APL-OPSD: INSTANCE-ADAPTIVE PRIVILEGED LANGUAGE SELECTION FOR MULTILINGUAL ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models acquire knowledge from multilingual corpora, but this knowledge is not equally accessible through every language. The most informative language can vary across questions, domains, and cultural contexts, making fixed English supervision suboptimal. We propose **Adaptive Privileged Language On Policy Self Distillation (APL-OPSD)**, a framework that selects the language of privileged context separately for each training instance. Given parallel question and answer pairs in multiple languages, APL-OPSD performs a one time offline probe that measures how each candidate context changes the likelihood of the target language reference answer and selects the most informative non target language. During training, the student observes only the target language question and generates an on policy response, while a frozen teacher additionally observes the selected crosslingual context and provides dense token level supervision over the same trajectory. Experiments across multiple languages, model scales, and tasks show that APL-OPSD achieves the strongest aggregate performance among the compared methods. Further ablations suggest that informative language selection contributes beyond merely adding privileged context. APL-OPSD jointly trains a single multilingual model and requires no additional routing, translation, or reference information at inference time.

1 INTRODUCTION

Large language models (LLMs) acquire knowledge from corpora spanning diverse languages and communities. Yet the knowledge encoded in their parameters is not equally accessible across languages: semantically equivalent queries can elicit substantially different reasoning behaviors and factual responses when expressed in different languages (Shi et al., 2022; Ahuja et al., 2023; Huang et al., 2024; Wang et al., 2025; Liu et al., 2025; Ai et al., 2025; Hwang et al., 2025; Li et al., 2025). As illustrated in Figure 1(a), the same culturally grounded question about the meaning of a thumbs-up gesture in Greece elicits different answers across languages: the model responds incorrectly in Chinese and English, but correctly in Greek. This contrast shows that, even when the underlying semantics and cultural context are unchanged, the language of the query can substantially affect whether the model successfully accesses the relevant knowledge.

Crucially, this does not imply that non-English languages are generally more informative than English. English often provides a strong linguistic pathway, particularly for reasoning-intensive tasks (Kang et al., 2026; Zhao et al., 2026b; Hwang et al., 2025; Zhang et al., 2026; Son et al., 2026), yet its advantage is not universal. Which language provides the most useful access to relevant knowledge may depend on the individual question, domain, and cultural context, with non-English languages sometimes eliciting complementary knowledge that is less readily accessed through English or the query language alone (Guo et al., 2025; 2026; Liu et al., 2025; Ai et al., 2025; Zheng et al., 2025). These observations suggest that multilingual competence is not a single capability uniformly shared across languages, but instead involves language-conditioned pathways for accessing the model's knowledge. Accordingly, the key problem is not merely to strengthen the model's capabilities in each language independently, but to make effective use of knowledge that becomes accessible through different linguistic contexts. A central challenge, therefore, is to identify the

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

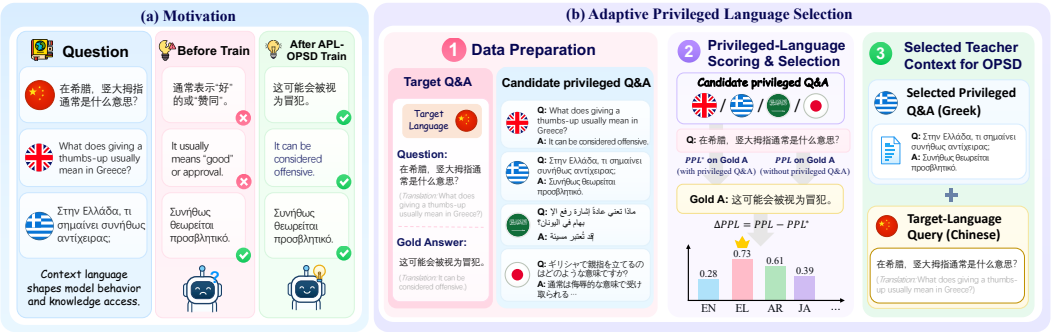

Figure 1: **Motivation and overview of APL-OPSD.** (a) Semantically equivalent queries in different languages can trigger different knowledge access and response correctness. (b) APL-OPSD scores candidate privileged languages for each instance, selects the most helpful one, and uses its Q&A as the teacher's privileged context for OPSD.

most informative linguistic pathway for each instance and transfer its advantage to the desired target language.

On-policy distillation (OPD) (Agarwal et al., 2024) has emerged as an effective post-training paradigm that combines the complementary strengths of supervised fine-tuning and reinforcement learning: it trains on trajectories sampled from the student's current policy, reducing the train–inference distribution mismatch of offline supervision, while retaining dense token-level supervision instead of relying solely on sparse sequence-level rewards. On-policy self-distillation (OPSD) (Zhao et al., 2026a) further removes the dependence on an external teacher by deriving both student and teacher from the same language model under different conditioning contexts. The student observes only the inference-time question, whereas the privileged teacher additionally observes a reference solution and evaluates the same student-generated trajectory. Its application to multilingual learning, however, remains limited. Recent extensions transfer reasoning across languages by providing the privileged teacher with an English translation of the question and an English reference solution, with some methods further emphasizing reasoning-critical token positions (Liu et al., 2026; Wang et al., 2026a). However, these approaches uniformly use English as the privileged language for every instance. While English provides strong privileged supervision for many instances, particularly on reasoning-intensive tasks, using it globally assumes that the same linguistic context remains optimal across questions and domains. This leaves privileged-language selection itself largely unexplored.

To this end, we propose **Adaptive Privileged-Language On-Policy Self-Distillation (APL-OPSD)**, which generalizes a globally fixed privileged language into an instance-specific language choice. We introduce adaptive privileged language selection (APLS), the core mechanism of APL-OPSD, to choose the privileged language separately for each training instance. As illustrated in Figure 1(b), we first construct a set of candidate privileged contexts from parallel question–answer pairs in different languages. For each candidate language, we use the frozen initial model to measure how much providing its parallel question–answer pair as privileged context reduces the perplexity of the target-language gold answer, relative to conditioning on the target-language query alone. The language with the largest improvement is selected as the privileged language for that instance. The corresponding question–answer pair is subsequently provided only to the privileged teacher, while the student continues to observe and answer the original target-language question. In this way, APL-OPSD adapts the teacher's linguistic context to each instance without changing the student's target language or requiring any language-selection component at inference time. Our contributions are summarized as follows:

❑ (1) We revisit multilingual on-policy self-distillation and identify an overlooked design choice in existing approaches: the teacher's privileged information is provided in a globally fixed language, typically English. We formulate privileged-language selection as an instance-specific decision, leading to adaptive privileged-language on-policy self-distillation.

❑ (2) We propose adaptive privileged language selection, an offline model-based mechanism that evaluates, for each training instance, which candidate language provides the most useful privileged context for predicting the target-language answer.

❑ (3) We integrate adaptive privileged-language selection into OPSD and conduct extensive experiments across multiple model scales. Our evaluation spans culturally grounded knowledge, mathematical reasoning, and broad-coverage multilingual understanding, demonstrating strong overall performance against a broad range of post-training baselines.

2 Related Work

2.1 Multilingual Capabilities and Crosslingual Knowledge Transfer

Although modern large language models are pretrained on corpora covering many languages, their capabilities remain unevenly distributed across them. Studies of multilingual reasoning and general language understanding reveal substantial performance differences when semantically equivalent questions are expressed in different languages (Shi et al., 2022; Ahuja et al., 2023; Kang et al., 2026; Zhao et al., 2026b). These differences are partly associated with imbalances in pretraining and post training data, but resource availability alone does not fully determine which language performs best. In particular, culturally grounded facts, social norms, and regional knowledge may be more readily elicited through languages associated with the relevant communities, while English may remain more effective for other categories of knowledge (Liu et al., 2025; Wang et al., 2025). Moreover, stronger alignment between representations of parallel inputs does not necessarily yield commensurate multilingual performance gains (Ravisankar et al., 2026). Multilingual performance therefore depends not only on whether a model encodes the required knowledge, but also on which linguistic context is used to access and express it.

Prior work has explored several ways to transfer and combine capabilities across languages. Multilingual instruction tuning constructs translated reasoning data or uses a pivot language to improve consistency across target languages (Lai & Nissim, 2024; Zhang et al., 2024). Crosslingual preference optimization aligns responses in other languages with those produced in a dominant language (She et al., 2024; Dang et al., 2024; Yang et al., 2025b; Zhao et al., 2025), while multilingual knowledge aggregation combines information elicited through different language contexts (Huang et al., 2024). More recently, language routing has treated the response language as a selectable variable and adapted language preferences during online optimization (Guo et al., 2026). These studies demonstrate that explicit use of crosslingual information can improve multilingual performance. However, they primarily operate on translated training sequences, complete generated responses, or response level rewards. In contrast, our work studies language selection within privileged teacher context and transfers the selected language advantage through dense token level supervision on responses generated by the current student policy.

2.2 On Policy Self Distillation

On policy distillation trains a student model on trajectories sampled from its current policy while using a teacher model to provide dense token level supervision over the same trajectories (Agarwal et al., 2024). Compared with supervised fine tuning on fixed reference responses, this paradigm reduces the mismatch between training and inference by exposing the student to prefixes that it actually generates. Compared with reinforcement learning based on outcome rewards, it provides a learning signal at every generated position rather than assigning the same scalar reward to an entire response. Conventional on policy distillation, however, typically relies on a separate and often stronger teacher model.

On policy self distillation removes this requirement by initializing the student and teacher policies from the same language model and differentiating them through their conditioning contexts. The student observes only the problem, while the teacher additionally receives a reference solution as privileged information and evaluates the response generated by the student (Zhao et al., 2026a). Recent crosslingual extensions provide the teacher with an English version of the problem and an English reference solution to transfer reasoning capabilities into other languages (Liu et al., 2026). Subsequent work further distinguishes reasoning decisions from surface realization and concentrates privileged supervision on positions that are more sensitive to the reference solution (Wang et al., 2026a). These approaches improve where token level supervision is applied, but retain English as the privileged language for every instance and typically train a separate model for each target language. Our work instead adapts which language supplies the privileged context for each instance and jointly

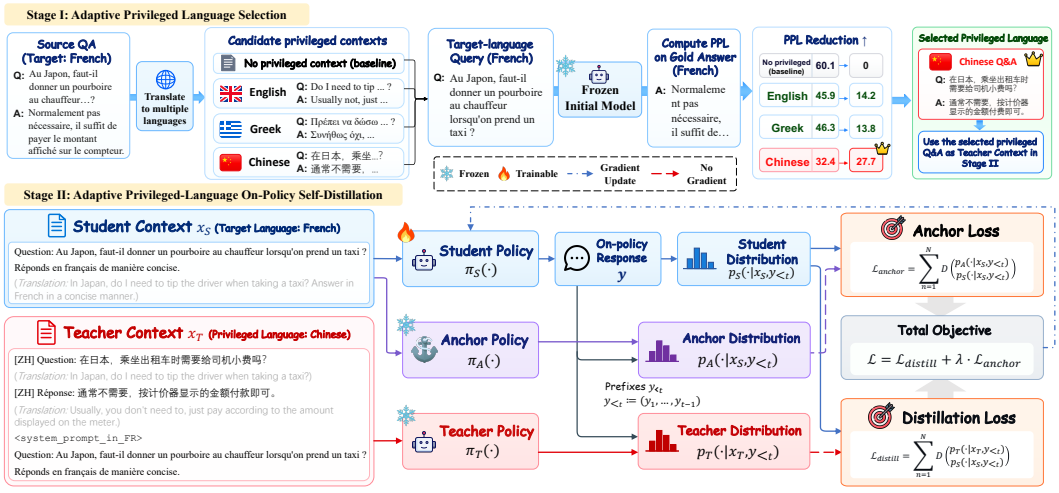

Figure 2: **Overview for Our APL-OPSD.** Unlike fixed English supervision, APL-OPSD selects an informative privileged language for each question, providing more culturally appropriate supervision in the target language.

trains a single model across all target languages. The resulting language selection mechanism is complementary to methods that determine which token positions should receive stronger distillation.

3 Methodology

3.1 Overview

We consider a multilingual question answering setting over a set of languages $\mathcal{L}$. For each instance, we construct parallel question–answer pairs across multiple languages and designate one language as the target language, in which the model is expected to answer. Our goal is to exploit complementary knowledge accessible through other languages during training while preserving target-language generation at inference time.

As illustrated in Figure 2, APL-OPSD consists of two stages. In Stage I, Adaptive Privileged Language Selection, we use the frozen initial model to evaluate non-target candidate languages and select the language that best improves prediction of the target-language reference answer relative to the target-query-only baseline. In Stage II, on-policy self-distillation, the student generates responses from the target-language query, while a frozen copy of the initial model acts as the teacher, additionally observing the selected privileged-language Q&A and providing token-level supervision along the same student-generated trajectory. The student is optimized with token-level distillation and initial-policy anchoring. At inference time, only the student model is retained, requiring neither privileged information nor language selection.

3.2 Adaptive Privileged Language Selection

Motivation. Existing cross-lingual OPSD methods Wang et al. (2026a); Liu et al. (2026) typically use English as a fixed privileged language for all training instances. However, the utility of privileged information need not be determined solely by the overall resource level of its language; it may also depend on how effectively a particular linguistic view exposes the knowledge required by each instance. Semantically equivalent question–answer pairs expressed in different languages can condition the same model differently, making some languages more useful than others for predicting a given target-language answer. This motivates treating privileged-language utility as instance-specific rather than globally fixed. Accordingly, as illustrated in Figure 2, we propose adaptive privileged language selection (APLS), which performs a one-time offline probe with the frozen initial model to identify, for each training instance, the privileged language that most improves the predictability of the target-language answer.

Multilingual Data Construction. As illustrated in the first part of Stage I in Figure 2, we construct the multilingual training data used for adaptive privileged-language selection. Following LRPO (Guo et al., 2026), we draw training instances from publicly available multilingual question-answering and instruction-following datasets (Guo et al., 2025; Wang et al., 2026b) and construct semantically aligned translations across 15 languages. We shuffle the resulting instances and assign target languages in a round-robin manner to maintain an approximately balanced distribution across languages. For each instance i, the assigned target-language pair $(x_i^{t_i}, y_i^{t_i})$ serves as the student training example, while its aligned variants in the remaining non-target languages form the candidate privileged contexts for subsequent language selection.

Privileged Language Scoring and Selection. As illustrated in Figure 2, we select the privileged language separately for each training instance from the constructed multilingual parallel data. For an instance i with target language t_i, we use the frozen initial model π_0 to evaluate the utility of each non-target language as privileged information. We first compute the perplexity of the target-language gold answer $y_i^{t_i}$ conditioned only on the target-language query $x_i^{t_i}$, which serves as the no-privilege baseline. For each candidate language $\ell \neq t_i$, we then prepend its aligned question–answer pair (x_i^{ℓ}, y_i^{ℓ}) as privileged context while keeping the prediction target unchanged. Specifically, we compute

$$\begin{aligned} \mathrm{PPL}_i^0 &= \mathrm{PPL}_{\pi_0}\left(y_i^{t_i} \mid x_i^{t_i}\right), & \mathrm{PPL}_i^{\ell} &= \mathrm{PPL}_{\pi_0}\left(y_i^{t_i} \mid x_i^{\ell}, y_i^{\ell}, x_i^{t_i}\right), \\ \Delta\mathrm{PPL}_i(\ell) &= \mathrm{PPL}_i^0 - \mathrm{PPL}_i^{\ell}, & \ell_i^* &= \arg\max_{\ell \in \mathcal{L}\setminus\{t_i\}} \Delta\mathrm{PPL}_i(\ell). \end{aligned} \tag{1}$$

A larger $\Delta\mathrm{PPL}_i(\ell)$ indicates that the privileged Q&A in language ℓ makes the target-language gold answer more predictable under the initial model. We use $(x_i^{\ell_i^*}, y_i^{\ell_i^*})$ as the privileged Q&A for instance i in the subsequent OPSD stage. This selection is performed once offline with the frozen initial model.

We exclude the target language to avoid the most direct form of same-language copying and restrict the study to cross-lingual privileged transfer. The selected privileged language is then cached and kept fixed throughout self-distillation; no language routing is performed within the training loop.

3.3 Adaptive Privileged-Language On-Policy Self-Distillation

After selecting an instance-specific privileged language, we use its question–answer pair to construct the privileged teacher context and perform on-policy self-distillation. As illustrated in Figure 2, for each training instance, we construct two conditioning contexts. The *student context* x_S contains only the target-language query and its response instruction, whereas the *teacher context* x_T additionally prepends the question–answer pair in the selected privileged language. Importantly, the target-language query and response instruction at the end of x_T are identical to those in x_S. Thus, the teacher receives additional cross-lingual knowledge while the student retains exactly the input interface used at inference time.

At each training step, the current student policy π_S first samples an on-policy response from the student context. We define the resulting trajectory and the student and teacher next-token distributions along this shared trajectory as

$$\begin{aligned} y = (y_1, \ldots, y_T) &\sim \pi_S(\cdot \mid x_S), \\ p_S^t(\cdot) &:= \pi_S(\cdot \mid x_S, y_{<t}), \qquad y_{<t} = (y_1, \ldots, y_{t-1}). \\ p_T^t(\cdot) &:= \pi_T(\cdot \mid x_T, y_{<t}), \end{aligned}$$

Here, π_T denotes the frozen privileged teacher policy, and $y_{<t}$ is the shared student-generated prefix at position t. Rather than generating a separate teacher response, the teacher evaluates the same trajectory sampled by the student, providing supervision directly on the states visited by the current student policy.

We additionally retain a frozen copy of the initial policy, denoted π_A, as an anchor policy. Using the same student context and response prefix, we define its next-token distribution as

$$p_A^t(\cdot) := \pi_A(\cdot \mid x_S, y_{<t}).$$

Only the student policy is updated during training, whereas both the privileged teacher and the anchor policy remain frozen. The resulting student, teacher, and anchor distributions provide the

signals used to train the student; we describe the corresponding training objectives in the following subsection.

3.4 Training Objective

Given an on-policy response $y = (y_1, \ldots, y_T)$ sampled from the current student policy, we optimize the student with two complementary objectives: a privileged distillation loss that transfers the teacher's token-level predictions under the selected privileged context, and an initial-policy anchoring loss that regularizes the student against excessive deviation from the initial policy.

Privileged Distillation. At each response position t, we obtain the full-vocabulary next-token distributions of the student and privileged teacher conditioned on the same prefix $y_{<t}$. We define the privileged distillation loss as the forward KL divergence from the teacher distribution to the student distribution:

$$\mathcal{L}_{\text{distill}} = \frac{1}{\sum_{t=1}^{T} m_t} \sum_{t=1}^{T} m_t D_{\text{KL}} \left(p_T^t \,\|\, p_S^t \right), \tag{2}$$

where the divergence is computed over the complete vocabulary and m_t masks non-response or padded positions. This objective encourages the student to reproduce the token-level predictive behavior induced by the adaptively selected privileged context while remaining aligned with states visited by the current student policy.

Initial-Policy Anchoring. Privileged distillation alone may cause the student policy to drift excessively from the initial model. We therefore regularize the student toward the frozen anchor policy π_A on the same on-policy trajectory. Rather than computing a full-vocabulary divergence, the anchor regularization is evaluated on the token y_t actually sampled by the student. We define

$$\begin{aligned} d_t &= \log p_A^t(y_t) - \log p_S^t(y_t), \\ \mathcal{L}_{\text{anchor}} &= \frac{1}{\sum_{t=1}^{T} m_t} \sum_{t=1}^{T} m_t \left[\exp(d_t) - 1 - d_t \right]. \end{aligned} \tag{3}$$

This anchoring term constrains policy drift while preserving the privileged teacher as the primary source of cross-lingual supervision.

Overall Objective. The final training objective is $\mathcal{L} = \mathcal{L}_{\text{distill}} + \lambda \mathcal{L}_{\text{anchor}}$, where λ controls the strength of initial-policy anchoring. Only the student policy is updated, while the teacher and anchor policy remain frozen.

4 Experiments

4.1 Experimental Setup

Baselines. We compare APL-OPSD with five baselines: base, the original model without additional training; SFT, supervised fine-tuning on target-language reference responses; GRPO, on-policy optimization with sequence-level rewards (Shao et al., 2024); LRPO, multilingual policy optimization with adaptive rollout-language routing (Guo et al., 2026); and COPSD, crosslingual on-policy self-distillation using English as the fixed privileged language (Liu et al., 2026).

Implementation Details. We conduct experiments on Qwen3-1.7B, Qwen3-4B, and Qwen3-8B (Yang et al., 2025a). Our training corpus contains 3,936 instances from CARE and HelpSteer3, with target languages approximately balanced across 15 languages (Guo et al., 2025; Wang et al., 2026b). For APL-OPSD, we jointly train a single model over all target languages for 300 steps, using an effective batch size of 32 and a learning rate of 7×10^{-6}. The privileged teacher remains frozen, and the reference-anchoring coefficient is set to $\lambda = 2.0$. Every training configuration is run with three random seeds, 42, 1042, and 2042, and we report the arithmetic mean across the three resulting runs. Additional configurations are provided in the appendix.

Evaluation Tasks. We evaluate models on five multilingual benchmarks. CARE and CARE-Pro assess cultural and fine-grained regional knowledge (Guo et al., 2025; 2026); mGSM-v2 evaluates

Table 1: **Main results** across five multilingual benchmarks and three Qwen3 model scales. *Seen*/*Unseen* denote languages included in/held out from training. Bold and underlined values indicate the best and second-best results.

Approach	CARE	CARE-Pro			mGSM-v2			Global-MMLU-Lite			Include-Lite			Overall		
	Avg.	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*
Qwen3-1.7B																
Base	26.98	18.15	4.11	9.33	79.06	38.57	64.34	46.93	32.42	39.36	43.98	36.58	38.87	43.02	27.92	35.78
SFT	23.56	13.89	1.44	6.06	61.71	14.67	44.61	**50.39**	**35.38**	**42.56**	**48.47**	**39.65**	**42.39**	39.60	22.79	31.84
GRPO	26.44	17.71	4.11	9.16	77.90	37.67	63.27	48.52	33.97	40.93	48.14	38.15	41.24	43.74	28.48	36.21
LRPO	25.56	14.93	4.11	8.13	78.67	**40.00**	64.61	32.93	26.80	29.73	31.16	28.65	29.42	36.65	24.89	31.49
COPSD	26.44	18.40	3.70	9.16	77.90	39.00	63.76	45.89	32.01	38.65	43.26	36.10	38.32	42.38	27.70	35.27
APL-OPSD	**29.11**	**20.49**	**4.52**	**10.45**	**79.81**	39.67	**65.21**	48.50	34.09	40.98	46.61	38.24	40.84	**44.90**	**29.13**	**37.32**
Qwen3-4B																
Base	39.28	27.37	10.61	16.84	**93.84**	61.35	82.03	62.35	46.08	53.86	60.55	49.88	53.19	56.68	41.98	49.04
SFT	36.00	24.65	7.39	13.81	81.14	40.67	66.42	62.68	45.80	53.88	59.77	49.80	52.89	52.85	35.92	44.60
GRPO	37.56	25.35	11.50	16.65	93.14	62.00	81.82	61.57	45.88	53.39	59.65	49.84	52.88	55.45	42.31	48.46
LRPO	38.67	27.43	7.80	15.10	93.52	62.33	82.18	60.14	44.66	52.06	59.05	49.21	52.26	55.76	41.00	48.05
COPSD	40.67	27.43	9.03	15.87	92.95	59.00	80.61	61.57	45.74	53.31	59.38	50.08	52.96	56.40	40.96	48.68
APL-OPSD	**42.44**	**27.78**	**11.70**	**17.68**	92.38	**66.33**	**82.91**	**63.25**	**47.24**	**54.90**	**61.47**	**50.13**	**53.65**	**57.46**	**43.85**	**50.32**
Qwen3-8B																
Base	46.84	**35.01**	15.88	22.99	95.07	77.24	88.59	66.84	49.01	57.54	63.29	**54.05**	**56.91**	61.41	49.05	54.57
SFT	42.89	29.86	11.09	18.06	84.76	59.33	75.52	66.05	47.84	56.55	**64.14**	52.57	56.16	57.54	42.71	49.84
GRPO	48.22	32.29	18.28	23.48	93.90	76.00	87.39	66.84	49.22	57.65	62.49	53.81	56.50	60.75	49.33	54.65
LRPO	49.11	32.64	**18.69**	**23.87**	94.48	75.67	87.64	65.14	46.63	55.48	60.81	52.27	54.92	60.44	48.32	54.20
COPSD	46.67	32.29	16.84	22.58	**95.24**	**77.33**	**88.73**	66.59	48.36	57.08	62.70	53.54	56.38	60.70	49.02	54.29
APL-OPSD	**51.11**	32.99	18.28	23.74	**95.24**	**77.33**	**88.73**	**67.57**	**50.24**	**58.53**	63.54	53.85	56.85	**62.09**	**49.93**	**55.79**

multilingual mathematical reasoning (Peter et al., 2026); and Global-MMLU-Lite and INCLUDE-Lite evaluate multilingual factual and regional knowledge (Singh et al., 2025; Romanou et al., 2025). CARE, CARE-Pro, and mGSM-v2 are open-ended generation tasks, whereas Global-MMLU-Lite and INCLUDE-Lite are multiple-choice tasks. We report performance on languages included in training (*Seen*) and held-out languages (*Unseen*). The overall results are macro-averaged across the applicable benchmarks.

4.2 Main Results

Table 1 presents the main results across five multilingual benchmarks. APL-OPSD achieves the highest overall average at all three model scales, reaching 37.32, 50.32, and 55.79 on Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, respectively. These results outperform the strongest competing baseline by 1.11, 1.28, and 1.14 points. APL-OPSD also achieves the best overall performance on unseen languages at every scale and on seen languages at 1.7B, 4B, and 8B. Compared with COPSD, which uses English as a fixed privileged language, APL-OPSD improves the overall average by 2.05, 1.64, and 1.50 points across the three scales, suggesting a benefit from selecting the privileged language adaptively for each instance.

Across individual benchmarks, the largest gains appear on CARE, where APL-OPSD outperforms the strongest baseline by 2.13, 1.77, and 2.00 points at 1.7B, 4B, and 8B, respectively. It also achieves the best CARE-Pro results at 1.7B and 4B, and the best or tied-best mGSM-v2 average at all three scales. On the multiple-choice benchmarks, APL-OPSD achieves the strongest Global-MMLU-Lite results at 4B and 8B while remaining competitive on INCLUDE-Lite. Taken together, these results show aggregate improvements at each tested model scale, with gains varying by benchmark and being most pronounced on culturally grounded knowledge and held-out languages.

5 Analysis

We further analyze APL-OPSD along three dimensions. First, we evaluate whether adaptive privileged-language selection improves downstream performance over fixed and uninformed alternatives. Second, we examine how the selected languages vary across target languages, instances, and model scales. Finally, we quantify how often different selection strategies introduce privileged-context interference. Additional ablations on privileged-information content and teacher–student thinking modes are reported in Appendix C.

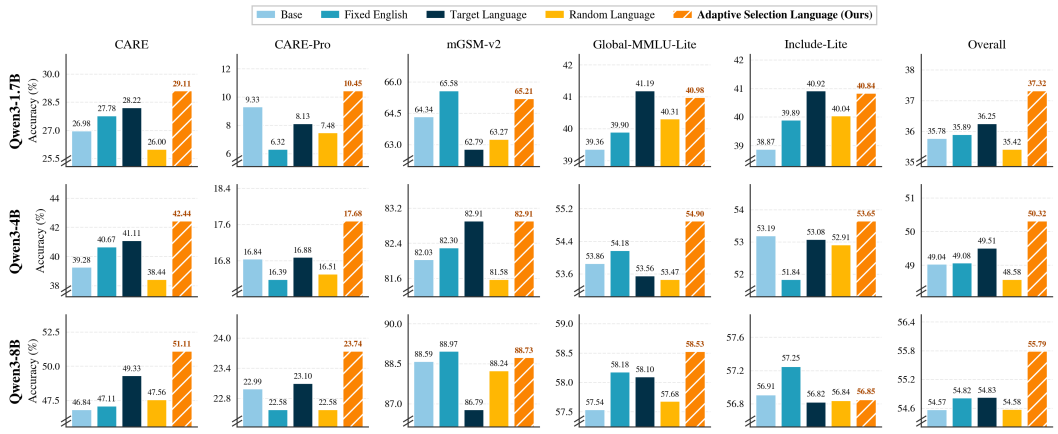

Figure 3: **Effect of privileged-language selection strategies.** Average performance across five multilingual benchmarks and three Qwen3 model scales. APLS achieves the strongest aggregate performance in this comparison.

5.1 Effectiveness of Privileged Language Selection

To isolate the effect of privileged-language selection, we compare APLS with three alternatives: *Fixed English*, which always uses English; *Target Language*, which uses the target-language question–answer pair; and *Random Language*, which samples a non-target language from the same candidate set as APLS. All variants differ only in the privileged language. As shown in Figure 3, APLS achieves the strongest aggregate performance across all three model scales, with higher scores than fixed-English and random selection. Notably, random selection remains close to the base model, suggesting that cross-lingual privileged information alone may be insufficient in this setting and that informative language selection matters. APLS also performs favorably against the target-language strategy in aggregate, supporting our choice to select privileged languages from non-target languages.

The benefit, however, varies across tasks. Adaptive selection yields particularly pronounced gains on CARE, whereas its advantage on mGSM-v2 is limited, with fixed English remaining highly competitive. This pattern is consistent with prior observations that reasoning-heavy tasks often favor dominant languages such as English, leaving less room for instance-specific language selection. In contrast, culturally and regionally grounded knowledge is more language-dependent, making adaptive privileged-language selection particularly beneficial.

5.2 Patterns of Adaptive Privileged-Language Selection

To better understand how APLS adapts its decisions beyond aggregate performance, we examine its selections across target languages and model scales. Figure 4 reports the selected privileged-language distribution conditioned on the target language, which is excluded from the candidate set. The distributions vary substantially across target languages: for example, Japanese instances frequently select Korean, whereas German instances often favor Dutch or English. Each row also remains distributed over multiple candidates, showing that instances with the same target language need not receive the same privileged language. Moreover, the relative preferences shift across model scales, suggesting that selection depends jointly on the instance, target language, and underlying model. Together with Figure 3, these patterns support instance-adaptive selection over a fixed language or deterministic target-to-language mapping. Marginal selection frequencies are reported in Appendix B.1.

5.3 Diagnostic Analysis of Privileged-Context Interference

To motivate instance-adaptive privileged-language selection, we examine how often a privileged context increases the perplexity (PPL) of the target-language reference answer relative to conditioning on the target-language query alone. A PPL increase indicates that the additional context interferes with the model's prediction of the reference answer. This analysis is intended to diagnose

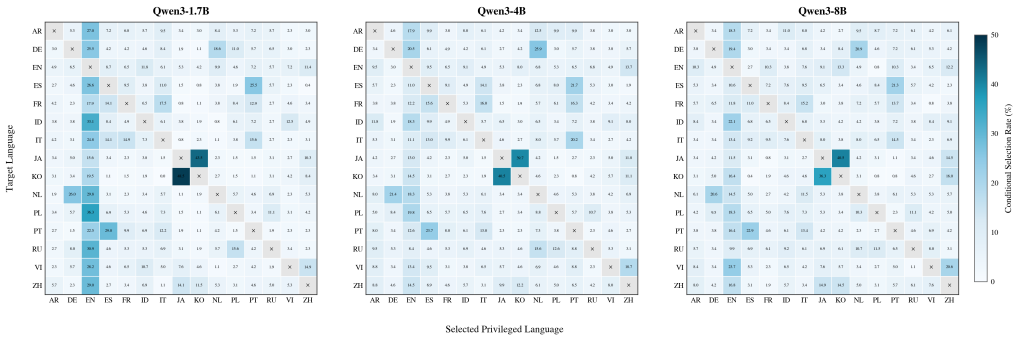

Figure 4: **Target-conditioned distribution of selected privileged languages.** Selection frequency of each privileged language conditioned on the target language. Each row corresponds to a target language and each column to a candidate privileged language.

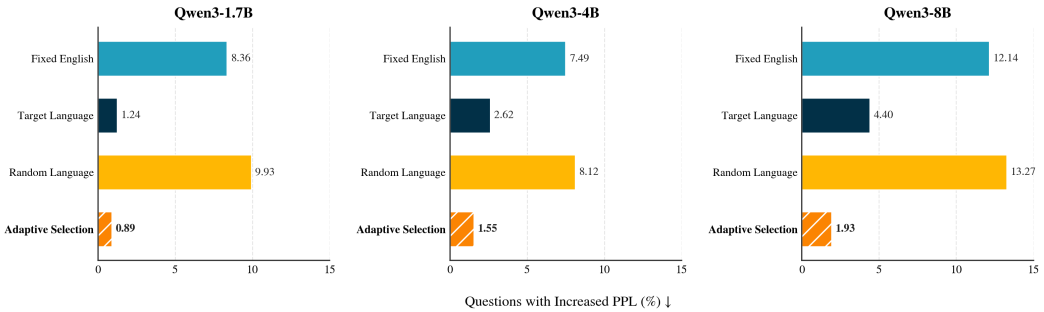

Figure 5: **Diagnostic of privileged-context interference.** Percentage of instances where the selected privileged context increases the target-answer perplexity (PPL) relative to the target-language query alone. Lower is better.

the potential interference caused by fixed or uninformed privileged-language choices, rather than to serve as an independent evaluation of selection quality. Since APLS directly selects the candidate with the lowest target-answer PPL, it is expected by construction to achieve a lower PPL-increase rate than fixed or random selection.

As shown in Figure 5, fixed-English selection increases PPL for 8.36%, 7.49%, and 12.14% of the training instances on Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, respectively. Random selection produces similarly high rates of 9.93%, 8.12%, and 13.27%, showing that an arbitrary cross-lingual context can be unhelpful or even introduce interference. The target-language strategy yields lower rates of 1.24%, 2.62%, and 4.40%, while APLS further reduces them to 0.89%, 1.55%, and 1.93%.

These observations quantify the risk of applying a globally fixed or uninformed privileged language to every instance and motivate screening candidate contexts according to their instance-level likelihood effects. Whether this selection criterion translates into downstream performance improvements is evaluated separately through the strategy comparisons in Figure 3, where APLS is compared with fixed-English, target-language, and random selection under the same training framework.

6 Conclusion

We introduced APL-OPSD, a multilingual on-policy self-distillation framework that adaptively selects the privileged language for each training instance. Its APLS component uses a one-time likelihood probe to identify the non-target language that most increases the model's support for the target-language reference response. A frozen teacher then uses the selected crosslingual question–answer pair to provide dense token-level supervision on trajectories generated by the student. Experiments across multiple benchmarks and model scales show that APL-OPSD achieves the strongest aggregate performance among the compared methods. Further analyses suggest that performance benefits from selecting informative privileged languages, providing complete question–answer contexts, and appropriately configuring teacher and student thinking modes.

AI USE STATEMENT

Generative AI tools were used to generate synthetic multilingual datasets, as described in Section A.2, and to improve the clarity, grammar, and style of the manuscript. They were not used to formulate the research hypotheses, design the methodology or experiments, generate or interpret experimental results, or make scientific claims. All AI-assisted outputs and edits were carefully reviewed and verified by the authors. The authors take full responsibility for the accuracy, originality, and final content of the paper.

REPRODUCIBILITY STATEMENT

We provide the implementation and evaluation details needed to reproduce our experiments throughout the paper and appendix. The Methodology section specifies the adaptive privileged-language selection procedure, student and teacher contexts, and training objectives. The Experimental Setup and Appendix A.1 report the model backbones, optimization hyperparameters, hardware, baseline implementations, random seeds, and result-aggregation protocol. Appendix A.2 documents the training-data sources, target-language assignment, translation model and prompt, and quality-control procedure. Appendix A.3 further specifies the evaluation datasets, language splits, retained example counts, and scoring protocols.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, et al. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 4232–4267, 2023.

Xi Ai, Mahardika Krisna Ihsani, and Min-Yen Kan. Are knowledge and reference in multilingual language models cross-lingually consistent? *arXiv preprint arXiv:2507.12838*, 2025.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. Rlhf can speak many languages: Unlocking multilingual preference optimization for llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13134–13156, 2024.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. Care: Multilingual human preference learning for cultural awareness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32854–32883, 2025.

Geyang Guo, Hiromi Wakaki, Yuki Mitsufuji, Alan Ritter, and Wei Xu. Learning to route languages for multilingual policy optimization. *arXiv preprint arXiv:2605.25360*, 2026.

Yue Huang, Chenrui Fan, Yuan Li, Siyuan Wu, Tianyi Zhou, Xiangliang Zhang, and Lichao Sun. 1+ 1¿ 2: Can large language models serve as cross-lingual knowledge aggregators? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13394–13412, 2024.

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

Deokhyung Kang, Seonjeong Hwang, Daehui Kim, Hyounghun Kim, and Gary Geunbae Lee. Why do multilingual reasoning gaps emerge in reasoning language models? In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31684–31716, 2026.

Huiyuan Lai and Malvina Nissim. 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)*, pp. 12012–12026, 2024.

Yihao Li, Jiayi Xin, Miranda Muqing Miao, Qi Long, and Lyle Ungar. The impact of language mixing on bilingual llm reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32519–32536, 2025.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Luu Anh Tuan, and Lidong Bing. Is translation all you need? a study on solving multilingual tasks with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9594–9614, 2025.

Yihong Liu, Raoyuan Zhao, Michael A Hedderich, and Hinrich Schütze. Crosslingual on-policy self-distillation for multilingual reasoning. *arXiv preprint arXiv:2605.09548*, 2026.

Jan-Thorsten Peter, David Vilar, Tobias Domhan, Dan Malkin, and Markus Freitag. Mind the gap... or not? how translation errors and evaluation details skew multilingual results. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 191–204, 2026.

Kartik Ravisankar, Hyojung Han, Sarah Wiegreffe, and Marine Carpuat. Can you map it to english? the role of cross-lingual alignment in the multilingual performance of llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4854–4872, 2026.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed Haggag, Alfonso Amayuelas, et al. Include: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations*, volume 2025, pp. 83291–83322, 2025.

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

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

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

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, 2025.

Guijin Son, Donghun Yang, Hitesh Laxmichand Patel, Amit Agarwal, Hyunwoo Ko, Srikant Panda, Minhyuk Kim, Dasol Choi, Kyong-Ha Lee, Youngjae Yu, et al. Pushing on multilingual reasoning models with language-mixed chain-of-thought. In *International Conference on Learning Representations*, volume 2026, pp. 79387–79407, 2026.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5075–5094, 2025.

Xinye Wang, Junxiao Liu, and Shujian Huang. Rp-opsd: Reasoning-pivot-guided on-policy self-distillation for multilingual reasoning transfer. *arXiv preprint arXiv:2608.06347*, 2026a.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages. *Advances in Neural Information Processing Systems*, 38, 2026b.

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

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. Language imbalance driven rewarding for multilingual self-improving. In *International Conference on Learning Representations*, volume 2025, pp. 43911–43936, 2025b.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Kaiyu Huang, Yufeng Chen, Jinan Xu, and Jie Zhou. Think natively: Unlocking multilingual reasoning with consistency-enhanced reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20766–20783, 2026.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. Plug: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7025–7046, 2024.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026a.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, et al. Mpo: Multilingual safety alignment via reward gap optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23564–23587, 2025.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, et al. When less language is more: Language-reasoning disentanglement makes llms better multilingual reasoners. *Advances in Neural Information Processing Systems*, 38:38608–38642, 2026b.

Weihua Zheng, Roy Ka-Wei Lee, Zhengyuan Liu, Kui Wu, AiTi Aw, and Bowei Zou. Ccl-xcot: An efficient cross-lingual knowledge transfer method for mitigating hallucination generation. 2025.

APPENDIX

A ADDITIONAL EXPERIMENTAL DETAILS

A.1 TRAINING AND IMPLEMENTATION DETAILS

We conduct experiments with Qwen3-1.7B, Qwen3-4B, and Qwen3-8B. For all model scales, the student is optimized with full-parameter on-policy self-distillation, while the teacher is a frozen copy of the initial model. The student is initialized from the same checkpoint but does not observe the teacher's privileged input. At each step, the student first samples its own response, and the token-level teacher and student distributions are then matched along this student-generated trajectory.

Unless otherwise specified, we train with bfloat16 precision on eight NVIDIA H20 GPUs using DeepSpeed ZeRO-2 with optimizer offloading. We use a learning rate of 7×10^{-6}, an effective batch size of 32, a maximum context length of 4,096 tokens, and a maximum generation length of 1,024 tokens. Student rollouts are sampled with temperature 1.1, top-p 0.95, and top-k 20. The distillation loss is the full-vocabulary forward KL divergence from the privileged teacher to the student and is combined with an anchor loss of weight 2.0 to limit deviation from the initial policy. Gradients are clipped at 1.0, and models are trained for 300 optimization steps.

Random Seeds and Result Aggregation. We independently train every post-training configuration—including APL-OPSD, all trainable baselines, and all ablation variants—with random seeds 42, 1042, and 2042. Unless explicitly stated otherwise, every reported score for a trained method is the arithmetic mean of the corresponding evaluations from these three runs.

Baseline Implementations. We reproduce COPSD and LRPO from their officially released codebases,[1][2] rather than reimplementing their algorithms from scratch. We replace the backbone, training data, supported-language set, and sequence-length settings as needed to match our Qwen3-based experimental protocol, while retaining the corresponding official objectives and algorithm-specific implementation. Thus, the reported results are obtained by running the official implementations under our common data and evaluation setting, rather than by copying numbers from the original papers.

Our GRPO baseline uses two independently computed binary rewards: a language-consistency reward and an answer-correctness reward. For a generated response y, target language t, and reference answer y^*, they are defined as

$$r_{\text{lang}}(y) = \mathbb{I}[\text{Lang}(y) = t]\,,$$
$$r_{\text{ans}}(y, y^*) = \mathbb{I}[\text{Correct}(y, y^*)]\,.$$

Both components take values in $\{0, 1\}$. We combine them by direct, unweighted addition,

$$r(y) = r_{\text{lang}}(y) + r_{\text{ans}}(y, y^*),$$

so the final reward is 0, 1, or 2. This design separately rewards answering in the requested language and producing the correct answer: a factually correct response in the wrong language, or an incorrect response in the correct language, receives a reward of 1. The resulting rewards are normalized within each rollout group and optimized with the standard GRPO objective.

A.2 MULTILINGUAL DATA CONSTRUCTION

The training set contains 3,936 instances, including 2,828 cultural question–answer examples from CARE Guo et al. (2025) and 1,108 open-ended instructions from HelpSteer3 Wang et al. (2026b). Each instance is available in the following 15 languages: Arabic (`ar`), Chinese (`zh`), Dutch (`nl`), English (`en`), French (`fr`), German (`de`), Indonesian (`id`), Italian (`it`), Japanese (`ja`), Korean (`ko`), Polish (`pl`), Portuguese (`pt`), Russian (`ru`), Spanish (`es`), and Vietnamese (`vi`). We assign the target languages in a balanced manner, resulting in 262 or 263 instances per target language. These 15 languages define the *Seen* set throughout evaluation; any benchmark language outside this

[1] `https://github.com/cisnlp/COPSD`
[2] `https://github.com/Guochry/LRPO`

set is categorized as *Unseen*. We follow the official train–test split of CARE Guo et al. (2025), using only the training split and its multilingual variants for training, while reserving the test split exclusively for evaluation.

Translation Procedure. We construct the multilingual views with DeepSeek-V4-Flash, with native thinking disabled, temperature set to 0, and a maximum output length of 8,000 tokens. The original human-written version in an instance's source language is retained rather than machine translated. For each instance, all remaining target-language translations are requested together in one JSON response to encourage consistent terminology across language views. CARE questions are translated directly from their original Chinese, Japanese, or Arabic text rather than through an English pivot; HelpSteer3 user messages are likewise translated directly from their original language.

Translation Prompt and Quality Control. The translation prompt is summarized as follows: "Translate the question or user message into each listed target language. Preserve every proper noun and place name exactly; preserve the original communicative form, line breaks, code, markup, and list structure; do not answer the input, add or remove information, or repair incomplete text; and output only a JSON object mapping each requested language code to its translation." For question-answering examples, the prompt additionally instructs the model to keep the input as a question without answering it; for open-ended HelpSteer3 messages, it forbids changing statements into questions or carrying out the requested task. We parse each response as JSON and require every requested language field to be non-empty and to pass a target-script check. Missing or wrong-script languages are retried for up to three attempts, while refusal-like placeholder responses are rejected. Instances without valid views in all 15 languages are excluded. Together, the prompt constraints and validation are intended to preserve the semantic content and structural form of the source while preventing incomplete or wrong-language views from entering training.

The student observes only the target-language question and the target-language response instruction. The teacher additionally receives a privileged question–answer pair that conveys the same information in another language. All structural labels, transition instructions, and response instructions are written in the target language; only the privileged question–answer content remains in the selected privileged language.

For adaptive privileged language selection, we score the target-language ground-truth answer under each candidate privileged-language context using teacher forcing. The target language itself is excluded, and the language yielding the lowest answer perplexity among the remaining 14 candidates is selected for that instance. Fixed-English, target-language, and uniformly random non-target-language variants use the same prompt structure and differ only in the language of the privileged question–answer pair.

A.3 Evaluation Protocols

We evaluate on five multilingual benchmarks: CARE Guo et al. (2025), CARE-Pro, mGSM-v2, Global-MMLU-Lite, and Include-Lite. Table 2 reports the exact Seen/Unseen partition and the number of examples actually used after dataset preprocessing. Across the five benchmarks, our evaluation contains 22,274 examples in 54 distinct languages. Per-benchmark scores are computed over all retained examples in the corresponding split, and the Overall score is the macro-average of the five benchmark scores.

CARE contains 150 examples in each of Arabic, Japanese, and Chinese. CARE-Pro contains 84 Spanish, 204 Chinese, and 487 Hindi examples. The evaluated mGSM-v2 subset contains 125 examples per language, except Telugu with 124. Global-MMLU-Lite contains 400 examples for every Seen language; its Unseen subsets contain between 319 and 400 examples per language. Include-Lite contains between 89 and 253 retained examples per language. The less common held-out codes in Table 2 denote Bengali (`bn`), Czech (`cs`), Welsh (`cy`), Hindi (`hi`), Hungarian (`hu`), Burmese (`my`), Odia (`or`), Slovak (`sk`), Albanian (`sq`), Swahili (`sw`), Tajik (`tg`), Yoruba (`yo`), Telugu (`te`), Thai (`th`), Azerbaijani (`az`), Belarusian (`be`), Bulgarian (`bg`), Greek (`el`), Estonian (`et`), Basque (`eu`), Persian (`fa`), Finnish (`fi`), Hebrew (`he`), Croatian (`hr`), Armenian (`hy`), Georgian (`ka`), Kazakh (`kk`), Lithuanian (`lt`), Macedonian (`mk`), Malayalam (`ml`), Malay (`ms`), Nepali (`ne`), Serbian (`sr`), Tamil (`ta`), Tagalog (`tl`), Turkish (`tr`), Ukrainian (`uk`), Urdu (`ur`), and Uzbek (`uz`).

Table 2: **Evaluation-set composition.** Language codes follow ISO 639 conventions. Counts are the numbers of examples actually evaluated, rather than the nominal sizes of the public dataset releases.

Benchmark	Seen languages	Unseen languages	Seen	Unseen	Total
CARE	ar, ja, zh	–	450	0	450
CARE-Pro	es, zh	hi	288	487	775
mGSM-v2	de, en, es, fr, ja, ru, zh	bn, sw, te, th	875	499	1,374
Global-MMLU-Lite	ar, de, en, es, fr, id, it, ja, ko, pt, zh	bn, cs, cy, hi, hu, my, or, sk, sq, sw, tg, yo	4,400	4,537	8,937
Include-Lite	ar, de, es, fr, id, it, ja, ko, nl, pl, pt, ru, vi, zh	az, be, bg, bn, el, et, eu, fa, fi, he, hi, hr, hu, hy, ka, kk, lt, mk, ml, ms, ne, sq, sr, ta, te, tl, tr, uk, ur, uz	3,316	7,422	10,738
Total			9,329	12,945	22,274

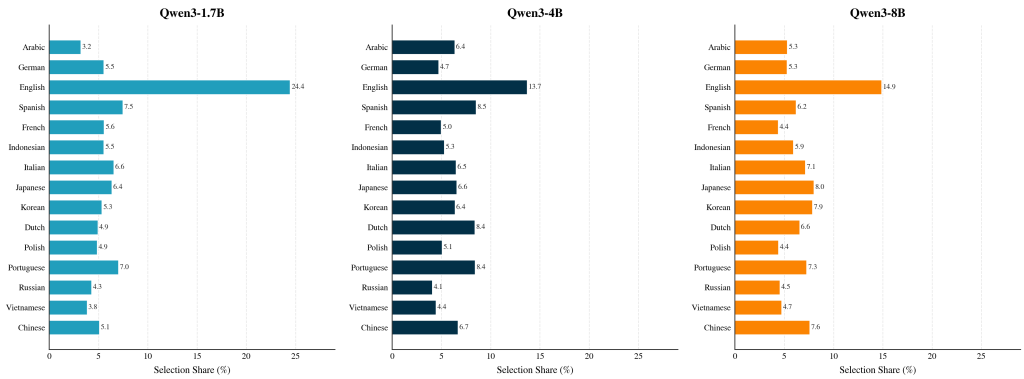

Figure 6: **Marginal distribution of selected privileged languages.** Selection frequency of each privileged language across all training instances for three Qwen3 model scales.

CARE responses are scored by `gpt-oss-120b` on a 1–10 scale and linearly rescaled to 0–100. CARE-Pro follows the official LRPO four-way judging protocol, where the judge assigns CORRECT, CORRECT-BUT-WRONG-LANGUAGE, INCORRECT, or NOT-ATTEMPTED; only CORRECT is counted as correct. For mGSM-v2, we extract and exactly compare the final numerical answer. Global-MMLU-Lite and Include-Lite are evaluated by exact matching of the predicted multiple-choice option.

B ADDITIONAL ANALYSIS OF PRIVILEGED-LANGUAGE SELECTION

B.1 DISTRIBUTION OF SELECTED PRIVILEGED LANGUAGES

We further analyze which languages are selected by APLS as privileged languages. Figure 6 shows the marginal selection frequency of each candidate language across all training instances, complementing the target-conditioned analysis in Figure 4. The distribution is clearly non-uniform and varies across model scales. English is selected most frequently, accounting for 24.4%, 13.7%, and 14.9% of the selections for Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, respectively. Nevertheless, the remaining selections are widely distributed across other languages, indicating that APLS does not simply collapse to English as a globally preferred privileged language. Moreover, the substantial changes in selection frequencies across model scales suggest that the most useful privileged language also depends on the knowledge-access patterns of the underlying model.

B.2 QUALITATIVE EXAMPLES OF PRIVILEGED-LANGUAGE SELECTION

Table 5 presents representative examples from Qwen3-4B and Qwen3-8B illustrating how the choice of privileged language affects target-answer perplexity. For each instance, we compare the student-

Table 3: **Ablation results** on different privileged-information contents across five multilingual benchmarks and three Qwen3 model scales. *Seen*/*Unseen* denote languages included in/held out from training. Bold and underlined values indicate the best and second-best results.

Approach	CARE	CARE-Pro			mGSM-v2			Global-MMLU-Lite			Include-Lite			Overall		
	Avg.	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*	*Seen*	*Unseen*	*Avg.*
Qwen3-1.7B																
Base	26.98	18.15	4.11	9.33	79.06	38.57	64.34	46.93	32.42	39.36	43.98	36.58	38.87	43.02	27.92	35.78
Question + Answer	**29.11**	**20.49**	**4.52**	**10.45**	**79.81**	39.67	**65.21**	**48.50**	**34.09**	**40.98**	46.61	**38.24**	**40.84**	**44.90**	**29.13**	**37.32**
Answer Only	27.71	18.75	3.70	9.29	78.29	39.00	64.00	48.36	33.88	40.81	**46.73**	37.80	40.57	43.97	28.60	36.48
Question Only	27.33	19.79	4.31	10.06	78.29	**41.00**	64.73	47.84	32.74	39.96	45.63	37.73	40.18	43.78	28.95	36.45
Qwen3-4B																
Base	39.28	27.37	10.61	16.84	**93.84**	61.35	82.03	62.35	46.08	53.86	60.55	49.88	53.19	56.68	41.98	49.04
Question + Answer	**42.44**	27.78	**11.70**	**17.68**	92.38	**66.33**	**82.91**	**63.25**	**47.24**	**54.90**	**61.47**	50.13	**53.65**	**57.46**	**43.85**	**50.32**
Answer Only	40.22	28.47	11.09	17.55	91.43	61.67	80.61	62.93	46.22	54.21	60.55	**50.36**	53.52	56.72	42.34	49.22
Question Only	38.67	**29.51**	10.47	17.55	90.86	62.00	80.36	62.52	45.63	53.71	60.64	49.89	53.22	56.44	42.00	48.70
Qwen3-8B																
Base	46.84	**35.01**	15.88	22.99	95.07	77.24	88.59	66.84	49.01	57.54	63.29	54.05	**56.91**	61.41	49.05	54.57
Question + Answer	**51.11**	32.99	**18.28**	**23.74**	**95.24**	77.33	88.73	67.57	**50.24**	**58.53**	**63.54**	53.85	56.85	**62.09**	49.93	**55.79**
Answer Only	48.00	32.29	17.04	22.71	94.10	**80.67**	**89.21**	**67.59**	49.89	58.35	62.91	**54.20**	56.90	60.98	**50.45**	55.03
Question Only	50.67	33.68	17.25	23.35	94.67	78.33	88.73	66.91	48.66	57.39	62.79	53.81	56.59	61.74	49.51	55.35

only input with fixed-English, target-language, and adaptively selected privileged contexts while keeping the target-language query unchanged. The selected non-target language can substantially reduce the perplexity of the target-language reference answer, whereas fixed-English or even target-language privileged context is not consistently beneficial.

In Case 1, on Qwen3-4B, the adaptively selected Korean question–answer pair reduces the PPL from 83.22 to 22.30. In contrast, fixed English increases it to 93.30, and the target-language Italian context further increases it to 139.42. Case 2 shows the same pattern on Qwen3-8B: the Korean privileged context reduces the PPL from 46.98 to 16.18, while fixed English provides only a modest reduction to 42.64 and the target-language Indonesian context increases it to 48.22. These cross-model examples show that semantically equivalent privileged information can have markedly different effects depending on the language in which it is expressed, motivating instance-specific privileged-language selection rather than reliance on a globally fixed language.

C Additional Ablation Studies

C.1 Effect of Privileged Information Content

To determine which components of privileged information contribute to self-distillation, we keep the privileged language selected by APLS fixed and vary the content provided to the teacher. We compare the complete question–answer pair with two restricted variants that provide only the answer or only the question.

As shown in Table 3, the complete question–answer context achieves the highest overall performance across all three model scales, reaching 37.32, 50.32, and 55.79 on Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, respectively. Answer-only supervision retains most of the gains but consistently underperforms the complete context, whereas question-only supervision yields weaker and less consistent improvements. This indicates that the answer provides the primary privileged signal, while the question contributes complementary context for grounding the answer and facilitating more effective knowledge transfer.

C.2 Effect of Teacher and Student Thinking Modes

In OPSD Zhao et al. (2026a), the teacher evaluates trajectories generated by the student rather than producing an independent response. The thinking configurations of the two policies should therefore be aligned. When one policy enables thinking while the other disables it, the teacher evaluates a trajectory whose format differs from that implied by its own prompt, introducing distributional discrepancies unrelated to privileged information. In principle, both *On/On* and *Off/Off* preserve this consistency, while the former additionally allows explicit reasoning during generation.

Table 4 compares all four student–teacher configurations. *On/On* achieves the highest overall score at every model scale, reaching 37.32, 50.32, and 55.79 on Qwen3-1.7B, Qwen3-4B, and Qwen3-

Table 4: **Ablation study** on reasoning configurations of the student and teacher models. Only average scores are reported. *Think (S/T)* denotes the reasoning configurations of the student/teacher. Bold and underlined values indicate the best and second-best results.

Backbone	Think (S/T)	CARE	CARE-Pro	mGSM-v2	Global-MMLU-Lite	Include-Lite	Overall
Qwen3-1.7B	On/On	**29.11**	**10.45**	**65.21**	<u>40.98</u>	<u>40.84</u>	**37.32**
	Off/On	<u>28.89</u>	9.68	<u>63.64</u>	39.15	39.79	36.23
	On/Off	28.67	9.29	<u>63.64</u>	**41.37**	**41.22**	<u>36.84</u>
	Off/Off	28.67	<u>10.32</u>	63.15	39.23	40.41	36.36
Qwen3-4B	On/On	**42.44**	<u>17.68</u>	<u>82.91</u>	**54.90**	**53.65**	**50.32**
	Off/On	41.33	17.16	81.94	54.17	52.80	49.48
	On/Off	<u>41.56</u>	17.42	<u>82.91</u>	<u>54.36</u>	<u>53.42</u>	49.93
	Off/Off	41.33	**18.32**	**83.52**	53.97	53.03	<u>50.03</u>
Qwen3-8B	On/On	**51.11**	**23.74**	<u>88.73</u>	58.53	56.85	**55.79**
	Off/On	48.89	<u>22.06</u>	88.48	<u>58.55</u>	<u>57.03</u>	55.00
	On/Off	48.67	<u>22.06</u>	87.52	**59.43**	**57.38**	<u>55.01</u>
	Off/Off	<u>49.78</u>	20.65	**89.70**	58.41	56.44	55.00

8B, respectively. Matched configurations also tend to be more stable than mismatched ones. Among the mismatched settings, keeping student thinking enabled generally performs better than enabling it only for the teacher, suggesting that the quality and format of student-generated trajectories are particularly important for on-policy distillation.

The differences nevertheless vary across benchmarks. For example, *Off/Off* performs strongly on mGSM-v2 at the larger model scales, while several mismatched configurations achieve the best result on individual tasks. Thus, the results do not imply that *On/On* dominates every benchmark. Instead, they support it as a strong default configuration in terms of aggregate multilingual performance. We therefore use *On/On* for the main APL-OPSD experiments.

Table 5: Qualitative examples of language-dependent privileged inputs on Qwen3-4B and Qwen3-8B. The left column shows the actual strategy-specific input; all teacher inputs append the same target-language question shown in the case header. The fixed template instruction is omitted only to avoid repetition. Lower perplexity is better, and arrows show changes relative to the student-only Base condition.

Actual input	**Answer PPL**
Case 1: The third great composer of Vienna (Qwen3-4B) `care/ja/1562` **Target question (Italian):** I tre grandi compositori di Vienna sono Mozart, Haydn e chi è il terzo? **Ground-truth answer:** È Beethoven.	
Base (student input only) The target Italian question above; no privileged question–answer pair.	83.22
Fixed English privileged input **[EN] Question:** The three great composers of Vienna are Mozart, Haydn, and who is the third? **[EN] Answer:** Beethoven.	93.30 (↑ 12.11%)
Target-language privileged input **[IT] Domanda:** I tre grandi compositori di Vienna sono Mozart, Haydn e chi è il terzo? **[IT] Risposta:** È Beethoven.	139.42 (↑ 67.53%)
Adaptive privileged input (Korean, ours) **[KO] 질문:** 빈의3대작곡가라면모차르트, 하이든, 그리고나머지한명은누구인가? **[KO] 답변:** 베토벤입니다.	**22.30** **(↓ 73.20%)**
Case 2: German philosopher associated with dialectics (Qwen3-8B) `care/ja/1054` **Target question (Indonesian):** Siapakah filsuf klasik Jerman yang mendirikan dialektika dan dikenal melalui karya-karya seperti "Fenomenologi Roh"? **Ground-truth answer:** Hegel.	
Base (student input only) The target Indonesian question above; no privileged question–answer pair.	46.98
Fixed English privileged input **[EN] Question:** Who is the German classical philosopher who established dialectics and is known for works such as the "Phenomenology of Spirit"? **[EN] Answer:** Hegel.	42.64 (↓ 9.23%)
Target-language privileged input **[ID] Pertanyaan:** Siapakah filsuf klasik Jerman yang mendirikan dialektika dan dikenal melalui karya-karya seperti "Fenomenologi Roh"? **[ID] Jawaban:** Hegel.	48.22 (↑ 2.63%)
Adaptive privileged input (Korean, ours) **[KO] 질문:**《정신현상학》등의저서로알려져있으며변증법을확립한독일고전철학자는누구입니까? **[KO] 답변:** 헤겔입니다.	**16.18** **(↓ 65.56%)**