Gained in Translation: Privileged Pairwise Judges Enhance Multilingual Reasoning

Lintang Sutawika[1] **Gokul Swamy**[2] **Zhiwei Steven Wu**[3] **Graham Neubig**[1]
[1]Carnegie Mellon University, Language Technologies Institute
[2]Carnegie Mellon University, Robotics Institute
[3]Carnegie Mellon University, Software and Societal Systems Department
{lsutawik, gswamy, zstevenwu, gneubig}@cs.cmu.edu

Abstract

When asked a question in a language less seen in its training data, current reasoning large language models (RLMs) often exhibit dramatically lower performance than when asked the same question in English. In response, we introduce SP3F (Self-Play with Privileged Pairwise Feedback), a two-stage framework for enhancing multilingual reasoning without *any* data in the target language(s). First, we supervise fine-tune (SFT) on translated versions of English question-answer pairs to raise base model correctness. Second, we perform RL with feedback from a pairwise judge in a self-play fashion (Swamy et al., 2024), with the judge receiving the English reference response as *privileged information*. Thus, even when none of the model's responses are completely correct, the privileged pairwise judge can still tell which response is better. End-to-end, SP3F greatly improves base model performance, even outperforming fully post-trained models on multiple math and non-math tasks with less than $1/8$ of the training data across the single-language, multilingual, and generalization to unseen language settings.

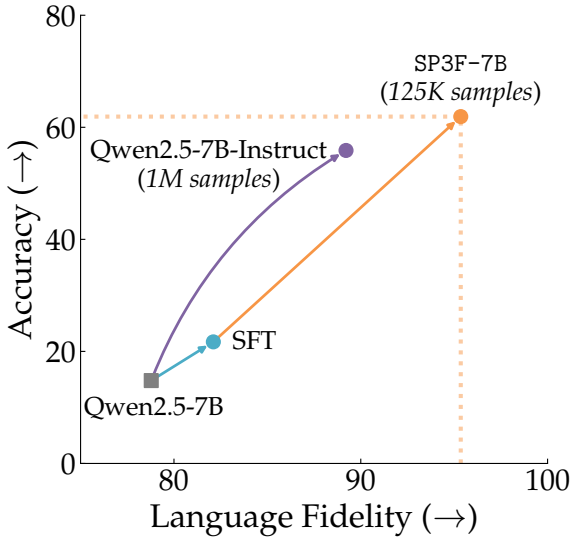

Figure 1: We propose SP3F: *Self-Play with Privileged Pairwise Feedback*: a method for training multilingual reasoning models without *any* data in the target language(s). SP3F-7B out-performs Qwen2.5-7B-Instruct across 4 tasks with roughly $1/8$ of the training data ($125,000$ for SP3F-7B vs. $1,000,000$ for Qwen2.5-7B-Instruct), both in terms of accuracy and language fidelity (did the model answer in the target language?).

1 Introduction

Current reasoning large language models (RLMs) are trained on data (e.g., chains of thought, CoTs) that is primarily in English (Ghosh et al., 2025). This means that when an RLM is asked the same question in a non-English language, it often exhibits dramatically lower performance than if it were asked the question in English (Yong et al., 2025; Muennighoff et al., 2023; Shi et al., 2022; Tam et al., 2025).

Improving reasoning performance in *lower resourced* languages (e.g., Indonesian, Swahili, Bengali) is challenging as we lack large amounts of data in the target language for supervised fine-tuning (SFT), and the base model's probability of generating the correct answer might be so low that getting positive signal for reinforcement learning (RL) to succeed is computationally challenging. Furthermore, for reasoning tasks, outcome-level *verifiable rewards* (Lambert et al., 2025) that consider just the final answer provide only indirect supervision on the CoT, making exploration challenging due to the sparsity of feedback (Kakade, 2003). Put together, we face a *cold start* problem we can't easily offline fine-tune our way out of.

In response, we propose SP3F (Self-Play with Privileged Pairwise Feedback): a two-stage framework for increasing reasoning performance in non-English target language(s) that doesn't require *any* data in the target language(s). First, we apply SFT on *translated* versions of English reference responses to raise our RLM's probability of gen-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8687–8705

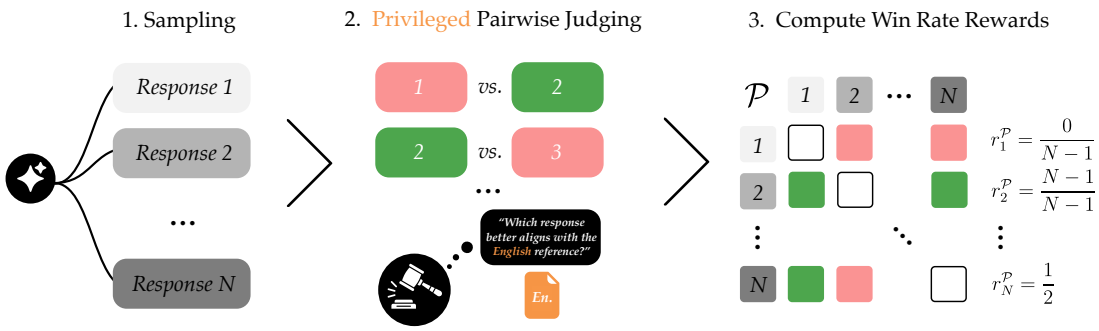

Figure 2: The second stage of the SP3F pipeline is to perform RL (GRPO, Shao et al. (2024)) with feedback from verifiable rewards (Lambert et al., 2025) and a pairwise judge. To aid in its judgments, the judge LLM is given access to *privileged information* in the form of an English reference response. Concretely, we sample N responses from the model (left), ask the privileged judge to pick a winner from each pair (center), and then use the average win-rate of each response against the other $N-1$ samples as the reward for RL (right, Swamy et al. (2024)).

erating correct answers. Second, we perform RL with a combination of verifiable rewards (e.g., answer correctness, language fidelity) and preference feedback from an LLM judge (Zheng et al., 2023). The LLM judge directly supervises the CoT of the RLM, giving it a direction of improvement even when it can't produce a correct final answer, which helps with the cold start issue.

While conceptually promising, the noisiness of the feedback provided by LLM judges makes incorporating them into RL training challenging. First, if the LLM judge itself is unfamiliar with a lower-resourced language, it may be unable to provide accurate feedback. In response, we provide the English reference response as *privileged information* (Vapnik and Vashist, 2009) to the judge, asking it to merely pick which of the two RLM responses more closely aligns with the English reference response, which is an easier *translation* task than judgment in the abstract. We find that the use of privileged information improves judgment quality.

Second, due to their pretraining on vast swathes of internet text, LLMs often exhibit *intransitive* (i.e., cyclic) preferences where they might rank $A \succ B$, $B \succ C$, and $C \succ A$ (Xu et al., 2025). Such intransitivity means no scalar reward function can faithfully represent the judge's preferences, making standard reward modeling fundamentally misspecified. Rather than fitting an inconsistent reward model, we adopt a *self-play* style approach that optimizes pairwise preferences directly: after sampling a batch of candidate responses, we use the judge to compare all pairs and assign each response a score equal to its empirical win rate. This aggregation converts pairwise judgments into a learning objective that reliably improves the model despite intransitivity (Swamy et al., 2024). Put together, we propose to use *privileged pairwise judges* to provide denser feedback to the RLM during RL.

Our key insight is that ***we can use English reference responses during both SFT and RL by framing both learning problems in terms of translation.*** In particular, we use reference responses as data for translation during SFT and as privileged information for the pairwise judge during downstream RL. More explicitly, our contribution is three-fold:

1. We introduce SP3F: a multi-step framework for increasing reasoning performance in a target language without data in said language. We find that RLMs trained via SP3F out-perform fully post-trained models on both in-domain math and out-of-domain non-math tasks in the target language.

2. We apply SP3F on data from 18 languages, producing a model that out-performs fully post-trained models using $\frac{1}{8}$ as much training data. We outperform Qwen2.5-7B-Instruct across math and on-math reasoning tasks. We find particularly large improvements on lower-resourced languages and see better generalization to unseen languages.

3. We perform an in-depth exploration of the benefits provided by privileged information. We find that privileged information is particularly helpful with reducing the intransitivity of the judge model, as well as in improving detection of correct reasoning chains, even if the final answer is incorrect. This helps mitigate cold-start issues.

Finally, we release our source-code[1] and artifacts[2].

[1] https://github.com/lintangsutawika/sp3f
[2] https://hf.co/collections/neulab/sp3f

Model	Overall		MGSM		MT Math100		Belebele		Global MMLU Lite	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	14.79	78.78	22.15	90.67	21.16	58.22	7.52	80.39	8.34	85.85
+ SFT	21.70	82.11	33.66	91.37	26.72	58.26	12.94	89.18	13.48	89.62
+ RLVR	57.79	**96.09**	65.34	**99.75**	44.50	**86.10**	**68.18**	98.73	53.15	**99.78**
SP3F-7B	**61.91**	95.35	**72.50**	99.38	56.84	82.93	67.54	**99.65**	50.76	99.45
Qwen2.5-7B-Instruct	55.87	89.21	66.36	98.38	52.12	65.66	56.79	96.59	48.20	96.21
+ Translate Test	57.01	85.98	66.15	95.81	**60.08**	59.34	48.09	92.27	**53.73**	96.49

Table 1: Across in-domain math tasks (MGSM and MT Math100) and out-of-domain tasks non-math tasks (Belebele and Global MMLU Lite) over a subset of 18 languages (Table 6) that were used to train SP3F-7B, we see SP3F-7B consistently outperforms the Qwen2.5-7B-Instruct. We measure performance in percentage by Accuracy (Acc) and Language Fidelity (Lang). Highest score presented in **bold** and second highest underlined. Notably, SP3F-7B required only $\frac{1}{8}$ as much data to post-train Qwen2.5-7B-Instruct. Full results in Appendix D.

2 SP3F: Self-Play with Privileged Pairwise Feedback for Multilingual Reasoning

In this section, we begin by describing SP3F in detail. SP3F is a two-step framework for improving reasoning performance in a target language without data in said language. SP3F only requires English reference responses, which can be relatively easily generated by a teacher model (e.g., o1 (Jaech et al., 2024), R1 (DeepSeek-AI et al., 2025)).

Below, we use $x \in \mathcal{X}$ to refer to prompts/questions and $y \in \mathcal{Y}$ to refer to responses, with $y^\star$ referring to an (English) reference response. We assume access to a dataset $\mathcal{D}$ of $(x, y^\star)$ pairs. Each response y consists of a chain-of-thought $z \in \mathcal{Z}$ and response $a \in \mathcal{A}$ (i.e., $y = (z, a) \in \mathcal{Y} = \mathcal{Z} \times \mathcal{A}$). We search over policies $\pi \in \Pi \subseteq \{\mathcal{X} \to \Delta(\mathcal{Y})\}$. We use $\circ$ to denote the concatenation of two strings and $\mathsf{tx}(\cdot)$ to denote translation into the appropriate target language. There are two stages of the SP3F pipeline: an SFT stage, followed by an RL stage.

Stage 1: SFT on Translated English Responses. Ideally, we would solve the cold-start problem of reasoning in a lower-resourced language by training on data in the target language. However, by definition, there is a relatively limited about of data available in a target language. Furthermore, it is often difficult to learn a strong policy given limited amounts of data to SFT (Swamy et al., 2025).

We propose a simple solution to this problem: SFT *translations* of relatively plentiful English reference responses $(x, y^\star)$. Explicitly, we maximize likelihood via a standard next-token prediction loss:

$$\pi_{\mathsf{sft}} = \arg\max_{\pi \in \Pi} \sum_{i=1}^{|\mathcal{D}|} \log\left(\pi(\mathsf{tx}(y_i^\star) | \mathsf{tx}(x_i))\right). \quad (1)$$

Performing this process raises our RLM's probability of generating the correct answer in the target language, aiding in downstream mode selection via online RL (Yue et al., 2025) in the next stage.

Stage 2: RL with Verifiable Rewards + Privileged Pairwise Judge Feedback. Next, we perform RL, with rewards given via a composition of four terms: three verifiable binary indicators, and one batch-level judge feedback term. Explicitly, given N responses $y_{1:N} \sim \pi(x)$, we compute:

$$r(x, y_i, y^\star) = r^{\mathsf{acc}}(x, y_i) + r^{\mathsf{fmt}}(y_i) + r^{\mathsf{lang}}(y_i) + r^{\mathcal{P}}(x, y_i, y_{1:N}, y^\star). \quad (2)$$

Verifiable Rewards. The first three terms, $r^{\mathsf{acc}}(x, y_i) \in \{0, 1\}$ (accuracy), $r^{\mathsf{fmt}}(y_i) \in \{0, 1\}$ (formatting), and $r^{\mathsf{lang}}(y_i) \in \{0, 1\}$ (language fidelity), are each verifiable rewards. In particular, $r^{\mathsf{acc}}(x, y_i)$ measures if the answer a_i is correct, $r^{\mathsf{fmt}}(y_i)$ measures if answer a_i was provided inside a `\boxed{}` template, and $r^{\mathsf{lang}}(y_i)$ measures whether the response was indeed in the target language. We use an automated language classifier to check what fraction of the response is in the target language. If the fraction $\geq 70\%$, we output a score of 1. We found that providing a binary indicator of the target language content, rather than a scalar in $[0, 1]$, helped avoid "reward-hacking" (Hadfield-Menell et al., 2017), where the model would learn to output a short response to achieve 100% language fidelity. We chose 70% as our threshold to account for the fact that math symbols are not counted as part of any particular language.

Judge Feedback Reward. Even after SFT, our RLM may still have a relatively low probability of generating a correct reasoning chain. As the verifiable rewards only focus on the correctness of

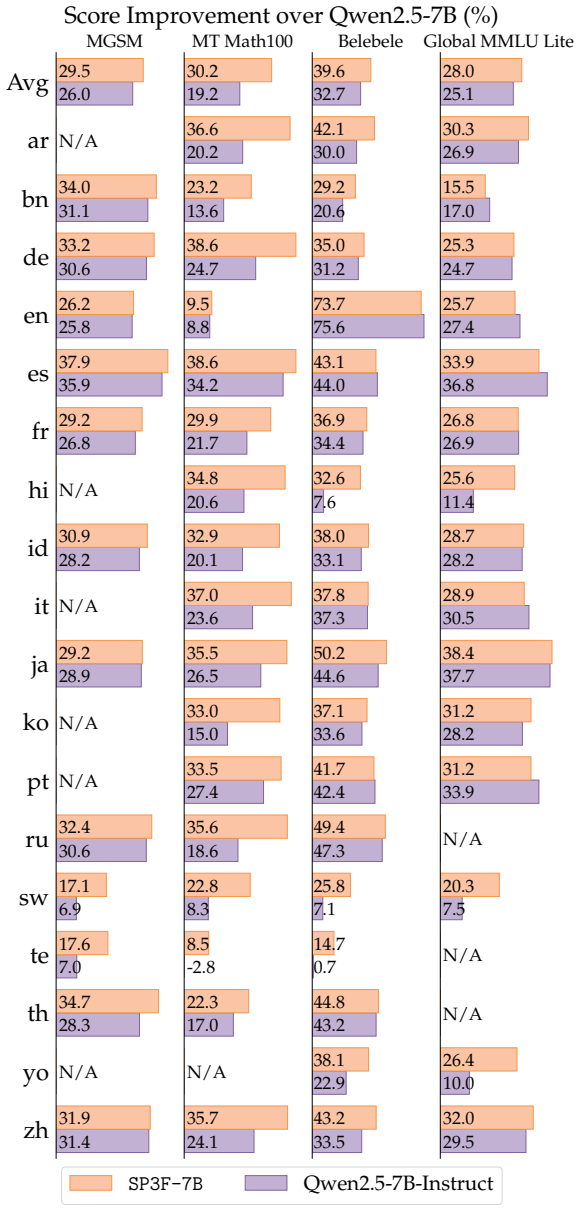

Figure 3: SP3F-7B generally outperforms Qwen2.5-7B-Instruct across most languages tested. We see particularly large gains for in-domain math tasks (left two columns) and on out-of-domain tasks in lower-resourced languages (e.g., Swahili). Each bar represents the gain in terms of absolute performance points compared to Qwen2.5-7B for a specific language. N/A denotes that the task is not available for that language.

the answer and the language fidelity of the CoT, they do not provide direct supervision on the the correctness of the CoT. In response, we propose using an LLM judge for supervision on the CoT z. This can provide a clear direction of improvement even when the RLM can't answer the question completely correctly. We use gpt-4o-mini as a judge

There are two key challenges with learning from LLM judge feedback for multilingual reasoning. The first is that the judge may struggle to evaluate responses in a lower resourced language that it doesn't understand well itself. In response, we propose to give the judge (but not the RLM) access to *privileged information* (Vapnik and Vashist, 2009) in the form of the English reference answer $y^\star$. Thus, rather than having to judge the model's response y in the abstract, the judge merely needs to assess how closely the solution y aligns with the English reference $y^\star$, which is an easier *translation*-style task. From another angle, we're *recycling* the data used during offline SFT during online RL, effectively squeezing more juice out of the same samples, akin to Jain et al. (2025).[3]

The second challenge is that due to their pretraining on a wide variety of text scraped from the internet, LLM judges often exhibit *intransitive* preferences (Xu et al., 2025), where they might rank responses $y_A, y_B, y_C \sim \pi(x)$ as $y_A \succ y_B$, $y_B \succ y_C$, and $y_C \succ y_A$. We find significant intransitivity in our LLM judge, as we explore below. Such inconsistent feedback can be a challenge to learn from. In response, given N samples in a batch, we use a *pairwise judge* to pick a winner from each of the $\binom{N}{2}$ pairs and use the win rate of each sample as the reward. Such a *self-play* approach is provably robust to intransitive preferences (Swamy et al., 2024). Put together, we optimize the following:

$$r^{\mathcal{P}}(x, y_i, y_{1:N}, y^\star) = \sum_{j \neq i}^{N} \frac{\widetilde{\mathcal{P}}(y_i \succ y_j | x, y^\star)}{N-1}, \quad (3)$$

where $\widetilde{\mathcal{P}}(y_i \succ y_j | x, y^\star) \in [0, 1]$, with a value of 1 denoting that y_i was preferred to y_j by the privileged pairwise judge. To account for the positional bias of pairwise judges (Zheng et al., 2023; Qin et al., 2024), we perform the standard averaging of judge preferences across both input orderings:

$$\widetilde{\mathcal{P}}(y_i \succ y_j | x, y^\star) = \frac{\mathcal{P}(y_i \succ y_j | x, y^\star) + (1 - \mathcal{P}(y_j \succ y_i | x, y^\star))}{2}. \quad (4)$$

RL Algorithm. We use the DR.GRPO (Liu et al., 2025b) variant of GRPO (Shao et al., 2024). While industry standard practice for RLM training, these

[3] In greater detail, when given privileged information, the judge is able to provide more accurate feedback without requiring parameter updates/training data. Thus, compared to non-privileged judges, we have reduced the *sample complexity* of learning a *verifier*. This directly translates to a reduction in the end-to-end sample complexity of learning a policy/*generator* via the arguments presented in Swamy et al. (2025).

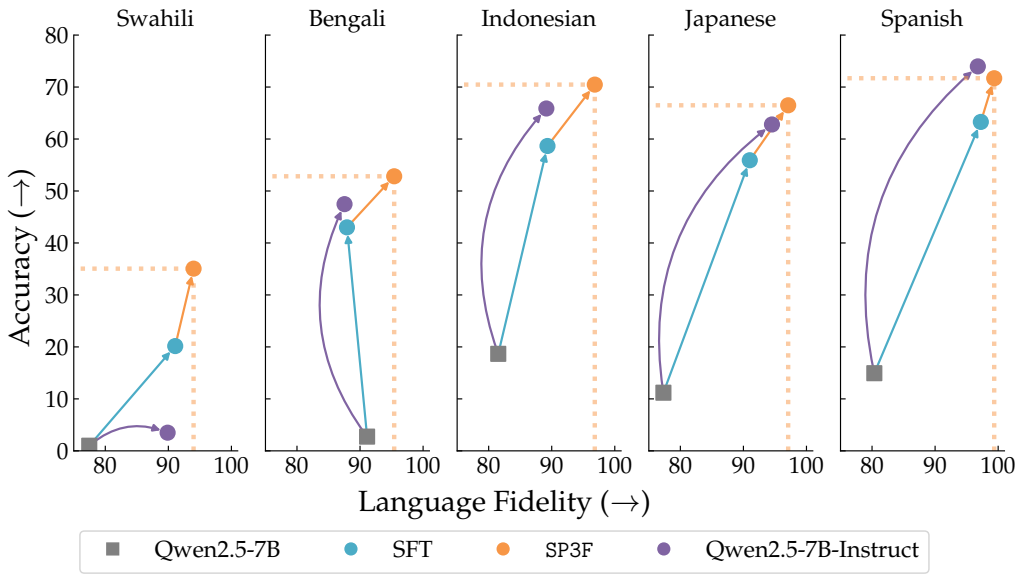

Figure 4: Across 5 target languages (ordered in terms of increasing resourced-ness), single-language SP3F training produces models that out-performs Qwen2.5-7B-Instruct. We find particularly large deltas on lower resourced languages like Indonesian, Bengali, and Swahili. Results are averaged across all four tasks considered.

RL algorithms are particularly prone to mode selection (Shao et al., 2025; Oertell et al., 2025), underscoring the need for a preliminary SFT step.

Multilingual Training. Finally, to take advantage of the repeatedly observed benefits of multilingual training (Yong et al., 2025; Shi et al., 2022; Muennighoff et al., 2023), we apply the above pipeline with data from 18 different languages (see Table 6 for full list). We refer to the model that results from this multilingual training process as SPF3-7B.

3 Experiment Setup

We now outline our experimental setup.

Dataset Construction. Our training data is generated from DeepScaleR (Luo et al., 2025), which contains math reasoning problems from the AIME, AMC, and other competitions. We translate both the query and response to 18 languages (full list in Table 6) using GPT-5-Nano. All translated versions are then merged into a single dataset of equal proportions. Each sample includes the original English response, which is provided as privileged information for the LLM judge during RL training.

Training. We train SP3F-7B on top of the Qwen2.5-7B (Qwen et al., 2025) base model. Our experiments were implemented in Verl (Sheng et al., 2024), with a slight modification to allow for pairwise judges during reward calculation. For the SFT stage, we perform 1000 gradient steps with a batch size of 16 and learning rate of 1×10^{-5}. For the secondary RL stage, we use the aforementioned combination of verifiable rewards and judge feedback for supervision. We train for 500 gradient steps using a batch size of 32 prompts, $N = 8$ responses per prompt, and a learning rate of 5×10^{-7}.

Metrics. We report two metrics: *Accuracy* (the correctness of the model's final \boxed{} answer) and *Language Fidelity* (whether the model's response is at least 70% in the target language). To measure language fidelity, we use lingua.[4] These metrics are precisely the r^{acc} and r^{lang} discussed in Sec. 2.

Evaluation. We evaluate all models on 2 math tasks and 2 non-math tasks. In the context of our math reasoning-based training data, the math tasks are in-domain, while the non-math tasks are out-of-domain tasks meant to provide an estimate of how well the model generalizes. For math tasks, we use MGSM (Shi et al., 2022) to test basic word math problems and MT-Math100 (Son et al., 2025) that is a translated subset of MATH500 (Lightman et al., 2023).[5] For non-math tasks, we use Global MMLU Lite (Singh et al., 2024) to evaluate world knowledge and Belebele (Bandarkar et al., 2024) for reading comprehension. Each score is a per-question average over 8 model responses.

Baselines. To compare against a strong post-training baseline, we choose Qwen2.5-7B-

[4] https://github.com/pemistahl/lingua-py

[5] To increase coverage, we additionally contribute a new, translated and manually verified (by an author) Indonesian version of MGSM.

Input Response-Answer Pair	$\mathcal{P}_{\text{priv}}$	$\mathcal{P}_{\text{no-priv}}$
✓CoT ∘ ✓Ans vs. ✗CoT ∘ ✗Ans	**85.77**	76.42
✓CoT ∘ ✓Ans vs. ✗CoT ∘ ✓Ans	77.16	**81.08**
✓CoT ∘ ✗Ans vs. ✗CoT ∘ ✗Ans	**59.90**	46.53

Table 2: Using sampled correct responses (✓CoT ∘ ✓Ans) and incorrect responses (✗CoT ∘ ✗Ans), we observe how well privileged information can help identify between CoT that were sourced from correct responses (✓CoT) and incorrect responses (✗CoT) with their final answers swapped. We evaluate the accuracy of the judge preferring the left-hand option when privileged information ($\mathcal{P}_{\text{priv}}$) and without priviledge information ($\mathcal{P}_{\text{no-priv}}$). **Row 1**: Privileged information increases judge accuracy on responses from Qwen2.5-7B+SFT. **Row 3**: Access to privileged information increases the judge's accuracy of identifying correct CoT even when the final answer is wrong. This can be important early in RL.

	Belebele				MT Math100			
	Q2.5-7B-I		SP3F-7B		Q2.5-7B-I		SP3F-7B	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Avg	39.9	93.4	**58.3**	**99.9**	48.3	62.7	**51.7**	**83.6**
af	19.2	98.8	**72.5**	**100.0**	55.9	62.9	**59.1**	**83.7**
gu	**30.6**	99.5	19.2	**100.0**	35.5	65.0	**38.0**	**68.9**
he	61.3	72.5	**64.1**	**99.2**	52.7	36.4	**56.4**	**75.5**
nl	67.2	99.1	**77.9**	**100.0**	58.5	58.8	**60.1**	**87.0**
pa	8.4	97.8	**23.0**	**100.0**	32.1	75.5	**35.1**	**80.3**
tl	16.6	81.3	**60.5**	**100.0**	44.4	56.7	**51.8**	**87.1**
tr	49.7	98.3	**69.1**	**100.0**	50.6	70.7	**53.3**	**92.0**
vi	66.5	99.8	**79.8**	**100.0**	56.9	75.4	**59.8**	**94.4**

Table 3: SP3F-7B consistently outperforms Qwen2.5-7B-Instruct even on languages that it was not explicitly trained on. We show languages that were not included in the training set that exist in both tasks.

Instruct (Qwen et al., 2025), a model post-trained by the Qwen team on 1M post-training examples. In addition, we compare against Translate-Test (Ponti et al., 2021; Artetxe et al., 2023), where the query is translated into English before being solved by the model. Specifically, we use Self-Translate Test (Etxaniz et al., 2023) where the translation are done using the model itself. This technique is a training-free inference time procedure to boost model reasoning performance.

Using Privileged Information. We use GPT-4o-mini as our LLM judge and provide it with the query in the target response language. Our system and user prompts instruct the model to deliberate over the responses A and B and decide which among them have closest sense to the included English reference response (full prompt available in Table 11). The judge is then directed to provide its final answer or either `\boxed{A}` or `\boxed{B}`.

4 SP3F Unlocks Data-Efficient Multilingual Reasoning

We begin by discussing the single-language gains of the SP3F pipeline, before discussing our multilingual results and more carefully exploring the benefits of privileged information for LLM judges.

4.1 SP3F Improves Lower-Resourced Language Reasoning

As seen in Figure 4, applying the SP3F pipeline consistently boosts model performance above Qwen2.5-7B-Instruct level, averaged across both in-domain math tasks and out-of-domain non-math tasks. We see particularly large gains on the left side of the figure in lower-resourced languages (e.g., Swahili, where the SP3F-trained model achieves more than three times the accuracy of Qwen2.5-7B-Instruct). Furthermore, we see that both stages of the SP3F pipeline are critical for strong final model performance. We emphasize that single-language SP3F uses significantly less data than the entire Qwen2.5-7B-Instruct post-training pipeline and no data in the target language.

4.2 SP3F Improves Multilingual Reasoning

We now explore the performance of SP3F-7B, which is trained by applying the SP3F pipeline on multilingual training data from 18 languages.

Aggregate Results. As seen in Table 1, SP3F-7B consistently out-performs Qwen2.5-7B-Instruct on both in-domain math tasks and out-of-domain not math tasks. This is impressive given SP3F-7B only required $\frac{1}{8}$ as much post-training data. Furthermore, even when we apply the inference-time translate test technique to Qwen2.5-7B-Instruct, SP3F-7B still usually beats the improved model.

By comparing the +RLVR and SP3F-7B rows of Table 1, we can more precisely identify the benefits of judge feedback. We see that judge feedback improves accuracy at the cost of slightly worse language fidelity. Zooming in further, we see particularly strong gains on in-domain math tasks, with a slight decrease in performance relative to RLVR-trained models on out-of-domain non-math tasks. This suggests that exploring regularization techniques (e.g., those proposed by Song et al. (2024)) may be a promising avenue for improving the out-of-domain generalization of judge-trained models.

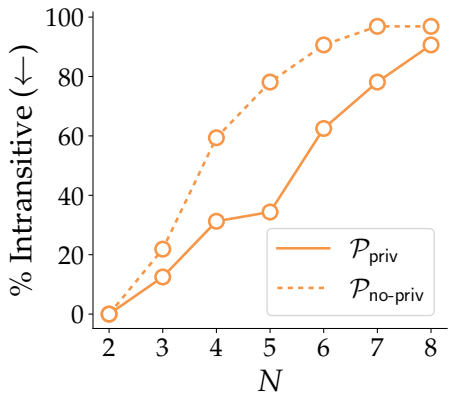

Figure 5: As N increases, it becomes increasingly likely for both privileged and non-privileged judges to have intransitive preferences. However, we consistently find that the privileged judge $\mathcal{P}_{\text{priv}}$ has more transitive preferences than the non-privileged judge $\mathcal{P}_{\text{no-priv}}$. We report the PNT metric proposed by Xu et al. (2025). We use $N = 8$ as the number of training rollouts per sample.

Judge Type	Avg	Math		Non-Math	
		MGSM	MT Math100	Belebele	Global MMLU Lite
$\mathcal{P}_{\text{priv}}$	**64.6**	**75.3**	**56.8**	**72.7**	53.7
$\mathcal{P}_{\text{no-priv}}$	63.5	74.2	53.6	71.5	**54.7**

Table 4: We see training models with feedback from privileged judges ($\mathcal{P}_{\text{priv}}$) performance over those trained with feedback from non-privileged judges ($\mathcal{P}_{\text{no-priv}}$). We see particularly strong gains on in-domain math tasks.

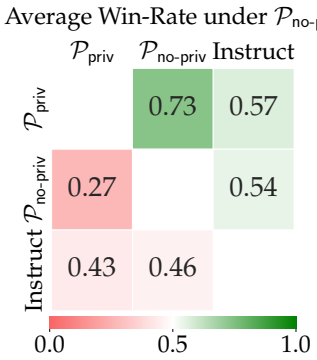

Figure 6: We see $\mathcal{P}_{\text{no-priv}}$ prefer models trained by $\mathcal{P}_{\text{priv}}$ to models trained under its own feedback. Each cell represents the win-rate of the row model against the column model, as evaluated by $\mathcal{P}_{\text{no-priv}}$.

Per-Language Results. In Figure 3, we display the gains over the base model broken down by target language. Echoing the results in Table 1, we see consistent positive deltas over Qwen2.5-7B-Instruct on all languages on in-domain tasks. We also see particularly large positive deltas on out-of-domain tasks in lower-resourced languages like Swahili, Hindi, Yoruba, and Telegu.

Unseen Language Generalization Results. When we evaluate `SP3F-7B` on eight languages outside of its training set (but not necessarily that of Qwen2.5-7B-Instruct), we see better performance across tasks compared to Qwen2.5-7B-Instruct. With the exception of Gujarati (gu), `SP3F-7B` out-performs Qwen2.5-7B-Instruct by around 18% on Belebele and 3.4% on MT Math100 in terms of accuracy. This potentially indicates that `SP3F-7B` is a more generally capable multilingual reasoning model than Qwen2.5-7B-Instruct and that our training pipeline doesn't somehow preclude generalization.

4.3 Privileged Information Aids LLM Judges

We now perform an in-depth exploration of the multiple benefits of privileged information for LLM judge performance. For conciseness, we use $\mathcal{P}_{\text{priv}}$ and $\mathcal{P}_{\text{no-priv}}$ to refer to judges with and without access to English reference responses, respectively.

Privileged Information Helps With Cold Starts. We introduced the pairwise judge reward $r^{\mathcal{P}}$ (Eq. 3) to provide supervision on the RLM's CoT, especially early on in training when the model may struggle to generate correct final answers. To understand the effect of privileged information on achieving this goal, we first sample responses from the Qwen2.5-7B+SFT model and group them by the correctness of the final answer into correct responses (✓CoT ∘ ✓Ans) and incorrect responses (✗CoT ∘ ✗Ans). We then graft together different CoTs and answers across correctness groups. Beyond increasing the base accuracy of the judge (Table 2, Row 1), we also observe a significant 13% increase in the ability of the judge model to detect the correctness of the CoT even when the answer is incorrect (Table 2, Row 3). Thus, privileged information appears to help mitigate the early-training cold start issue by increasing the efficacy of CoT supervision. This is particularly important in lower-resourced languages where the accuracy of the SFT model is likely to be relatively low.

Privileged Information Reduces Intransitivity. While our self-play approach is robust to intransitivity, more consistent preferences can still simplify our learning problem. In Figure 5, we see a clear reduction in intransitivity when we provide the pairwise judge with privileged information. Thus, beyond merely increasing the accuracy of the judge, privileged information enhances the ability of the judge to provide more consistent *global rankings* (i.e., total orderings) over responses.

Privileged Information Helps Most In-Domain. In Table 4, we see models trained with feedback from privileged judges outperform those trained with feedback from non-privileged judges, potentially as a result of the reduction in intransitivity and increase in correct CoT detection capabilities privileged information seems to provide. We see consistent gains on in-domain math tasks and mixed results on out-of-domain non-math tasks.

Models Trained by Judges With Privileged Information Are Preferred Even By Non-Privileged Judges. As a final evaluation, we stack the deck in the favor of models trained my $\mathcal{P}_{\text{no-priv}}$ and use $\mathcal{P}_{\text{no-priv}}$ as an oracle to perform model-to-model comparisons. In Figure 6, we see that $\mathcal{P}_{\text{no-priv}}$ prefers models trained via $\mathcal{P}_{\text{priv}}$ to models it trained *itself*, overcoming a *self-preference* bias (Panickssery et al., 2024). Thus, these results potentially indicate that the models trained by $\mathcal{P}_{\text{priv}}$ are qualitatively better models, rather than just narrowly optimizing the preferences of $\mathcal{P}_{\text{priv}}$.

5 Related Work

We provide a brief overview of some related work.

Multilingual Language Modeling. Data scarcity is a core problem in multilingual language modeling (Joshi et al., 2020). A variety of approaches have been proposed to deal with this concern. First, various authors scaled up dataset sizes via Internet scraping (Xue et al., 2021; BigScience Workshop et al., 2023), crowd-sourced high quality data (Cahyawijaya et al., 2023), manually translated English training data (Lai and Nissim, 2024; Ng et al., 2025), or proposed using synthetic training data (Kautsar et al., 2025). Beyond quantity, quality of data also matters: Zheng et al. (2025) propose using rubrics to curate datasets. In parallel, training-free approaches like few-shot prompting (Cahyawijaya et al., 2024) and representation editing (Zhao et al., 2025) have been demonstrated to elicit multilingual reasoning capability.

A plethora of techniques have been explored for training multilingual language models. For example, Huang et al. (2024) merged smaller specifically-trained components into off-the-shelf language models, Zhu et al. (2024) trained models to do question translation as a way to improve multilingual performance, and Barua et al. (2025) trained models using machine translated or distilled responses from teacher models. Contemporary work has shown how RL training exclusively in high resource languages improves performance on other languages (Huang et al., 2025), and explored optimizing language fidelity rewards via RL (Hwang et al., 2025). In contrast, our work trains on questions in lower-resourced languages and goes beyond sparse verifiable rewards.

Reinforcement Learning Post-Training. Popularized by RLHF (Ouyang et al., 2022), reinforcement learning techniques have gained wide adoption in LLM post-training. A wide spectrum of policy optimization algorithms have been proposed, from off-policy regression-based losses (Rafailov et al., 2023; Gao et al., 2024a; Azar et al., 2024; Gao et al., 2024b), to policy gradient techniques (Ahmadian et al., 2024; Shao et al., 2024; Liu et al., 2025b). We opt for GRPO-style policy gradients in our work due to their relative simplicity and on-policy nature, which fits in cleanly to the self-play algorithmic template (Swamy et al., 2024).

Supervision via Privileged Information. Privileged information has provable benefits for reducing the complexity of learning (Vapnik and Vashist, 2009). For example, privileged information has been the key ingredient in recent successes in robotics (Choudhury et al., 2017; Chen et al., 2019; Kumar et al., 2021; Swamy et al., 2022; Song et al., 2025). However, because these approaches focus on direct imitation, it is not immediately obvious how to directly apply them in the multilingual reasoning context where we lack sufficiently large amounts of data to imitate in the target language. Thus, we instead provide the privileged information to an LLM judge, similar to the work of Ye et al. (2024); Kim et al. (2024); Zhou et al. (2025). In particular, we propose providing relatively plentiful English reference answers as privileged information to aid LLM judges in their evaluation of CoTs, especially in lower-resourced languages.

6 Conclusion

We introduce `SP3F`: a data-efficient two-stage framework for improving multilingual reasoning performance without data in target language(s). We find that `SP3F`-trained models out-perform fully post-trained models across the single language, multilingual, and unseen languages setting while only requiring $\frac{1}{8}$ as much post-training data. We ablate the use of *privileged information* to improve the quality of LLM judgments and find it provides multiple benefits. Thus, an interesting direction for future work is to explore other uses of privileged

pairwise judges beyond multilingual reasoning.

Acknowledgements

LS and GN were supported in part by a grant from Apple and a compute grant from the CMU FLAME center. GKS and ZSW were supported in part by a STTR grant. We thank Sean Welleck for providing references to other uses of privileged LLM judges and Drew Bagnell for stimulating conversations about the sample complexity benefits of privileged verifiers.

Limitations

First, while we did not explicitly use any human-generated data in target languages during training, it is possible that the judge model saw some during its own training process. A more careful effort to understand the composition of judge training data would strengthen our argument. Second, while we tested all models on multiple math and non-math reasoning tasks, a wider set of evaluations and training sets would allow us to make stronger claims. Third, providing a formal understanding of the sample complexity benefits of privileged information would dovetail with our empirical results. Finally, we did not conduct human studies related to the stylization of the model responses nor has it be tested for any user application. Thus, further evaluations should be conducted before it is integrated into user-facing systems to avoid unwanted harms.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

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

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

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. 2025. Long chain-of-thought reasoning across languages. *Preprint*, arXiv:2508.14828.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, and 374 others. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, and 29 others. 2023. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. In *North American Chapter of the Association for Computational Linguistics*.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2019. Learning by cheating. *Preprint*, arXiv:1912.12294.

Sanjiban Choudhury, Mohak Bhardwaj, Sankalp Arora, Ashish Kapoor, Gireeja Ranade, Sebastian Scherer, and Debadeepta Dey. 2017. Data-driven planning via imitation learning. *Preprint*, arXiv:1711.06391.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english? *Preprint*, arXiv:2308.01223.

Zhaolin Gao, Jonathan Chang, Wenhao Zhan, Owen Oertell, Gokul Swamy, Kianté Brantley, Thorsten Joachims, Drew Bagnell, Jason D Lee, and Wen Sun. 2024a. Rebel: Reinforcement learning via regressing relative rewards. *Advances in Neural Information Processing Systems*, 37:52354–52400.

Zhaolin Gao, Wenhao Zhan, Jonathan D Chang, Gokul Swamy, Kianté Brantley, Jason D Lee, and Wen Sun. 2024b. Regressing the relative future: Efficient policy optimization for multi-turn rlhf. *arXiv preprint arXiv:2410.04612*.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. A survey of multilingual reasoning in language models. *Preprint*, arXiv:2502.09457.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. 2017. Inverse reward design. *Advances in neural information processing systems*, 30.

Shulin Huang, Yiran Ding, Junshu Pan, and Yue Zhang. 2025. Beyond english-centric training: How reinforcement learning improves cross-lingual reasoning in llms. *Preprint*, arXiv:2509.23657.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. Mindmerger: Efficient boosting llm reasoning in non-english languages. *ArXiv*, abs/2405.17386.

Jaedong Hwang, Kumar Tanmay, Seok-Jin Lee, Ayush Agrawal, Hamid Palangi, Kumar Ayush, I. Fiete, and Paul Pu Liang. 2025. Learn globally, speak locally: Bridging the gaps in multilingual reasoning. *ArXiv*, abs/2507.05418.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Arnav Kumar Jain, Vibhakar Mohta, Subin Kim, Atiksh Bhardwaj, Juntao Ren, Yunhai Feng, Sanjiban Choudhury, and Gokul Swamy. 2025. A smooth sea never made a skilled sailor: Robust imitation via learning to search. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Sham Machandranath Kakade. 2003. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom).

Muhammad Dehan Al Kautsar, Aswin Candra, Muhammad Alif Al Hakim, Maxalmina Satria Kahfi, Fajri Koto, Alham Fikri Aji, Peerat Limkonchotiwat, Ekapol Chuangsuwanich, and Genta Indra Winata. 2025. Seadialogues: A multilingual culturally grounded multi-turn dialogue dataset on southeast asian languages. *Preprint*, arXiv:2508.07069.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. *Preprint*, arXiv:2310.08491.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. 2021. Rma: Rapid motor adaptation for legged robots. *Preprint*, arXiv:2107.04034.

Huiyuan Lai and Malvina Nissim. 2024. mCoT: Multilingual instruction tuning for reasoning consistency in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12012–12026, Bangkok, Thailand. Association for Computational Linguistics.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

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

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. 2025a. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *Preprint*, arXiv:2505.24864.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025b. Understanding r1-zero-like training: A critical perspective. *Preprint*, arXiv:2503.20783.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Erran Li, Raluca Ada Popa, and Ion Stoica. 2025. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. Notion Blog.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulagadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Wayne Lau, Choon Meng Tan, and 12 others. 2025. Sea-lion: Southeast asian languages in one network. *Preprint*, arXiv:2504.05747.

O Oertell, W Zhan, G Swamy, ZS Wu, K Brantley, J Lee, and W Sun. 2025. Heuristics considered harmful: Rl with random rewards should not make llms reason. *Notion Blog*.

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. *Preprint*, arXiv:2203.02155.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. 2021. Modelling latent translations for cross-lingual transfer. *Preprint*, arXiv:2107.11353.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, and 1 others. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. 2025. Spurious rewards: Rethinking training signals in rlvr. *Preprint*, arXiv:2506.10947.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

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

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Guijin Son, Jiwoo Hong, Hyunwoo Ko, and James Thorne. 2025. Linguistic generalizability of test-time scaling in mathematical reasoning. *arXiv preprint arXiv:2502.17407*.

Yuda Song, Dhruv Rohatgi, Aarti Singh, and J. Andrew Bagnell. 2025. To distill or decide? understanding the algorithmic trade-off in partially observable reinforcement learning. *Preprint*, arXiv:2510.03207.

Yuda Song, Gokul Swamy, Aarti Singh, J. Andrew Bagnell, and Wen Sun. 2024. The importance of online data: Understanding preference fine-tuning via coverage. In *Advances in Neural Information Processing Systems*, volume 37, pages 12243–12270. Curran Associates, Inc.

Gokul Swamy, Sanjiban Choudhury, J Bagnell, and Steven Z Wu. 2022. Sequence model imitation learning with unobserved contexts. *Advances in Neural Information Processing Systems*, 35:17665–17676.

Gokul Swamy, Sanjiban Choudhury, Wen Sun, Zhiwei Steven Wu, and J. Andrew Bagnell. 2025. All roads lead to likelihood: The value of reinforcement learning in fine-tuning. *Preprint*, arXiv:2503.01067.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. 2024. A minimaximalist approach to reinforcement learning from human feedback. *Preprint*, arXiv:2401.04056.

Zhi Rui Tam, Cheng-Kuang Wu, Yu Ying Chiu, Chieh-Yen Lin, Yun-Nung Chen, and Hung yi Lee. 2025. Language matters: How do multilingual input and reasoning paths affect large reasoning models? *Preprint*, arXiv:2505.17407.

Vladimir Vapnik and Akshay Vashist. 2009. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. 2025. Investigating non-transitivity in llm-as-a-judge. *Preprint*, arXiv:2502.14074.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. Flask: Fine-grained language model evaluation based on alignment skill sets. *Preprint*, arXiv:2307.10928.

Zheng-Xin Yong, M. Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff,

Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H. Bach, and Alham Fikri Aji. 2025. Crosslingual reasoning through test-time scaling. *Preprint*, arXiv:2505.05408.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Preprint*, arXiv:2504.13837.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025. When less language is more: Language-reasoning disentanglement makes llms better multilingual reasoners. *Preprint*, arXiv:2505.15257.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Weihua Zheng, Xin Huang, Zhengyuan Liu, Tarun Kumar Vangani, Bowei Zou, Xiyan Tao, Yuhao Wu, Ai Ti Aw, Nancy F. Chen, and Roy Ka-Wei Lee. 2025. Adamcot: Rethinking cross-lingual factual reasoning through adaptive multilingual chain-of-thought. *Preprint*, arXiv:2501.16154.

Jin Peng Zhou, Séb Arnold, Nan Ding, Kilian Q Weinberger, Nan Hua, and Fei Sha. 2025. Graders should cheat: privileged information enables expert-level automated evaluations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16583–16601.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

A Contribution Statements

- **LS** initiated the project, performed all experiments, and wrote the first draft of the paper.
- **GS** came up with the core algorithmic idea of privileged judges, wrote most of the final paper, and helped advise the project.
- **ZSW** and **GN** advised the project, provided computational resources, and helped with writing.

B Training Hyperparameters

Hyperparameter	Value
Supervised Finetuning	
Batch Size	16
LR	1e-5
Optimizer	AdamW
Training Iterations (Multilingual)	1000
Training Iterations (Single Language)	250
Reinforcement Learning	
Batch Size	32
LR	5e-7
Rollouts (N)	8
Sampling Temperature	1.0
Max Response Length	2048
ε_{low}	0.2
$\varepsilon_{\text{high}}$	0.28
Training Iterations (Multilingual)	500
Training Iterations (Single Language)	250

Table 5: We increase the SFT iterations for our multilingual to account for multiple languages. For RL, we use slightly higher $\varepsilon_{\text{high}}$ following Liu et al. (2025a) to encourage more exploration.

C Evaluated Language

Train Languages		**Unseen Languages**	
Code	Language	Code	Language
ar	Arabic	af	Afrikaans
bn	Bengali	nl	Dutch
de	German	gu	Gujarati
en	English	pa	Punjabi
es	Spanish	tr	Turkish
fr	French	tl	Tagalog
hi	Hindi	he	Hebrew
id	Indonesian	vi	Vietnamese
it	Italian		
ja	Japanese		
ko	Korean		
pt	Portuguese		
ru	Russian		
sw	Swahili		
te	Telugu		
th	Thai		
yo	Yoruba		
zh	Chinese		

Table 6: Train and Unseen Languages

D Full Table Results

In this section, we present per-language evaluation scores for the tasks that we present in the Table 1. All languages presented here were included in the training dataset. Note that for Translate Test, English is left unevaluated. For the aggregate score, Translate Test is averaged without English.

	MGSM									
	Avg		bn		de		en		es	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	22.1	90.7	2.6	97.7	27.6	88.2	40.6	100.0	28.2	83.5
+ SFT	33.7	91.4	4.7	99.2	35.4	92.7	72.5	99.9	41.6	86.9
+ RLVR	65.3	99.8	53.5	100.0	75.1	99.4	91.0	100.0	83.0	99.8
SP3F-7B	72.5	99.4	68.2	100.0	82.7	99.7	93.1	100.0	87.5	100.0
Qwen2.5-7B-Instruct	66.4	98.4	63.0	99.4	80.3	96.8	92.2	100.0	84.2	99.3
+ Translate Test	66.2	95.8	42.9	97.2	80.2	97.5	N/A	N/A	88.0	100.0

	fr		id		ja		ru		sw	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	45.5	79.8	31.4	91.3	21.8	92.8	26.5	94.3	0.8	85.2
+ SFT	51.5	81.2	45.6	86.8	27.2	96.2	43.0	90.8	0.6	88.2
+ RLVR	78.8	99.7	79.2	99.8	64.0	100.0	81.2	100.0	12.2	100.0
SP3F-7B	83.6	100.0	85.0	99.5	73.1	100.0	85.8	99.9	20.5	99.7
Qwen2.5-7B-Instruct	80.2	98.8	82.3	96.8	72.5	99.8	82.5	99.5	0.8	99.0
+ Translate Test	78.0	99.8	81.8	95.1	75.2	97.2	79.7	96.9	18.0	97.8

	te		th		zh	
	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	0.1	98.8	19.7	78.0	20.8	98.5
+ SFT	0.6	99.1	33.1	75.9	48.0	99.5
+ RLVR	14.3	100.0	70.3	98.5	81.3	100.0
SP3F-7B	34.1	100.0	72.6	94.5	83.8	99.3
Qwen2.5-7B-Instruct	13.9	99.0	62.2	92.0	82.2	100.0
+ Translate Test	29.7	99.5	73.6	73.1	80.8	100.0

Table 7: Evaluation scores per language for Global MGSM

	MT Math100									
	Avg		ar		bn		de		en	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	21.2	58.2	21.6	35.1	4.3	72.1	20.7	56.2	48.7	100.0
+ SFT	26.7	58.3	25.5	36.2	8.3	67.4	27.3	61.5	56.1	99.8
+ RLVR	44.5	86.1	41.8	78.3	29.9	88.9	52.4	86.9	61.6	100.0
SP3F-7B	56.8	82.9	57.6	72.2	48.1	74.6	62.0	92.0	67.8	100.0
Qwen2.5-7B-Instruct	52.1	65.7	45.8	51.3	47.5	56.2	61.2	65.0	66.4	100.0
+ Translate Test	60.1	59.3	58.2	46.7	57.3	48.0	65.7	63.4	N/A	N/A
	es		fr		hi		id		it	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	25.8	57.8	33.7	64.5	9.6	51.8	23.6	52.8	28.3	48.4
+ SFT	32.1	61.2	38.0	62.2	9.8	56.6	31.6	46.8	34.5	48.9
+ RLVR	51.5	98.1	53.9	92.4	36.6	84.7	49.4	79.8	51.1	84.7
SP3F-7B	61.9	98.9	62.5	95.5	52.8	78.3	61.1	81.1	61.9	88.8
Qwen2.5-7B-Instruct	63.3	88.6	61.2	80.3	43.8	58.8	54.7	62.0	60.0	63.9
+ Translate Test	71.6	80.0	59.6	77.7	62.9	58.2	64.9	59.5	63.6	68.4
	ja		ko		pt		ru		sw	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	15.7	65.7	25.1	44.4	34.2	57.8	28.2	41.4	2.0	56.9
+ SFT	21.6	57.5	26.3	44.4	41.9	61.6	33.8	35.9	2.5	63.0
+ RLVR	42.9	92.2	43.8	75.9	58.1	91.0	54.3	74.5	11.9	96.0
SP3F-7B	57.2	95.1	59.5	76.1	62.4	96.6	63.0	77.8	27.1	77.4
Qwen2.5-7B-Instruct	53.8	80.4	55.7	43.8	63.9	83.0	59.9	46.8	7.7	67.9
+ Translate Test	67.5	73.1	56.4	40.3	67.7	81.8	63.6	45.5	32.5	68.7
	te		th		zh					
	Acc	Lang	Acc	Lang	Acc	Lang				
Qwen2.5-7B	1.0	86.1	17.3	29.2	20.5	63.9				
+ SFT	0.8	85.5	23.6	31.1	33.6	64.5				
+ RLVR	12.6	97.2	37.9	46.1	55.6	91.6				
SP3F-7B	35.9	68.2	56.3	34.7	63.1	92.7				
Qwen2.5-7B-Instruct	25.1	56.3	52.9	27.5	57.6	75.0				
+ Translate Test	51.5	59.9	58.2	19.1	60.23	64.39				

Table 8: Evaluation scores per language for MT Math100

	Belebele									
	Avg		ar		bn		de		en	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	7.5	80.4	11.9	73.7	1.6	97.0	18.4	89.7	3.7	35.6
+ SFT	12.9	89.2	12.1	86.4	1.6	99.1	20.4	94.4	27.9	99.8
+ RLVR	68.2	98.7	74.6	99.8	57.8	100.0	77.8	99.9	86.8	100.0
SP3F-7B	67.5	99.7	69.8	99.9	57.0	100.0	78.2	100.0	86.7	100.0
Qwen2.5-7B-Instruct	56.8	96.6	66.2	79.4	42.1	97.6	70.9	99.6	90.4	100.0
+ Translate Test	48.1	92.3	21.6	92.1	44.9	97.9	58.6	99.7	N/A	N/A

	es		fr		hi		id		it	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	3.0	91.2	15.4	93.6	0.5	90.0	11.2	91.0	11.8	94.1
+ SFT	7.9	94.7	22.8	93.9	0.6	94.7	23.9	91.8	15.2	92.3
+ RLVR	80.4	100.0	81.7	99.9	55.6	99.9	76.6	99.7	78.9	99.8
SP3F-7B	80.4	100.0	82.7	100.0	55.7	100.0	78.3	100.0	81.5	100.0
Qwen2.5-7B-Instruct	82.5	99.8	77.8	99.9	5.9	99.8	69.0	99.5	80.8	99.8
+ Translate Test	60.2	99.4	61.1	99.5	46.3	98.5	59.4	98.8	62.5	99.2

	ja		ko		pt		ru		sw	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	2.2	73.9	23.9	79.0	11.8	87.7	6.2	72.2	0.5	82.6
+ SFT	4.2	78.3	25.3	84.8	23.4	91.0	16.7	86.6	1.4	88.8
+ RLVR	74.2	99.9	76.5	99.7	79.7	100.0	80.7	100.0	35.6	99.5
SP3F-7B	76.6	100.0	77.1	100.0	83.0	100.0	77.4	100.0	34.8	99.9
Qwen2.5-7B-Instruct	66.1	99.2	74.9	95.2	84.5	99.9	74.3	98.8	2.4	95.0
+ Translate Test	61.3	96.8	61.2	90.2	59.8	99.8	37.7	94.7	32.9	99.3

	te		th		yo		zh	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	0.3	99.3	9.5	60.9	0.5	47.5	3.0	87.9
+ SFT	0.5	98.6	15.2	77.1	1.0	61.5	12.9	91.5
+ RLVR	35.4	100.0	69.6	80.0	24.5	99.1	81.0	100.0
SP3F-7B	29.0	100.0	63.1	96.9	24.6	99.6	80.0	97.5
Qwen2.5-7B-Instruct	1.4	99.7	74.8	82.0	0.3	93.5	57.9	100.0
+ Translate Test	7.3	28.3	60.6	75.5	20.8	98.8	61.3	100.0

Table 9: Evaluation scores per language for Belebele

	Global MMLU Lite									
	Avg		ar		bn		de		en	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	8.3	85.8	8.9	72.8	2.5	97.7	17.2	91.0	18.1	100.0
+ SFT	13.5	89.6	10.6	84.1	2.2	99.4	17.8	95.5	40.3	100.0
+ RLVR	53.1	99.8	52.0	99.8	38.7	99.9	62.0	99.8	70.5	100.0
SP3F-7B	50.8	99.5	42.9	99.4	31.8	99.3	59.1	99.8	69.5	100.0
Qwen2.5-7B-Instruct	48.2	96.2	53.5	81.9	37.2	96.9	59.4	98.2	72.9	100.0
+ Translate Test	53.7	96.5	25.2	88.9	44.8	94.8	64.2	98.4	N/A	N/A
	es		fr		hi		id		it	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	2.8	88.9	17.2	90.7	1.0	82.2	8.4	91.0	11.6	92.5
+ SFT	7.8	92.9	21.3	90.9	1.1	92.7	17.9	87.9	13.8	92.9
+ RLVR	61.6	99.9	64.4	99.9	39.9	99.9	58.8	99.3	63.2	99.8
SP3F-7B	59.6	100.0	61.4	100.0	34.6	99.8	56.8	99.9	61.9	99.8
Qwen2.5-7B-Instruct	66.0	99.3	62.1	99.6	7.7	98.3	57.4	98.3	66.3	98.7
+ Translate Test	68.7	98.3	66.1	99.7	45.4	97.2	67.3	96.3	69.5	98.1
	ja		ko		pt		sw		yo	
	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang	Acc	Lang
Qwen2.5-7B	5.1	76.8	16.2	78.2	10.5	88.6	0.6	85.1	1.1	65.1
+ SFT	8.4	83.2	18.8	83.8	23.5	92.0	1.6	83.8	1.9	74.1
+ RLVR	57.0	99.8	54.7	99.3	62.1	99.9	25.7	99.7	24.3	99.7
SP3F-7B	58.8	100.0	57.5	99.3	61.6	100.0	26.5	99.8	19.4	99.5
Qwen2.5-7B-Instruct	58.7	98.7	59.3	91.6	67.3	99.6	3.1	97.6	1.0	85.2
+ Translate Test	65.9	96.5	53.1	87.4	64.9	99.5	31.5	97.9	19.8	98.4
	zh									
	Acc	Lang								
Qwen2.5-7B	4.2	87.1								
+ SFT	15.3	91.1								
+ RLVR	62.3	100.0								
SP3F-7B	60.0	95.2								
Qwen2.5-7B-Instruct	50.9	99.3								
+ Translate Test	65.8	99.4								

Table 10: Evaluation scores per language for Global MMLU

E Prompts

E.1 Pairwise Judge Prompts

System Message
You are an expert judge in evaluating the quality of responses to user queries. Your task is to determine which response (A or B) is preferable. You will be provided with the user query and the correct solution. The responses may be in various languages, but the solution will always be in English. Decide based on how well does each response align with the correct solution. The best response should have the closest meaning and intent to the correct solution. Write your analysis and end it by answering with either \boxed{A} or \boxed{B}.
User Message
<Query> ... </Query> <Correct Solution> ... </Correct Solution> <Response A> ... </Response A> <Response B> ... </Response B> First, using the solution as reference, decide which of the two responses is the closest to the solution. Finally, choose which is better by answering with either \boxed{A} or \boxed{B}. You MUST provide your reasoning before the answer.

Table 11: System and User Prompts for Privileged Pairwise Judge ($\mathcal{P}_{\text{priv}}$)

System Message
You are an expert judge in evaluating the quality of responses to user queries. Your task is to determine which response (A or B) is preferable. You will be provided with the user query and the correct solution. The responses may be in various languages. Write your analysis and end it by answering with either \boxed{A} or \boxed{B}.
User Message
<Query> ... </Query> <Response A> ... </Response A> <Response B> ... </Response B> First, decide which of the two responses is preferable. Finally, choose which is better by answering with either \boxed{A} or \boxed{B}. You MUST provide your reasoning before the answer.

Table 12: System and User Prompts for Non-Privileged Pairwise Judge ($\mathcal{P}_{\text{no-priv}}$)

E.2 System Messages

Code	Language	System Prompt (system_message)
af	Afrikaans	Dink stap vir stap na en plaas jou finale antwoord binne \boxed{}.
ar	Arabic	فكر خطوة بخطوة وضع إجابتك النهائية داخل \boxed{}.
bn	Bengali	ধাপে ধাপে যুক্তি দিন এবং আপনার চূড়ান্ত উত্তর \boxed{} এর মধ্যে লিখুন।
de	German	Begründen Sie dies Schritt für Schritt und geben Sie Ihre endgültige Antwort in \boxed{} ein.
en	English	Reason step by step and put your final answer within \boxed{}.
es	Spanish	Razona paso a paso y coloca tu respuesta final dentro de \boxed{}.
fr	French	Raisonner étape par étape et mettre votre réponse finale dans \boxed{}.
gu	Gujarati	પગલું દ્વારા પગલું કારણ આપો અને તમારા અંતિમ જવાબ \boxed{} માં મૂકો.
hi	Hindi	कदम दर कदम सोचें और अपना अंतिम उत्तर \boxed{} के भीतर लिखें।
id	Indonesian	Berpikir langkah demi langkah dan tuliskan jawaban akhir di dalam \boxed{}.
it	Italian	Ragiona passo dopo passo e scrivi la tua risposta finale all'interno di \boxed{}.
ja	Japanese	段階的に推論し、最終的な答えを \boxed{}内に記入してください。
jv	Javanese	Pikirake langkah demi langkah lan lebokake jawaban pungkasan sampeyan ing \boxed{}.
ko	Korean	단계별로 추론하고 최종 답을 \boxed{} 안에 넣으세요.
nl	Dutch	Denk stap voor stap na en plaats je uiteindelijke antwoord binnen \boxed{}.
pa	Punjabi	ਕਦਮ ਦਰ ਕਦਮ ਸੋਚੋ ਅਤੇ ਆਪਣਾ ਅੰਤਿਮ ਜਵਾਬ \boxed{} ਦੇ ਅੰਦਰ ਲਿਖੋ।
pt	Portuguese	Raciocine passo a passo e coloque sua resposta final dentro de \boxed{}.
ru	Russian	Рассудите шаг за шагом и поместите окончательный ответ в \boxed{}.
sw	Swahili	Sababu hatua kwa hatua na uweke jibu lako la mwisho ndani ya \boxed{}.
te	Telugu	దశలవారీగా తర్కించండి మరియు మీ తుది సమాధానాన్ని \boxed{} లో ఉంచండి.
th	Thai	อธิบายเหตุผลทีละขั้นตอนและใส่คำตอบสุดท้ายของคุณไว้ใน \boxed{}.
tl	Tagalog	Mag-isip nang hakbang-hakbang at ilagay ang iyong panghuling sagot sa loob ng \boxed{}.
tr	Turkish	Adım adım düşünün ve son cevabınızı \boxed{} içine yazın.
vi	Vietnamese	Hãy suy nghĩ từng bước một và đặt câu trả lời cuối cùng của bạn vào trong \boxed{}.
yo	Yoruba	Ṣe àlàyé ìdí rẹ̀ ní ìpele kọọkan kí o sì fi ìdáhùn ikẹhin rẹ sínú \boxed{}.
zh	Chinese	逐步推理并将您的最 终答案放在 \boxed{} 内。

Figure 7: System prompts for all the evaluated languages. Each prompt directs to think step-by-step and write a final answer inside \boxed{}.