DuDi: Dual-Signal Distillation with Cross-Lingual Verbalizer

Patomporn Payoungkhamdee[1*†], Tinnakit Udsa[1*], Jian Gang Ngui[2], Sarana Nutanong[1], Alham Fikri Aji[3], Peerat Limkonchotiwat[2]
[1]School of Information Science and Technology, VISTEC [2]AI Singapore [3]MBZUAI
{patomporn.p_s21,tinnakit.u_s24}@vistec.ac.th, peerat@aisingapore.org
GitHub Hugging Face

Abstract

Small language models (SLMs) are efficient and scalable, but their multilingual capabilities degrade severely at sub-billion scales, especially for Southeast Asian (SEA) languages. We introduce DuDi, a dual-signal multilingual distillation framework that combines an online sequence-level signal with off-policy and on-policy token-level signals. DuDi further uses a cross-lingual verbalizer to refine teacher feedback and improve teacher-student transferability in multilingual settings. Experiments on SEA-HELM across multiple model families, scales, and teacher–student settings show that DuDi consistently outperforms competitive distillation baselines. Ablations and analyses confirm that sequence-level optimization, token-level supervision, and cross-lingual verbalization provide complementary and transferable learning signals for multilingual SLMs.

1 Introduction

Small language models (SLMs) have recently attracted growing attention due to their efficiency and scalability (Hu et al., 2024; Nguyen et al., 2024; Subramanian et al., 2025; Wang et al., 2024). For instance, Qwen2.5-1.5B and -0.5B achieve about 1.5× and 2.2× higher inference throughput than Qwen2.5-7B, while reducing memory use by 48% and 81%, respectively.[1] These gains reduce deployment costs and enable more efficient large-scale serving. Compared with larger models, SLMs offer practical advantages in scalability, computation, and memory usage (Hu et al., 2024). These properties make them suitable for resource-constrained and edge-device deployment (Liu et al., 2024), while supporting real-world applications at scale (Pham et al., 2025; Chen et al., 2025).

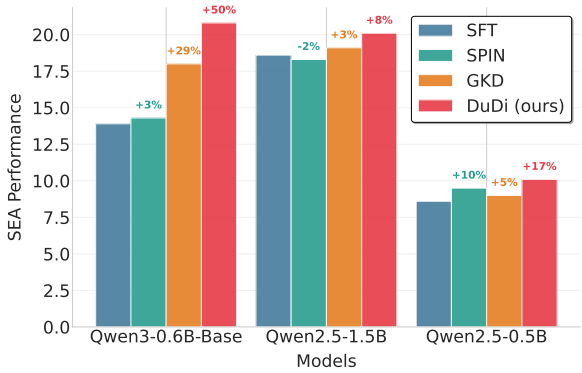


Figure 1: Comparison of SEA performance across different model scales and training frameworks, evaluated using the SEA-HELM benchmark (↑). Details of each model are provided in Section 4.

However, multilingual capabilities in SLMs remain limited (Qin et al., 2025; Xuan et al., 2025), especially for Southeast Asian (SEA) languages, a highly diverse region with hundreds of millions of speakers. As shown in Figure 1, SEA performance drops substantially as model size falls below the billion-scale regime. In particular, the Supervised Fine-Tuning (SFT) variant of Qwen2.5-0.5B drops sharply relative to Qwen2.5-1.5B, while the newer Qwen3-0.6B still shows limited SEA performance under standard SFT. These results suggest that scaling down weakens multilingual understanding, motivating training strategies tailored for SLMs.

A common approach to improve SLMs' performance is knowledge distillation (KD), which transfers knowledge from a larger teacher to a smaller student (Hinton et al., 2015; Kim and Rush, 2016; Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024, 2025). Despite recent advances in KD, multilingual distillation remains largely limited to task-specific or data-centric settings (Payoungkhamdee et al., 2024; Zhang et al., 2024), leaving general-purpose multilingual distillation for SLMs underexplored. This gap is especially pronounced for SEA languages due to their linguistic diversity and

*Equal contribution
†Work was conducted while Patomporn Payoungkhamdee was a visiting scholar at AI Singapore
[1]Measured with BF16 precision and input length 6144, based on benchmarks from https://qwen.readthedocs.io/en/v2.5/benchmark/speed_benchmark.html

2023
Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2023–2037
November 6-10, 2026 ©2026 Association for Computational Linguistics

limited high-quality training data, motivating multilingual KD strategies designed for SLMs.

To address this limitation, we propose **Du**al-Signal **Di**stillation with Cross-Lingual Verbalizer (DuDi), a general-purpose framework designed for multilingual distillation in SLMs. DuDi builds a unified KD framework around three components: *Sequence Signal*, *Token Signal*, and *Cross-Lingual Verbalizer*. For the sequence signal, DuDi introduces an online sequence-level objective that guides the student policy toward the ground-truth direction. For the token signal, DuDi uses both off-policy and on-policy supervision: off-policy signals come from the training corpus, while on-policy signals come from student-generated responses. To facilitate knowledge transferability, DuDi uses a cross-lingual verbalizer to refine teacher logits during on-policy distillation, aligning student responses with the ground-truth demonstrations. This design enables us to better facilitate student learning in a multilingual environment.

To evaluate DuDi, we compare it with competitive methods under the SEA training and evaluation framework, using SEA-Instruct[2] and SEA-HELM (Susanto et al., 2025). The experimental results demonstrate that DuDi achieves the strongest overall performance under the Qwen2.5-0.5B setting, with gains across most SEA languages. This trend generalizes across scales and architectures, demonstrating scalability and robustness. Ablations show consistent degradation when any DuDi component is removed, highlighting the need to jointly optimize sequence-level objective, dual-policy token signals, and the cross-lingual verbalizer. Finally, analysis of the DuDi verbalizer demonstrates that it provides richer learning signals for teacher-student distillation.

In conclusion, our contributions are as follows:

- We propose **DuDi**, a multilingual knowledge distillation framework that integrates sequence-level signals and token-level, improving SEA performance in small LMs.
- We introduce a **cross-lingual verbalizer** that better facilitates on-policy distillation.
- We conduct **ablations and analyses** to assess each component, showing the effectiveness of dual-signal and verbalizer designs.

2 Background

The multilingual training corpus consists of an input x, a ground-truth y, and a language l. Formally, this dataset is defined as $\mathcal{D} = \{(x_i, y_i, l_i)\}_{i=1}^{N}$, where N denotes the total number of training samples. To learn from the data, a fine-tuning methodology is employed to optimize the policy π_θ. This process involves minimizing an objective function, denoted as $\mathcal{L}(x_i, y_i, l_i; \pi_\theta)$, which serves as a metric for the difference between the model's stochastic predictions and the ground-truths. Existing methods structure this objective differently to address distinct learning dynamics.

Off-Policy Fine-Tuning. This method represents a straightforward approach, typically grounded in a cross-entropy objective. Given a model policy π_θ, the loss function is formulated as

$$\mathcal{L}_{\text{Off-FT}} = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[-w \log \pi_\theta(y|x) \right], \quad (1)$$

where w serves as a weighting coefficient to modulate the training signal. In standard Supervised Fine-Tuning (SFT), $w = 1$, treating all tokens with equal importance. In Dynamic Fine-Tuning (DFT) (Wu et al., 2026), w is defined as $\text{sg}(\pi_\theta(y|x))$, where $\text{sg}(\cdot)$ denotes the stop-gradient operator. This token-weighting is designed to stabilize gradient magnitudes and improve generalization during the fine-tuning process.

Iterative Fine-Tuning. To address the limitation of static training data, prior work has explored iterative self-play for policy refinement (Tesauro, 1995; Silver et al., 2017). Chen et al. (2024b) proposed Self-Play Fine-Tuning (SPIN), a bootstrapping framework that improves the model by distinguishing ground-truths and self-generated responses sampled from an SFT-initialized reference policy $y' \sim \pi_{\theta_{\text{Ref}}}(x)$. This approach optimizes the policy by maximizing an Integral Probability Metric against a previous iteration of the self. The iterative fine-tuning objective is defined as

$$\mathcal{L}_{\text{IFT}} = \mathbb{E}_{(x,y)\sim\mathcal{D},y'} \left[\ell \left(\lambda \log \frac{\pi_{\theta_{\text{Ref}}}(y|x)\pi_\theta(y'|x)}{\pi_\theta(y|x)\pi_{\theta_{\text{Ref}}}(y'|x)} \right) \right], \quad (2)$$

where $\ell(t) = \log(1 + \exp(-t))$ is the logistic loss and $\lambda > 0$ is the regularization parameter. By contrasting the log-likelihood ratios of target responses against its own generations, the model increasingly aligns its policy with the ground-truth distribution through successive iterations.

[2] https://huggingface.co/datasets/aisingapore/SEA-Instruct-2602

Method	Teacher	Off-Policy Token-Signal	On-Policy Token-Signal	Sequence-Signal	Verbalizer
SFT	×	✓	×	×	×
DFT (Wu et al., 2026)	×	✓	×	×	×
SPIN (Chen et al., 2024b)	×	×	×	✓	×
SDFT (Shenfeld et al., 2026)	Self	×	✓	×	English
SeqKD (Kim and Rush, 2016)	Larger	✓	×	×	×
GKD (Agarwal et al., 2024)	Larger	✓	✓	×	×
DuDi (Ours)	Larger	✓	✓	✓	Cross-lingual

Table 1: Comparison of training paradigms across different frameworks.

Self-Distillation. To mitigate the generalization of off-policy fine-tuning, several studies adopt a self-distillation paradigm (Yang et al., 2024; Zhang et al., 2024; Hübotter et al., 2026). Specifically, Shenfeld et al. (2026) introduced Self-Distillation Fine-Tuning (SDFT), which transforms off-policy signals into an on-policy paradigm by employing an exponential moving average student as the teacher $\pi_{\theta_T} \sim \text{EMA}(\pi_\theta)$. A central component of SDFT is an English verbalizer function, $z \sim v_{\text{en}}(x, y)$, which converts an input and ground-truth pair into a structured demonstration prompt for the teacher model (illustrated in Figure 3a). This verbalized demonstration is subsequently used to guide the student-generated response, $\tilde{y} \sim \pi_\theta(x)$. The optimization objective minimizes the divergence between the teacher, conditioned on the verbalized demonstration, and the student policy:

$$\mathcal{L}_{\text{SD}} = \mathbb{E}_{x\sim D,\tilde{y},z}\left[D\left(\pi_{\theta_T}(\tilde{y}|z)||\pi_\theta(\tilde{y}|x)\right)\right]. \quad (3)$$

By leveraging this temporary ensemble, SDFT regularizes the optimization path and improves generalization through token-level guidance.

Teacher Distillation. The teacher knowledge distillation (Hinton et al., 2015; Lin et al., 2020; Ko et al., 2024, 2025) paradigm leverages signals from a stronger teacher model (π_{θ_T}) to guide a student model (π_θ), typically a parameter-efficient counterpart. This approach abstractly optimizes two objectives that could utilize static ground-truths with stochastic explorations:

$$\begin{aligned}\mathcal{L}_{\text{TD}} =&(1-\lambda)\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[D(\pi_{\theta_T}(y|x)||\pi_\theta(y|x))\right] \\ &+\lambda\mathbb{E}_{x\sim\mathcal{D},\tilde{y}\sim\pi_\theta(x)}\left[D\left(\pi_{\theta_T}(\tilde{y}|x)||\pi_\theta(\tilde{y}|x)\right)\right],\end{aligned} \quad (4)$$

where D denotes a divergence function, and $\lambda \in [0, 1]$ balances the distillation signals from ground-truths and newly-generated responses. Here, $\tilde{y}$ is sampled from either the teacher or student policy. In off-policy KD, SeqKD (Kim and Rush, 2016) trains the student on teacher-generated sequences, i.e., $\tilde{y} \sim \pi_{\theta_T}(x)$. However, this approach often suffers from training-inference mismatches when the output sequences generated by the student at inference time deviate significantly from those encountered during training. To address this mismatch, Generalized Knowledge Distillation, GKD (Agarwal et al., 2024) introduces an on-policy mechanism. In the GKD framework, the student generates its own responses $\tilde{y} \sim \pi_\theta(x)$, while the teacher provides token-level signals for guiding the student output logits to correct their self-generated mistakes.

As summarized in Table 1, although these approaches have advanced fine-tuning methodologies, their multilingual extension remains insufficiently explored. Additionally, existing methods typically treat sequence-level and token-level supervision independently, limiting the complementarity of both learning signals within a unified framework.

3 DuDi

As illustrated in Figure 2, DuDi consists of three core components: a sequence-level signal (Section 3.1), a token-level signal (Section 3.2), and a cross-lingual verbalizer (Section 3.3). The framework enables SLMs to jointly leverage sequence-level and token-level supervision, while the cross-lingual verbalizer improves teacher-student knowledge transferability in multilingual settings.

3.1 Sequence Signal

To improve the efficiency of SLMs, we integrate a sequence-level objective inspired by the SPIN framework to ensure the student policy converges toward the ground-truth demonstration. This objective requires the model to differentiate its current policy from the ground-truth by maximizing the relative likelihood of ground-truths. Distinct from static iterations of SPIN, we sample responses $y' \sim \pi_\theta(x)$ in *real-time* to provide dynamic on-policy feedback with the objective in Equation 2.

3.2 Token Signal

Complementing token-level supervision, DuDi leverages both off- and on-policy objectives to enhance performance through knowledge distillation.

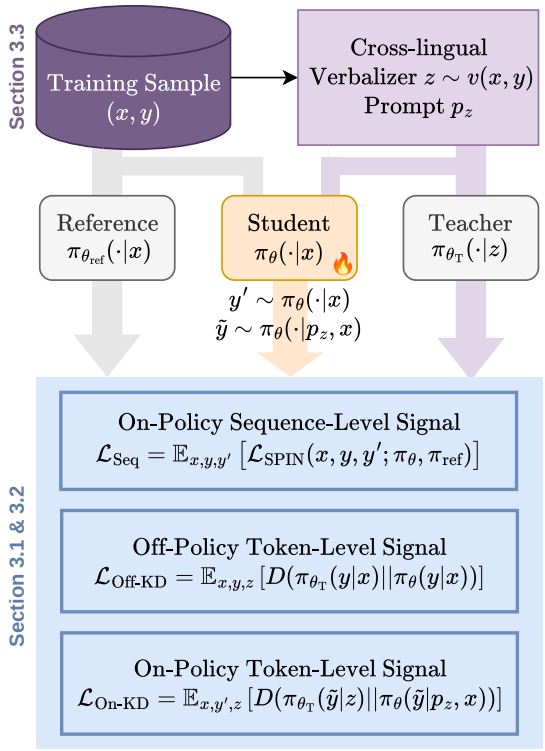


Figure 2: Overview of the DuDi framework, which integrates the sequence-level objective, token-level signals of off-policy, and on-policy knowledge distillation with teacher-guided cross-lingual verbalization.

Off-Policy KD. While on-policy distillation mitigates teacher-student mismatches (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024, 2025), relying solely on student-generated responses may reduce exposure to the ground-truth data distribution, potentially causing the learned policy to drift away from it. To address this problem, we adopt an off-policy distillation signal that leverages teacher-provided logits as guidance. This mechanism ensures that the supervisory signal remains strictly grounded in the ground-truth distribution. The corresponding objective can be represented as

$$\mathcal{L}_{\text{Off-KD}} = D\left(\pi_{\theta_T}(y|x)||\pi_\theta(y|x)\right). \quad (5)$$

On-policy KD. To enable teacher-guided refinement of the student-generated responses, we further adopt an on-policy knowledge distillation objective, where the student generates a response $\tilde{y} \sim \pi_\theta(x)$. The on-policy distillation objective is defined as

$$\mathcal{L}_{\text{On-KD}} = D\left(\pi_{\theta_T}(\tilde{y}|x)||\pi_\theta(\tilde{y}|x)\right). \quad (6)$$

3.3 Cross-Lingual Verbalizer

Optimizing on-policy learning with token-level supervision in multilingual settings requires a carefully designed framework for effective teacher-student knowledge transfer. To further improve the distillation process, inspired by English verbalizer (Shenfeld et al., 2026), we introduce a cross-lingual verbalizer. This component converts an input prompt x, a ground-truth y as demonstration, source language l, and target language l_z into a verbalized prompt $z = v(x, y, l, l_z)$ for the teacher, along with a corresponding prompt p_z for the student. The verbalizer prompt is the same language l as the sample native language. The prompt p_z instructs the student model to generate responses in the target language l_z, where l_z sampled uniformly from the set of training languages excluding l, with English included as an additional language. An example of the cross-lingual verbalizer prompt template is shown in Figure 3c. Furthermore, Figure 4 illustrates a Thai sample in which verbalized for generating responses in Vietnamese.

Consequently, with a cross-lingual verbalizer, the on-policy token-level distillation objective in Equation 6 could be modified with a cross-lingual verbalizer z for the teacher together with its associated prompt template p_z for the student. Based on this verbalized input, the student generates an additional response $\tilde{y} \sim \pi_\theta(p_z, x)$. The on-policy distillation objective with the cross-lingual verbalizer on the teacher is formally expressed as:

$$\mathcal{L}_{\text{On-KD}} = D\left(\pi_{\theta_T}(\tilde{y}|z)||\pi_\theta(\tilde{y}|p_z, x)\right). \quad (7)$$

The use of the cross-lingual verbalizer enables knowledge transfer across languages, thereby improving downstream performance in multilingual settings. The cross-lingual verbalizer configuration in DuDi is detailed in Section 4.1.

3.4 Unify Training Objective

The DuDi framework is optimized by integrating on-policy sequence-level alignment (Equation 2), off-policy token-level distillation (Equation 5), and on-policy token-level guidance via the multilingual verbalizer (Equation 7). Formally, the learning objective is abstractly represented as:

$$\begin{aligned}\mathcal{L}_{\text{DuDi}} &= \alpha\mathbb{E}_{(x,y)\sim\mathcal{D},y'\sim\pi_\theta(x)}\left[\mathcal{L}_{\text{SPIN}}(x, y, y'; \pi_\theta, \pi_{\theta_{\text{Ref}}})\right] \\ &+ (1-\lambda)\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}_{\text{Off-KD}}(x, y; \pi_\theta, \pi_{\theta_T})\right] \\ &+ \lambda\mathbb{E}_{x\sim\mathcal{D},\tilde{y}\sim\pi_\theta(p_z,x),z\sim v(x,y)}\left[\mathcal{L}_{\text{On-KD}}(x, \tilde{y}, z; \pi_\theta, \pi_{\theta_T})\right],\end{aligned} \quad (8)$$

where α denotes the weighting coefficient for sequence-level student policy optimization and λ denotes the weighting coefficient to balance token-level off-policy and on-policy distillation loss. This joint objective allows the student model to leverage both the sequence-level optimization towards

Teacher Prompt for English Verbalizer

{Prompt}

This is an example for a response to the question:

{Demonstration}

Now answer with a response of your own, including the thinking process.

(a) The English verbalizer from (Shenfeld et al., 2026)

Teacher Prompt for Multilingual Verbalizer

{Prompt}

นี่คือตัวอย่างคำตอบสำหรับคำถาม:
This is an example for a response to the question:

{Demonstration}

ตอนนี้ให้ตอบด้วยคำตอบของคุณเอง พร้อมทั้งแสดงกระบวนการคิดของคุณด้วย
Now answer with a response of your own, including the thinking process.

(b) our extended multilingual verbalizer

Teacher Prompt for Cross-Lingual Verbalizer

{Prompt}

นี่คือตัวอย่างคำตอบสำหรับคำถามในภาษา{Prompt Language}:
This is an example for a response to the question in {Prompt Language}:

{Demonstration}

ตอนนี้ให้ตอบด้วยคำตอบของคุณเองเป็นภาษา {Target Language} พร้อมทั้งแสดงกระบวนการคิดของคุณด้วย
Now answer with a response of your own, including the thinking process in {Target Language}

Student Prompt for Cross-Lingual Verbalizer

{Prompt}

ตอบด้วยคำตอบเป็นภาษา{Target Language} พร้อมทั้งแสดงกระบวนการคิดของคุณด้วย
Answer with a response, including the thinking process in {Target Language}.

(c) The proposed cross-lingual verbalizer, along with its corresponding student prompt example in Thai.

Figure 3: Comparison of verbalizer templates for the teacher prompt. The examples demonstrate the teacher prompt (z) and its corresponding student prompt (p_z) for a Thai training instance for a cross-lingual verbalizer.

Teacher Prompt

เดือนแรกของปีคือเดือนอะไร
What is the first month of the year?

นี่คือตัวอย่างคำตอบสำหรับคำถามในภาษาไทย:
This is an example for a response to the question in Thai:

เดือนแรกของปีคือเดือนมกราคม
The first month of the year is January.

ตอนนี้ให้ตอบด้วยคำตอบของคุณเองเป็นภาษาเวียดนามพร้อมทั้งแสดงกระบวนการคิดของคุณด้วย
Now answer with a response of your own, including the thinking process in Vietnamese.

Student Prompt

เดือนแรกของปีคือเดือนอะไร
What is the first month of the year?

ตอบด้วยคำตอบเป็นภาษาเวียดนาม พร้อมทั้งแสดงกระบวนการคิดของคุณด้วย
Answer with a response, including the thinking process in Vietnamese.

Figure 4: Example of cross-lingual verbalized teacher and student prompts. In this example, the original sample is in Thai, where the prompt is expressed in Thai, while the target response is generated in Vietnamese. For this sample, the target response language is uniformly sampled from seven languages: English, Indonesian, Vietnamese, Tamil, Tagalog, Malay, and Burmese, excluding the sample's original language (Thai).

the ground-truth data distribution and fine-grained token-level supervision from the teacher.

3.5 Differentiation from Previous Work

DuDi jointly leverages sequence-level and token-level signals, with two key distinctions from prior approaches. We integrate a self-play mechanism (Chen et al., 2024b) as a sequence-level signal, transitioning from offline to online generation to reflect the student's evolving policy. Furthermore, we extend the on-policy distillation objective (Agarwal et al., 2024; Shenfeld et al., 2026) by introducing cross-lingual prompting for student rollouts, paired with a teacher equipped with a cross-lingual verbalizer, thereby improving knowledge transferability between teacher and student.

Compared to prior methods summarized in Table 1, this design enables the integration of dual-signal supervision with a tailored verbalizer, achieving state-of-the-art performance in multilingual settings (Section 5). Ablation results in Section 6.1 further show that these components are not independently effective but require joint integration in multilingual scenarios. Moreover, Section 7 demonstrates that the cross-lingual verbalizer improves on-policy teacher-student knowledge transferability, consistent with our hypothesis.

4 Experimental Setup

4.1 Setup

Models and Datasets. We center our study on Qwen2.5 (Qwen et al., 2025), using Qwen2.5-3B-Instruct as the teacher model and Qwen2.5-0.5B and 1.5B as student models. To cover other families, we also evaluate Qwen3 (4B→0.6B) and Llama3.2 (3B→1B). All student models are initialized from base pretrained checkpoints, whereas the corresponding teacher models use instruction-tuned variants. For the training dataset, we use SEA-Instruct, which covers seven SEA languages: Indonesian, Vietnamese, Thai, Tamil, Tagalog, Malay, and Burmese. The dataset contains open-source prompts, each paired with a synthetic response and quality estimate. We sample 4,000 high-quality examples per language, as labeled by the original dataset, yielding 28,000 samples. Random sampling constraints preserve the distribution of domains, task types, and prompt complexity.

Framework Setup. The cross-lingual verbalizer in DuDi incorporates 7 SEA languages and English. During on-policy training, the target response lan-

Method	Indonesian	Vietnamese	Thai	Tamil	Tagalog	Malay	Burmese	Average
Qwen2.5-3B-Instruct (Teacher)	42.0	38.7	32.3	9.8	24.6	40.0	6.5	27.7
Qwen2.5-0.5B (Student)								
SFT	10.6	11.8	10.8	4.9	5.6	12.0	4.4	8.6
DFT	9.3	11.6	10.2	**10.0**	6.1	12.4	**8.3**	9.7
SPIN	**12.6**	11.4	10.0	5.3	**8.0**	14.7	4.8	9.5
SDFT	4.2	4.6	4.7	2.3	3.3	5.1	1.3	3.6
SeqKD	7.1	9.1	7.7	4.2	3.4	8.1	2.9	6.1
GKD	11.7	13.2	10.9	4.9	4.8	13.6	3.7	9.0
DuDi (ours)	11.7$^{\diamond}$	**14.4**$^{\diamond\ddagger}$	**12.8**$^{\diamond\ddagger}$	6.5	6.1	**14.8**$^{\diamond}$	4.6	**10.1**$^{\diamond}$

Table 2: Downstream evaluation across seven Southeast Asian languages. "$\diamond$" denotes a statistically significant improvement in each language, where DuDi outperforms DFT, the second-best overall performance. "‡" denotes a statistically significant improvement of DuDi over the second-best baseline within the corresponding language.

Method	Indonesian	Vietnamese	Thai	Tamil	Tagalog	Malay	Burmese	Average
Qwen2.5-3B-Instruct (Teacher)	42.0	38.7	32.3	9.8	24.6	40.0	6.5	27.7
Qwen2.5-1.5B (Student)								
SFT	21.4	26.0	**23.8**	10.0	16.6	26.1	**6.2**	18.6
DFT	22.1	17.8	16.6	9.6	13.6	22.8	**6.2**	15.6
SPIN	21.0	25.3	20.2	**10.1**	**19.2**	26.7	5.3	18.3
GKD	**28.6**	28.8	19.9	7.3	15.4	29.1	4.7	19.1
DuDi	27.9	**30.3**$^{\diamond}$	19.8	8.5$^{\diamond}$	19.0$^{\diamond}$	**30.7**$^{\diamond}$	4.8	**20.1**$^{\diamond}$
Qwen3-4B (Teacher)	54.1	52.7	50.7	43.1	45.9	53.0	20.0	45.6
Qwen3-0.6B-Base (Student)								
SFT	15.6	19.5	17.3	10.1	10.7	17.9	6.4	13.9
DFT	18.4	18.1	17.3	**14.0**	14.0	20.1	7.1	15.6
SPIN	14.4	20.5	17.7	10.4	14.0	16.1	6.8	14.3
GKD	20.6	27.0	21.9	10.9	15.1	23.9	6.5	18.0
DuDi	**24.2**$^{\diamond}$	**30.4**$^{\diamond}$	**23.4**$^{\diamond}$	13.2$^{\diamond}$	**17.6**$^{\diamond}$	**28.1**$^{\diamond}$	**8.4**$^{\diamond}$	**20.8**$^{\diamond}$
Llama3.2-3B-Instruct (Teacher)	32.4	30.3	38.2	16.7	28.3	42.8	5.0	27.7
Llama3.2-1B (Student)								
SFT	4.7	5.3	3.6	3.9	3.3	6.5	**5.7**	4.7
DFT	0.4	0.4	0.9	0.4	0.3	1.1	0.4	0.6
SPIN	2.9	3.6	3.8	3.1	2.9	7.4	2.4	3.7
GKD	11.8	**17.7**	**11.4**	5.6	7.3	16.3	4.2	10.6
DuDi	**14.4**$^{\diamond}$	16.5	10.4	**6.6**$^{\diamond}$	**8.1**	**18.5**$^{\diamond}$	5.2$^{\diamond}$	**11.4**$^{\diamond}$

Table 3: Results across different teacher-student model configurations. "$\diamond$" denotes a statistically significant improvement in each language, where DuDi outperforms GKD, the second-best overall performing framework.

guage for the student is uniformly sampled from this set, excluding the original language of the training sample. In addition, we adopt a two-stage training framework in which the base model is first SFT on the SEA-Instruct dataset (cold-start SFT), after which DuDi training is initialized from the resulting SFT checkpoints. For the sequence-level objective, the reference policy π_{Ref} is set as the cold-start SFT checkpoint. The importance of cold-start SFT initialization is further discussed in Section 7.2. Additional details regarding the training configurations are provided in Appendix B.

4.2 Evaluation

We evaluate on SEA-HELM (Susanto et al., 2025), which covers multiple Southeast Asian languages and diverse tasks. The languages are Indonesian, Vietnamese, Thai, Tamil, Tagalog, Malay, and Burmese. SEA-HELM includes natural language understanding (NLU), natural language generation (NLG), natural language reasoning (NLR), safety, linguistic diagnostics, instruction following, and Southeast Asian knowledge. All results are averaged over four seeds. We also use Almost Stochastic Order (ASO)[3] (Del Barrio et al., 2018; Dror et al., 2019) to test statistical significance between DuDi and the second-best performing framework.

4.3 Competitive Methods

We compare DuDi with all comparative methods for SLMs, as we discussed in Section 2. For off-policy fine-tuning, we use SFT and DFT (Wu et al., 2026) as standard fine-tuning baselines. For iterative fine-tuning, we adopt SPIN (Chen et al., 2024b) as a representative sequence-level optimization method. For self-distillation strategies, we include SDFT (Shenfeld et al., 2026), which uses temporary prefix prompting for self-guided on-policy refinement. For knowledge distillation, we evaluate SeqKD (Kim and Rush, 2016) and GKD (Agarwal et al., 2024). Furthermore, all methods, except off-policy fine-tuning, are initialized from a cold-start SFT checkpoint, following the same setup as DuDi. Additional implementation details for all

[3] Using the implementation from Ulmer et al. (2022)

competitive methods are provided in Appendix B.

5 Main Results

DuDi outperforms all methods. Overall, DuDi achieves the strongest performance among all frameworks. As shown in Table 2, DuDi obtains the highest average score of 10.1 across seven SEA languages, statistically significantly outperforming the strongest baseline, DFT, by 0.4 points and SPIN by 0.6 points. DuDi ranks first in Vietnamese, Thai, and Malay, and second in Indonesian, Tamil, and Tagalog, demonstrating consistent improvements across both high- and mid-resource SEA languages. The only exception is Burmese, where the gain is limited, likely due to the smaller teacher-student gap and teacher performance in this language.

Comparison with off-policy fine-tuning. DFT is a strong off-policy baseline, achieving the second-best overall score, only 0.4 points below DuDi. It outperforms the larger model in Tamil and Burmese, suggesting that direct fine-tuning is beneficial when the teacher is unreliable. However, DFT does not consistently surpass SFT across languages and is limited as a cold-start initialization for further distillation (Reported in Appendix 7.3).

Contrast with other methods. SPIN emerges as the third strongest approach, suggesting that self-play remains effective. GKD is also a strong KD baseline, achieving 9.0 on average, while SeqKD performs moderately but remains below both GKD and SPIN. In contrast, SDFT substantially underperforms despite its self-distillation design; its English-based verbalizer limits the teacher-student transferability. Further analyses regarding verbalizer configurations and teacher–student knowledge transferability are detailed in Sections 6.2 and 7.1.

Robustness across model families. DuDi also generalizes consistently across different teacher–student configurations. As shown in Table 3, DuDi achieves the best average performance for all three student model families. Compared with GKD, the strongest prior baseline, DuDi improves average performance by 5.2%, 15.6%, and 7.5% on Qwen2.5-1.5B, Qwen3-0.6B-Base, and Llama3.2-1B, respectively. These gains demonstrate that DuDi remains effective across variations in model scale, version, and architecture family.

Stability over competing methods. Among prior methods, GKD is the most competitive baseline and consistently outperforms SPIN and DFT across model families. However, SPIN exhibits unstable behavior, it provides little improvement over SFT on Qwen2.5-1.5B and Llama3.2-1B, and only modest gains on Qwen3-0.6B-Base. DFT is even less stable, failing to outperform SFT and collapsing on Llama3.2-1B. In contrast, DuDi consistently improves over these baselines, highlighting its robustness over comparative fine-tuning approaches.

6 Ablation Studies

6.1 Critical Components

DuDi comprises sequence-level supervision, off-policy KD, and on-policy KD with a cross-lingual verbalizer. We perform an ablation study by removing one component at a time and measuring its impact on overall performance. Additionally, we examine whether excluding English from the set of target responses leads to performance degradation.

Table 4 reports the ablation results. Eliminating the sequence-level loss yields only a minor performance decline, though it still helps align the student with the ground-truth trajectory. Excluding off-policy KD causes the largest performance drop, highlighting its critical role in providing ground-truth supervision and guiding SLMs toward the target trajectory. Disabling on-policy KD results in the second-largest degradation, as the teacher can no longer refine student-generated responses.

For the verbalizer component, their absence noticeably reduces performance by 5%. Lastly, removing English generation from the cross-lingual verbalizer degrades performance, suggesting the inclusion of the English language facilitates better cross-lingual transfer. Overall, all components are complementary and jointly contribute to DuDi's optimal performance.

Method	SEA-HELM	Δ	% Difference
DuDi	10.1	-	-
w/o sequence	9.7	-0.4	-4.7%
w/o off-policy KD	7.6	-2.5	-24.6%
w/o on-policy KD	9.5	-0.6	-6.6%
w/o verbalizer	9.6	-0.5	-5.0%
w/o English	9.8	-0.3	-3.3%

Table 4: Ablation results of DuDi's components.

6.2 Design Choices

Verbalizer Modes. For each training sample (x, y, l), there are three verbalizer modes: **(i) English**, where verbalizer prompt template are English, following Shenfeld et al. (2026) (Figure 3a); **(ii) Multilingual**, where l_z is the sample language l (Figure 3b); and **(iii) Cross-lingual**, where verbalizer prompt template language matches the sam-

ple's native language l, while l_z is sampled uniformly from the set consisting of English and all training languages excluding l (Figure 3c).

As described in Section 3, DuDi uses a cross-lingual verbalizer. We compare three variants: *English*, *Multilingual*, and *Mix*, which uses 50% multilingual mode for native-language learning and 50% cross-lingual mode for knowledge transfer.

Table 5 shows that our cross-lingual verbalizer is the only variant outperforming the no-verbalizer baseline. The multilingual verbalizer performs worst overall, while the English-only verbalizer slightly surpasses it despite lacking language-specific supervision. The mixed verbalizer further improves upon the multilingual setting, underscoring the importance of cross-lingual verbalization. Overall, these results support the effectiveness of the proposed cross-lingual verbalizer in facilitating teacher-student knowledge transfer.

KD Objective. We compare DuDi's reverse KL objective against forward KL and Jensen-Shannon Divergence (JSD), which interpolates between the two. Table 5 shows that reverse KL achieves the best SEA-HELM score (10.1), substantially outperforming JSD and forward KL, corresponding to relative drops of 23.8% and 37.8%, respectively. The weaker performance of forward KL may stem from teacher-student mismatch and overestimation of low-probability regions in the teacher distribution (Gu et al., 2024). In contrast, reverse KL yields more stable and effective knowledge transfer.

Method	SEA-HELM	Δ	% Difference
DuDi	10.1	-	-
Verbalizer Mode			
No verbalizer	9.6	-0.5	-5.0%
English	8.2	-1.9	-18.8%
Multilingual	7.9	-2.2	-21.8%
Mix	9.0	-1.1	-10.9%
KD Objective			
Forward-KL	6.3	-3.8	-37.8%
JSD	7.7	-2.4	-23.8%

Table 5: Evaluation of performance under different verbalizers and knowledge distillation objectives.

7 Analyses

To further investigate the properties of DuDi, we conduct two analyses. In Section 7.1, we examine how cross-lingual verbalizers facilitate the transferability between teacher and student. Subsequently, in Section 7.2, we analyze the role of cold-start fine-tuning and demonstrate its importance for effective multilingual knowledge distillation. Finally, in Section 7.3, we examine the limitations of using DFT checkpoints as the cold-start initialization.

7.1 Why DuDi Verbalizer is Optimal?

To better understand the effectiveness of the proposed cross-lingual verbalizer, we employ the overlap ratio analysis (Li et al., 2026), which evaluates the top-k overlap between student and teacher output logits during on-policy token-level distillation. This metric quantifies the degree of agreement between the two distributions, thereby facilitating the on-policy gradient signal to the student model.

As shown in Figure 5, the proposed cross-lingual verbalizer achieves the highest overlap ratio throughout training, indicating that cross-lingual rollouts provide informative supervision signals for on-policy distillation. For the remaining variants, the overlap ratio ranking follows the performance trend in Table 5, with no verbalizer performing second best, followed by the mix, English, and multilingual verbalizers. Notably, the results suggest that multilingual verbalizers can increase teacher-student mismatches, highlighting the challenges of verbalizer design in multilingual settings.

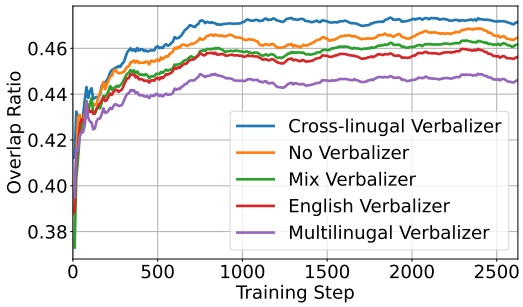


Figure 5: Overlap ratio between teacher and student logits during on-policy rollouts across different verbalizers. The mix verbalizer denotes a uniform random combination of multilingual and cross-lingual verbalizers.

7.2 Why is Cold-Start SFT Important?

DuDi is trained under a cold-start SFT setting, where models are initialized from SFT checkpoints using ground-truth responses as supervision. In parallel, Li et al. (2026); Zhu et al. (2026) adopts a related initialization strategy based on teacher-generated responses rather than ground-truth data, demonstrating that SFT initialization improves the effectiveness of on-policy distillation. As shown in Table 6, training DuDi directly from the base model without cold-start initialization results in the weakest performance. Although initialization with teacher-generated outputs yields moderate improvements, it remains lower than initialization using the original ground-truth training data. These findings indicate that cold-start initialization also plays an important role in effective multilingual knowledge

distillation for SLMs. Furthermore, similar trends are consistently observed across all comparative methods, as reported in Appendix G.

Method	SEA-HELM	Δ	% Difference
DuDi	10.1	-	-
Teacher generated	9.1	-1.0	-10.2%
No cold-start	8.6	-1.5	-15.1%

Table 6: Evaluation of downstream performance under different cold-start setting initialization.

7.3 Limitations of DFT as Cold-Start

As shown in Table 2, DFT outperforms SFT on downstream tasks. Motivated by this, we examine whether DFT provides a better initialization checkpoint than SFT as a cold-start. We compare off-policy fine-tuning (cold-start) initialized from SFT and DFT checkpoints across three methods: SPIN, GKD, and DuDi. As shown in Table 7, using a DFT checkpoint as the cold-start degrades performance across all three frameworks. Thus, while DFT is a strong standalone baseline, it is less compatible with subsequent training and yields limited additive gains.

Method	id	vi	th	ta	tl	ms	my	Avg. (Δ)
SFT	10.9	12.1	11.2	5.1	5.8	12.1	4.3	8.8
DFT	9.3	11.6	10.2	10.0	6.1	12.4	8.3	9.7
SFT → SPIN	12.6	11.4	10.0	5.3	8.0	14.7	4.8	9.5 (+0.7)
DFT → SPIN	11.5	8.5	9.1	9.3	5.0	12.4	8.2	9.1 (-0.6)
SFT → GKD	11.7	13.2	10.9	4.9	4.8	13.6	3.7	9.0 (+0.2)
DFT → GKD	6.6	7.8	7.9	3.9	4.5	7.2	3.6	5.9 (-3.8)
SFT → DuDi	12.8	14.6	11.5	5.4	6.4	14.6	3.4	10.1 (+1.3)
DFT → DuDi	7.9	10.8	8.9	6.4	4.9	9.1	4.1	7.4 (-2.3)

Table 7: Comparison of results using alternative SFT- or DFT-based checkpoints as a cold-start across different training frameworks.

To better understand the underlying mechanism between SFT and DFT as the cold-start, we compare the top-k overlap ratio of SFT and DFT as the cold-start initializations for DuDi. As illustrated in Figure 6, the cold-start DFT consistently exhibits a substantially lower overlap ratio than SFT throughout training across methods. According to Li et al. (2026), a low overlap ratio is indicative of degraded performance, suggesting that the cold-start DFT student assigns probability mass to a token set that is largely disjoint from that of the teacher. These findings indicate that DFT alters the token distribution too aggressively, biasing the model toward different token preferences.

8 Conclusion

In this work, we introduced DuDi, a general-purposed multilingual distillation framework for

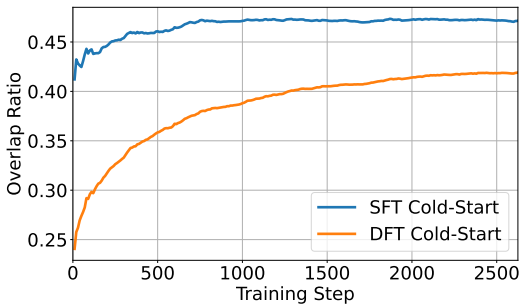


Figure 6: Overlap ratio between student and teacher model logits for on-policy token-level distillation of SFT vs DFT as a cold-start.

SLMs that integrates sequence-level and token-level supervisions, along with a cross-lingual verbalization mechanism. Extensive experiments across diverse model families and parameter scales demonstrate that DuDi consistently achieves the highest average SEA-HELM performance, substantially outperforming strong fine-tuning and distillation baselines. Ablation studies further indicate that jointly optimizing sequence-level and token-level objectives, together with the proposed verbalizer design, yields complementary benefits. Our analysis also shows that the cross-lingual verbalizer improves teacher-student knowledge transferability. Overall, DuDi offers an effective fine-tuning framework for SLMs in a multilingual environment. In addition, for the open-research purpose, we will release all artifacts in this paper, including training code, datasets, and models.

Limitations

The experimental setup of this study primarily focuses on Southeast Asian (SEA) languages, with models trained on SEA-Instruct and evaluated using SEA-HELM (Susanto et al., 2025), which covers seven SEA languages across a diverse set of tasks. Consequently, the findings related to DuDi may not generalize to tasks beyond those included in the current evaluation framework. Nevertheless, SEA-HELM remains a gold-standard benchmark for the comprehensive evaluation of language model capabilities in SEA languages. Future work will focus on extending both the training data and evaluation benchmarks to encompass a wider range of contemporary language modeling tasks.

Another limitation of this work concerns the availability of suitable teacher models. In particular, the teacher model must be more capable, typically larger in scale than the student model while sharing the same output vocabulary space. This constraint arises because the divergence func-

tion employed in knowledge distillation requires aligned teacher–student logit dimensions.

Acknowledgments

This project is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. We thank Trevor Cohn for his helpful feedback, and Ngee Chia Tai and Raymond Ng for their support and valuable comments.

References

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

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

Xiao Chen, Changyi Ma, Wenqi Fan, Zhaoxiang Zhang, and Li Qing. 2025. C2KD: Cross-layer and cross-head knowledge distillation for small language model-based recommendation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17827–17838, Vienna, Austria. Association for Computational Linguistics.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. Self-play fine-tuning convertsweak language models to strong language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. 2018. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. Association for Computational Linguistics.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Xinrong Zhang, Zhen Leng Thai, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, and 5 others. 2024. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*.

Jonas Hübotter, Frederike Lübeck, Lejs Deen Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. 2026. Reinforcement learning via self-distillation. In *The 1st Workshop on Scaling Post-training for LLMs*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. 2025. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In *Forty-second International Conference on Machine Learning*.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. 2024. DistiLLM: Towards streamlined distillation for large language models. In *Forty-first International Conference on Machine Learning*.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang Gao, Wenkai Yang, Zhiyuan Liu, and 1 others. 2026. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. 2020. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6121–6133, Online. Association for Computational Linguistics.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. 2024. MobileLLM: Optimizing sub-billion parameter language models for on-device use cases. In *Forty-first International Conference on Machine Learning*.

Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir Parmar, Sasidhar Kunapuli, Joe Barrow, Junda Wu,

Ashish Singh, Yu Wang, Jiuxiang Gu, Franck Dernoncourt, Nesreen K. Ahmed, Nedim Lipka, Ruiyi Zhang, Xiang Chen, and 9 others. 2024. A survey of small language models. *Preprint*, arXiv:2410.20011.

Patomporn Payoungkhamdee, Peerat Limkonchotiwat, Jinheon Baek, Potsawee Manakul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. An empirical study of multilingual reasoning distillation for question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7739–7751, Miami, Florida, USA. Association for Computational Linguistics.

Thang M. Pham, Phat T. Nguyen, Seunghyun Yoon, Viet Dac Lai, Franck Dernoncourt, and Trung Bui. 2025. SlimLM: An efficient small language model for on-device document assistance. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 436–447, Vienna, Austria. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1):101118.

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, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. 2026. Self-distillation enables continual learning. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359.

Shreyas Subramanian, Vikram Elango, and Mecit Gungor. 2025. Small language models (slms) can still pack a punch: A survey. *Preprint*, arXiv:2501.05465.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xianbin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. SEA-HELM: Southeast Asian holistic evaluation of language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12308–12336, Vienna, Austria. Association for Computational Linguistics.

Gerald Tesauro. 1995. Temporal difference learning and td-gammon. *Commun. ACM*, 38(3):58–68.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022. deep-significance: Easy and meaningful signifcance testing in the age of neural networks. In *ML Evaluation Standards Workshop at the Tenth International Conference on Learning Representations*.

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

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. 2024. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *Preprint*, arXiv:2411.03350.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. 2026. On the generalization of SFT: A reinforcement learning perspective with reward rectification. In *The Fourteenth International Conference on Learning Representations*.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. 2025. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *The Thirteenth International Conference on Learning Representations*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043, Bangkok, Thailand. Association for Computational Linguistics.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. LangBridge: Multilingual reasoning without multilingual supervision. In *Proceedings of the 62nd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7502–7522, Bangkok, Thailand. Association for Computational Linguistics.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11189–11204, Bangkok, Thailand. Association for Computational Linguistics.

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. 2026. When less language is more: Language-reasoning disentanglement makes LLMs better multilingual reasoners. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. 2026. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *Preprint*, arXiv:2605.11182.

A Related Work

A.1 Knowledge Distillation

Knowledge distillation (KD) (Hinton et al., 2015) is a training paradigm that transfers knowledge from a larger teacher model to a smaller student model, enabling compact models to benefit from the capabilities of stronger models. Early sequence-level distillation methods, including SeqKD (Kim and Rush, 2016) and ImitKD (Lin et al., 2020), demonstrate that teacher-generated outputs provide effective supervision signals for student training. Subsequent work has focused on improving the stability and efficiency of distillation. For example, MiniLLM (Gu et al., 2024) introduces a policy-gradient-based framework that mitigates the high variance commonly encountered in reinforcement learning optimization. Similarly, Xu et al. (2025) combines static datasets with on-policy distillation through speculative decoding for synthetic data generation. Among recent approaches, DistiLLM (Ko et al., 2025) achieves strong performance and training efficiency by employing symmetric KL divergence together with an adaptive off-policy distillation strategy.

More recent studies extend distillation to on-policy settings, where student models learn directly from their own generated responses. For instance, Agarwal et al. (2024) proposes on-policy optimization objectives based on reverse KL divergence and Jensen-Shannon divergence (JSD). Beyond conventional teacher distillation paradigms, self-distillation methods exploit model-generated responses as supervision signals to iteratively improve reasoning and downstream capabilities. Specifically, offline self-distillation methods (Yang et al., 2024) utilize self-generated responses from a seed model to better align the model with its own output distribution, whereas online variants (Shenfeld et al., 2026) employ in-context learning to acquire new capabilities while retaining the original competencies of the base model, alongside related reinforcement learning formulations proposed by Hübotter et al. (2026). In parallel, sequence-level optimization methods such as self-play approaches like SPIN (Chen et al., 2024b), and DistiLLM-2 (Ko et al., 2025) leverage contrastive objectives and trajectory-level regularization to enhance sample efficiency in reasoning-focused tasks.

A.2 Multilingual Distillation

Recent studies have primarily focused on constructing high-quality datasets for training smaller models on reasoning tasks. Payoungkhamdee et al. (2024) propose a distillation framework that transfers teacher capabilities through response generation, leveraging both positive and negative rationales to fine-tune smaller models for question-answering tasks. From a data-centric perspective, (Zhang et al., 2024) present self-distillation from transferring rich-resources to low-resources and MathOctopus (Chen et al., 2024a) translates mathematical training data into target languages to improve multilingual mathematical reasoning performance. On the architectural side, Yoon et al. (2024) introduce a multilingual encoder integrated with reasoning-capable LLMs for solving multilingual mathematics problems. In addition, Zhao et al. (2024) investigate the disentanglement of language and reasoning by identifying and exploiting language-specific neurons, thereby enhancing multilingual capabilities. Similarly, Zhao et al. (2026) propose causal intervention methods to improve downstream multilingual reasoning performance. Despite these advances, existing knowledge distillation frameworks remain largely centered on English settings and single task-specific setting, while general-purpose training frameworks for multilingual LLMs are still underexplored.

B Training Configuration

We fine-tune all models and all methods with a learning rate of 2e-5, batch size of 32, and 3 training epochs. The maximum sequence length is set to 1024 tokens, while on-policy rollouts are limited to a maximum of 256 generated tokens. For off-policy fine-tuning, we apply loss only on assistant tokens to better align with the next-token prediction objective of language models. In the on-policy rollout, following Agarwal et al. (2024), we enable stochastic exploration by setting[4] `do_sample=True` and `top-k=0`, with temperature 0.9 for all approaches. All knowledge distillation objectives use a temperature of 1.0. For hyperparameters in DuDi, we use fixed values of $\lambda = 0.5$ and $\alpha = 0.1$ across all experiments without per-model tuning.

For other methods, we evaluate SeqKD (Kim and Rush, 2016) with $\lambda = 1.0$, corresponding to training solely on teacher-generated responses, and

[4] https://huggingface.co/docs/transformers/main_classes/text_generation

GKD (Agarwal et al., 2024) using its default setting of $\lambda = 0.5$. Both SeqKD and GKD use JSD as the objective. We implement DuDi and all comparative methods using the TRL trainer framework (von Werra et al., 2020), except for SDFT (Shenfeld et al., 2026), where we adopt the original codebase provided by the authors.

C Computing Resources

We trained small models on 8× NVIDIA H200 (140GB) GPUs, completing fine-tuning within approximately 7 hours for SDFT, SeqKD, SPIN, GKD, and DuDi, and within an hour for SFT and DFT. For evaluation, we used 2× NVIDIA H200 (140GB) GPUs, completing SEA-HELM within 1 hour. In total, our experiments required approximately 1,704 GPU hours.

D Top-K Overlap Analysis Details

For visualization purposes, the top-k overlap analysis presented in this paper is smoothed using an exponential moving average with a smoothing factor of 0.99.[5]

E GKD with Reversed-KL

To further examine the effect of the divergence objective, we evaluate an RKL-based variant of GKD on Qwen3-0.6B, with the results reported in Table 8. GKD with RKL substantially underperforms the original GKD formulation with JSD, achieving an average score of 9.0 compared with 18.0. This result suggests that the performance improvements of DuDi cannot be explained solely by the choice of the f-divergence function.

Method	id	vi	th	ta	tl	ms	my	Avg.
GKD (JSD)	20.6	27.0	21.9	10.9	15.1	23.9	6.5	18.0
GKD (RKL)	11.5	12.7	9.8	5.2	4.2	16.2	3.5	9.0
DuDi	24.2	30.4	23.4	13.2	17.6	28.1	8.4	20.8

Table 8: Comparison of GKD with JSD and RKL objectives on Qwen3-0.6B.

F Computational Overhead

Similar to GKD, DuDi generates an additional cross-lingual response from the student model in addition to the standard student-generated response. Table 9 reports the relative wall-clock time and memory consumption of different training approaches. Here, T and S denote the memory required for the teacher and student model parameters, respectively. DuDi requires two rollouts and consequently incurs higher computational and memory costs than GKD.

Approach	Wall Clock Time (1×SFT)	Memory Cost	# Rollouts
SFT	1.0	S	0
DFT	1.2	S	0
SPIN	45.4	$2S$	1
GKD	39.4	$T + S$	1
DuDi	77.0	$T + 2S$	2

Table 9: Relative computational and memory costs of different training approaches.

G Additional Results

Task-Level Performance. Table 10 shows downstream evaluation for Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen3-0.6B-Base, and Llama-3.2-1 on multiple tasks in SEA-HELM, showing capability of each training methods. DuDi consistency achieves either the best or second-best performance across most tasks and model families. Notably, several tasks, Natural Language Reasoning (NLR), Safety, Linguistic Diagnostics (LD), and Knowledge, exhibit near-zero to very low absolute performance for Qwen2.5-0.5B and Llama-3.2-1B (often below 3 points), despite the teacher demonstrating some capability on these tasks. In contrast, larger or more capable base models such as Qwen2.5-1.5B and Qwen3-0.6B-Base retain non-trivial performance on these tasks. We hypothesize that this gap arises from limited coverage of these task distributions in the training data, resulting in weak teacher–student transfer and indicating that such tasks may require stronger base model capabilities rather than distillation alone.

Cold-Start SFT. Consistent with our findings in Section 7.2 and prior studies (Li et al., 2026; Zhu et al., 2026), Table 11 highlights the importance of cold-start SFT in knowledge distillation. The performance of both self-distillation methods, such as SDFT, and teacher-distillation approaches (SeqKD and GKD) declines when the student model is not initialized with cold-start SFT.

[5] https://docs.wandb.ai/models/app/features/panels/line-plot/smoothing.

Method	NLU	NLG	NLR	Safety	LD	IF	Knowledge	Average
Qwen2.5-3B-Instruct (Teacher)	32.5	34.8	16.4	19.8	13.5	47.2	15.1	27.7
Qwen2.5-0.5B (Student)								
SFT	7.2	12.5	0.4	0.2	0.0	28.8	2.2	8.6
DFT	7.3	9.5	0.2	0.0	0.4	**38.5**	0.9	9.7
SPIN	7.5	13.9	**1.1**	**0.4**	**0.5**	28.6	**5.2**	9.5
SDFT	1.1	3.8	0.0	0.0	0.0	16.6	0.4	3.6
SeqKD	3.2	9.5	0.0	0.0	0.0	22.4	1.3	6.1
GKD	8.1	15.6	0.3	0.3	0.0	28.1	1.4	9.0
DuDi (ours)	**9.0**	**15.8**	0.6	0.1	0.2	33.6	1.4	**10.1**
Qwen2.5-3B-Instruct (Teacher)	32.5	34.8	16.4	19.8	13.5	47.2	15.1	27.7
Qwen2.5-1.5B (Student)								
SFT	22.0	**26.3**	6.2	3.6	0.6	40.7	13.3	18.6
DFT	17.4	9.9	**11.9**	4.5	**4.1**	35.2	**15.4**	15.6
SPIN	20.4	23.7	8.0	5.1	0.4	38.8	14.6	18.3
GKD	24.9	22.0	7.8	7.9	2.3	41.3	10.5	19.1
DuDi	**26.5**	23.7	6.1	**8.8**	2.5	**44.1**	10.2	**20.1**
Qwen3-4B (Teacher)	58.6	25.2	47.3	37.9	36.2	68.3	33.1	45.6
Qwen3-0.6B-Base (Student)								
SFT	16.6	22.8	4.5	0.7	0.0	35.6	2.2	13.9
DFT	20.7	8.0	7.6	2.0	1.2	40.6	**12.6**	15.6
SPIN	20.0	18.6	5.6	0.9	0.2	36.3	2.3	14.3
GKD	24.6	26.2	8.1	3.0	0.0	41.5	1.9	18.0
DuDi	**28.5**	**26.3**	**14.0**	**3.7**	**2.8**	**44.7**	5.5	**20.8**
Llama-3.2-3B-Instruct (Teacher)	34.2	44.2	7.9	15.8	1.0	47.2	16.0	27.7
Llama-3.2-1B (Student)								
SFT	1.5	6.2	0.0	0.0	0.0	20.0	0.3	4.7
DFT	0.0	3.3	0.0	0.0	0.0	0.0	0.0	0.6
SPIN	0.0	21.7	0.0	0.0	0.0	0.0	0.0	3.7
GKD	9.3	23.1	0.4	0.5	0.0	28.1	**2.2**	10.6
DuDi	**9.7**	**24.0**	**0.4**	**2.9**	**0.4**	**30.0**	1.2	**11.4**

Table 10: Task evaluations for Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen3-0.6B-Base, and Llama-3.2-1B.

Method	Indonesian	Vietnamese	Thai	Tamil	Tagalog	Malay	Burmese	Average
SDFT	4.2	4.6	4.7	2.3	3.3	5.1	1.3	3.6
SDFT w/o SFT	0.7	0.7	1.8	1.1	0.6	0.9	1.1	1.0
SeqKD	7.1	9.1	7.7	4.2	3.4	8.1	2.9	6.1
SeqKD w/o SFT	4.4	7.2	6.1	3.6	2.1	6.1	3.2	4.7
GKD	11.7	13.2	10.9	4.9	4.8	13.6	3.7	9.0
GKD w/o SFT	6.5	7.5	7.2	3.9	3.5	7.7	3.3	5.6
DuDi	11.7	14.4	12.8	6.5	6.1	14.8	4.6	10.1
DuDi w/o SFT	9.4	13.7	9.7	7.2	5.3	10.8	4.1	8.6

Table 11: Results of comparative knowledge distillation methods without cold-start SFT.