99.85)	95.98(99.84)
	SFT	97.36(99.96)	100.00(100.00)	99.76(100.00)	98.83(98.67)	99.91(100.00)	99.17(99.72)
	DPO	91.30(99.95)	94.50(99.95)	97.36(99.98)	99.22(99.61)	99.47(99.90)	96.37(99.88)
	ORPO	89.01(99.87)	99.00(99.99)	97.12(99.97)	98.86(99.05)	99.75(99.92)	96.75(99.76)
	TLPO(ours)	95.96(99.97)	99.50(100.00)	98.80(99.99)	99.53(99.94)	100.00(100.00)	98.76(99.98)
ar	Baseline	83.81(98.64)	95.00(99.95)	96.90(99.98)	98.23(99.81)	99.42(99.97)	94.67(99.67)
	SFT	95.26(99.72)	99.67(100.00)	99.76(100.00)	96.60(98.35)	100.00(100.00)	98.26(99.61)
	DPO	82.73(98.44)	97.00(99.98)	98.33(99.99)	98.20(99.80)	99.50(99.97)	95.15(99.64)
	ORPO	86.59(99.14)	97.33(99.84)	98.57(99.99)	98.33(99.85)	99.50(99.99)	96.07(99.76)
	TLPO(ours)	95.40(99.79)	98.67(99.99)	99.05(99.99)	98.84(99.90)	99.42(99.99)	98.28(99.93)
ko	Baseline	96.74(99.96)	98.00(99.98)	96.90(99.95)	98.30(99.85)	93.65(99.87)	96.72(99.92)
	SFT	97.04(99.92)	100.00(100.00)	99.75(100.00)	98.14(99.75)	99.91(100.00)	98.97(99.93)
	DPO	96.38(99.96)	100.00(100.00)	97.37(99.94)	98.33(99.86)	93.78(99.88)	97.17(99.93)
	ORPO	97.07(99.98)	99.00(100.00)	96.90(99.93)	98.30(99.85)	99.92(100.00)	98.24(99.95)
	TLPO(ours)	97.43(99.97)	100.00(100.00)	98.09(99.95)	98.48(99.88)	97.10(99.95)	98.22(99.95)
ja	Baseline	91.67(99.95)	92.00(99.91)	94.98(99.89)	99.14(99.92)	97.23(99.84)	95.00(99.90)
	SFT	97.10(99.96)	100.00(100.00)	98.01(99.53)	95.55(99.55)	100.00(100.00)	98.13(99.81)
	DPO	94.58(99.84)	93.00(99.83)	95.22(99.85)	99.10(99.75)	97.73(99.97)	95.93(99.85)
	ORPO	97.45(99.99)	95.00(99.96)	97.14(99.92)	99.17(99.88)	99.83(100.00)	97.72(99.95)
	TLPO(ours)	97.46(99.99)	100.00(100.00)	98.57(99.99)	99.38(99.90)	99.33(99.99)	98.95(99.98)
avg.	Baseline	90.24(99.62)	95.00(99.95)	96.65(99.95)	98.70(99.77)	97.38(99.88)	95.59(99.83)
	SFT	96.69(99.89)	99.92(100.00)	99.32(99.88)	97.28(99.08)	99.95(100.00)	98.63(99.77)
	DPO	91.25(99.55)	96.12(99.94)	97.07(99.94)	98.71(99.75)	97.62(99.93)	96.16(99.82)
	ORPO	92.53(99.74)	97.58(99.95)	97.43(99.95)	98.67(99.66)	99.75(99.97)	97.19(99.86)
	TLPO(ours)	96.56(99.93)	99.54(100.00)	98.63(99.98)	99.06(99.90)	98.96(99.98)	98.55(99.96)

(a) Average Response Pass Rate(RPR) and Word Pass Rate(WPR). Values are presented as RPR(WPR) in %.

Lang.	Method	MIF (en)	MIF (target)	MMMLU (target)	GPQA (en)	GPQA-D (en)	ARC-C (en)	BBH (en)	MATH (en)	GSM8K (en)	GSM8K (cross)	Mean
zh	Baseline	68.39	49.35	55.50	31.70	30.81	75.09	55.86	21.38	77.25	81.72	54.71
	SFT	53.42	36.23	52.14	23.21	23.23	74.06	57.53	41.44	23.57	23.49	40.83
	DPO	69.87	50.83	56.03	29.24	29.29	75.09	55.98	20.90	76.76	82.55	54.65
	ORPO	68.76	50.65	54.64	27.90	28.28	76.02	55.83	20.18	77.17	77.75	53.72
	TLPO(ours)	70.43	51.76	55.77	28.79	29.80	74.40	56.55	15.34	78.74	81.89	54.35
ar	Baseline	67.65	60.26	53.79	31.70	30.81	75.09	55.86	21.38	77.25	77.58	55.14
	SFT	56.38	37.15	48.95	27.68	25.25	74.57	57.15	38.22	25.89	46.82	43.81
	DPO	69.32	61.18	53.43	28.57	29.80	75.09	55.80	21.28	77.09	76.92	54.85
	ORPO	69.32	58.96	52.83	25.67	24.75	75.77	56.29	21.02	77.09	78.25	53.99
	TLPO(ours)	68.76	58.60	53.30	27.23	28.79	74.66	55.41	18.06	76.18	77.83	53.88
ko	Baseline	68.58	50.46	53.45	31.70	30.30	75.09	55.86	20.82	77.25	72.54	53.61
	SFT	52.50	31.24	49.95	24.33	25.76	74.23	57.06	38.16	24.48	26.63	40.43
	DPO	69.50	49.72	53.52	29.69	29.29	75.09	55.89	20.68	77.17	71.88	53.24
	ORPO	68.21	47.13	53.60	27.68	31.31	75.34	55.35	13.16	77.75	57.24	50.68
	TLPO(ours)	70.79	49.54	53.44	25.89	28.79	74.57	56.12	19.58	77.09	72.29	52.81
ja	Baseline	68.21	50.09	54.66	31.70	30.81	75.09	55.86	21.38	77.25	73.86	53.89
	SFT	51.20	31.42	51.03	22.10	27.78	74.15	56.18	37.84	24.65	45.82	42.22
	DPO	71.53	49.91	54.68	26.79	26.77	75.09	55.95	20.62	77.50	73.95	53.28
	ORPO	69.50	47.32	54.13	26.79	29.80	75.68	56.63	15.66	77.34	75.68	52.85
	TLPO(ours)	68.76	48.80	54.82	27.46	26.26	74.91	55.15	17.78	77.34	76.59	52.79
avg.	Baseline	68.21	52.54	54.35	31.70	30.68	75.09	55.86	21.24	77.25	76.43	54.33
	SFT	53.37	34.01	50.52	24.33	25.51	74.25	56.98	38.91	24.65	35.69	41.82
	DPO	70.06	52.91	54.42	28.57	28.79	75.09	55.91	20.87	77.13	76.32	54.01
	ORPO	68.95	51.02	53.80	27.01	28.53	75.70	56.02	17.50	77.34	72.23	52.81
	TLPO(ours)	69.69	52.17	54.33	27.34	28.41	74.64	55.81	17.69	77.34	77.15	53.46

(b) Average accuracy after fine-tuning.

Table 9: Detailed RPR, WPR, and accuracy results for the Gemma3-4B-it model after fine-tuning, in a setting where English output is regarded as neutral.

Lang.	Method	LCB (cross-lingual)	LCB (monolingual)	MIF (target)	MMMLU (target)	GSM8K (cross)	Mean
zh	Baseline	96.15(99.98)	98.25(99.98)	99.22(99.99)	99.55(99.86)	99.78(99.96)	98.59(99.96)
	SFT	98.42(99.99)	100.00(100.00)	99.88(100.00)	99.63(99.66)	99.94(100.00)	99.57(99.93)
	DPO	96.70(99.78)	98.50(99.99)	99.22(99.99)	99.58(99.86)	99.85(99.93)	98.77(99.91)
	ORPO	95.57(99.95)	99.38(99.99)	99.04(99.99)	99.48(99.75)	99.88(99.98)	98.67(99.93)
	TLPO(ours)	98.24(99.99)	99.88(100.00)	99.70(100.00)	99.77(99.98)	100.00(100.00)	99.52(99.99)
ar	Baseline	87.84(99.61)	95.00(99.97)	96.60(99.96)	97.16(99.84)	97.08(99.92)	94.74(99.86)
	SFT	96.22(99.91)	99.58(100.00)	99.88(99.97)	98.84(99.54)	99.35(99.98)	98.78(99.88)
	DPO	93.13(98.89)	97.75(99.55)	98.99(99.43)	98.18(98.45)	99.01(99.48)	97.41(99.16)
	ORPO	90.11(99.75)	95.42(99.93)	98.04(99.98)	96.95(99.88)	97.39(99.90)	95.58(99.89)
	TLPO(ours)	96.09(99.93)	99.25(100.00)	99.64(100.00)	99.28(99.93)	99.55(99.99)	98.76(99.97)
ko	Baseline	93.66(99.91)	94.75(99.94)	96.05(99.96)	96.38(99.90)	95.81(99.83)	95.33(99.91)
	SFT	97.44(99.94)	100.00(100.00)	99.32(99.99)	99.19(99.91)	99.66(99.97)	99.12(99.96)
	DPO	97.34(99.98)	99.50(100.00)	98.98(99.94)	98.64(99.51)	97.89(99.95)	98.47(99.87)
	ORPO	93.37(99.50)	96.50(99.96)	96.30(99.74)	95.54(99.68)	97.40(99.73)	95.82(99.72)
	TLPO(ours)	97.62(99.96)	99.75(100.00)	99.10(99.98)	99.17(99.96)	98.92(99.96)	98.91(99.97)
ja	Baseline	96.98(99.98)	96.50(99.97)	98.21(99.96)	99.53(99.98)	99.20(99.96)	98.08(99.97)
	SFT	97.57(99.78)	100.00(100.00)	99.07(99.87)	98.79(99.88)	99.93(100.00)	99.07(99.91)
	DPO	98.07(99.96)	97.75(99.95)	98.50(99.95)	99.37(99.91)	99.32(99.97)	98.60(99.95)
	ORPO	98.35(99.99)	98.75(99.99)	98.74(99.94)	99.45(99.96)	99.83(99.99)	99.02(99.97)
	TLPO(ours)	98.78(100.00)	100.00(100.00)	99.40(99.98)	99.73(99.97)	99.83(100.00)	99.55(99.99)
avg.	Baseline	93.66(99.87)	96.12(99.97)	97.52(99.97)	98.15(99.89)	97.97(99.92)	96.68(99.92)
	SFT	97.41(99.90)	99.90(100.00)	99.54(99.96)	99.11(99.75)	99.72(99.99)	99.14(99.92)
	DPO	96.31(99.65)	98.37(99.87)	98.92(99.83)	98.94(99.43)	99.02(99.83)	98.31(99.72)
	ORPO	94.35(99.80)	97.51(99.97)	98.03(99.91)	97.85(99.82)	98.63(99.90)	97.27(99.88)
	TLPO(ours)	97.68(99.97)	99.72(100.00)	99.46(99.99)	99.49(99.96)	99.58(99.99)	99.19(99.98)

(a) Average Response Pass Rate(RPR) and Word Pass Rate(WPR). Values are presented as RPR(WPR) in %.

Lang.	Method	MIF (en)	MIF (target)	MMMLU (target)	GPQA (en)	GPQA-D (en)	ARC-C (en)	BBH (en)	MATH (en)	GSM8K (en)	GSM8K (cross)	Mean
zh	Baseline	69.69	53.14	58.70	33.93	32.58	82.57	50.01	49.43	78.56	84.84	59.34
	SFT	57.35	41.22	49.99	27.62	25.13	82.79	56.38	33.86	54.26	62.53	49.11
	DPO	70.29	52.91	58.15	32.59	32.32	82.59	49.77	49.14	77.81	84.88	59.05
	ORPO	68.90	50.42	55.97	30.97	30.93	83.15	50.46	49.47	74.59	84.10	57.90
	TLPO(ours)	69.36	54.21	58.00	30.97	34.09	82.34	52.00	47.74	79.98	84.76	59.35
ar	Baseline	69.45	52.54	52.06	33.65	33.08	82.53	50.12	49.57	78.47	79.90	58.14
	SFT	64.70	42.88	43.99	30.75	29.55	82.94	57.65	49.22	65.28	63.90	53.09
	DPO	70.10	51.80	50.51	32.14	33.33	82.62	50.22	49.10	78.23	79.76	57.78
	ORPO	68.44	50.60	50.01	31.58	33.46	83.13	49.21	49.72	76.43	79.22	57.18
	TLPO(ours)	69.18	50.14	49.86	31.92	34.22	82.51	50.31	47.82	79.67	78.41	57.40
ko	Baseline	69.73	47.92	54.04	33.76	32.20	82.55	49.94	49.27	78.64	80.11	57.82
	SFT	63.54	39.60	45.43	27.96	29.80	82.74	56.88	48.70	64.72	62.55	52.19
	DPO	70.43	47.37	53.00	31.25	32.83	82.53	49.50	48.78	78.53	78.53	57.27
	ORPO	68.99	44.04	49.84	31.70	32.32	82.83	48.99	47.18	76.41	75.54	55.78
	TLPO(ours)	69.69	48.43	53.52	32.03	34.47	82.55	50.38	48.96	78.85	79.09	57.80
ja	Baseline	69.78	48.29	55.50	33.31	32.32	82.55	50.20	49.56	78.25	81.24	58.10
	SFT	58.96	35.95	46.72	26.62	27.02	82.79	55.18	33.64	53.14	64.39	48.44
	DPO	58.41	34.66	49.39	27.40	27.65	82.64	47.36	32.14	65.67	71.11	49.64
	ORPO	57.67	30.22	50.51	27.85	27.53	82.87	48.24	32.55	65.16	73.51	49.61
	TLPO(ours)	68.67	46.07	55.55	31.42	32.70	82.68	50.91	48.31	79.71	81.68	57.77
avg.	Baseline	69.66	50.47	55.07	33.66	32.54	82.55	50.07	49.46	78.48	81.52	58.35
	SFT	61.14	39.91	46.54	28.24	27.87	82.81	56.52	41.35	59.35	63.34	50.71
	DPO	67.31	46.68	52.76	30.85	31.53	82.59	49.21	44.79	75.06	78.57	55.94
	ORPO	66.00	43.82	51.58	30.52	31.06	82.99	49.22	44.73	73.14	78.09	55.12
	TLPO(ours)	69.22	49.71	54.24	31.58	33.87	82.52	50.90	48.21	79.55	80.99	58.08

(b) Average accuracy after fine-tuning.

Table 10: Detailed RPR, WPR, and accuracy averaged across the 4 models after fine-tuning, in a setting where English output is regarded as neutral.