arXiv:2604.11290v3 [cs.CL] 28 Aug 2026

Polyglot Teachers: Evaluating Language Models for Multilingual Synthetic Data Generation

Lester James V. Miranda **Ivan Vulić** **Anna Korhonen**
Language Technology Lab, University of Cambridge
ljvm2@cam.ac.uk

🤗 **Collection** ljvmiranda921/polyglot-teachers **Code** ljvmiranda921/polyglot-teachers

Abstract

Synthesizing supervised finetuning (SFT) data from language models (LMs) to teach smaller models multilingual tasks has become increasingly common. However, teacher model selection is often ad hoc, typically defaulting to the largest available option, even though such models may have significant capability gaps in non-English languages. This practice can result in poor-quality synthetic data and suboptimal student downstream performance. In this work, we systematically characterize what makes an effective multilingual teacher. We combine intrinsic measures of data quality with extrinsic student model performance in a metric we call POLYGLOT SCORE. We evaluate 10 LMs across 6 typologically diverse languages, generating over 1.4M SFT examples and training 240 student models. Our analyses reveal that model scale alone does not significantly predict teacher effectiveness: the most effective teachers we identify are consistently smaller than the largest models evaluated, and their ranking is stable across student base model families. Instead, data qualities such as prompt diversity, length, and response fluency capture 93.3% of the variance in intrinsic data quality and predict student performance. Finally, we provide practical recommendations, including matching the model families of teacher-student pairs and generating responses to existing prompts or translating them from English, which can yield improvements for less-resourced languages. We hope that our work advances data-centric research in multilingual synthetic data and LM development.

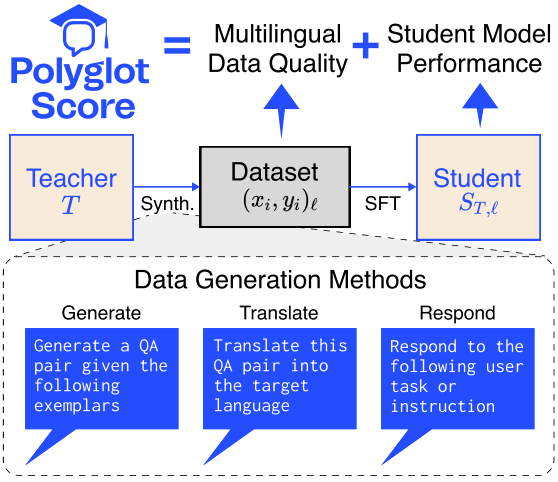


Figure 1: **Systematic evaluation of language models as multilingual teachers.** Across a synthetic data pipeline for language ℓ, we evaluate a teacher model T on their ability to generate supervised finetuning data $(x_i, y_i)_\ell$ in order to train a capable student model $S_{T,\ell}$. We examine three data generation methods: generate, translate, and respond. Our metric, POLYGLOT SCORE, incorporates both intrinsic data quality metrics and extrinsic student model performance to assess the effectiveness of a teacher model for a target language.

1 Introduction

Supervised finetuning (SFT, Ouyang et al., 2022) has emerged as a standard approach for adapting language models (LMs) to specific target languages (Aryabumi et al., 2024; Zhang et al., 2026b, *inter alia*). Central to the success of SFT is the availability of high-quality training data, consisting of pairs of user prompts and corresponding responses, which is often scarce for less-resourced languages (Kunchukuttan et al., 2025). Generating prompt-response pairs for these languages demands substantial human effort (Singh et al., 2024; Kapania et al., 2025), creating a bottleneck for language-specific model development.

To alleviate the challenge of human effort and data scarcity, synthetic data generation using LMs has gained traction as a promising solution for multilingual LM development (Cahyawijaya et al., 2024; Ng et al., 2025; Martins et al., 2025; Hammoud et al., 2026, *inter alia*). This approach involves leveraging a typically larger **teacher model** to generate training examples, which are then used to finetune a smaller **student model** to replicate the

knowledge of the teacher (Kim and Rush, 2016). However, existing works often select teacher models arbitrarily, defaulting to the largest state-of-the-art models that excel on benchmarks (Xu et al., 2025b; Li et al., 2025; Zhang et al., 2026a). This practice is problematic because these models, despite strong performance, may have significant capability gaps in non-English languages, leading to poor-quality synthetic data that propagates the teacher's weaknesses rather than its strengths. We therefore ask: *"what makes an effective multilingual teacher for synthetic data generation, and how can we systematically measure it?"*

In this work, we conduct a comprehensive analysis of 10 LMs across 6 typologically diverse languages on three common synthetic data generation methods: responding to a user query or instruction, translating prompts from English to a target language, and generating prompt-response pairs given in-context examples (§2.2). To systematically assess teacher model effectiveness, we evaluate LMs using both **intrinsic measures of data quality** (§2.3, i.e., the diversity of prompts and responses, the perplexity of the base model on the response, and response quality based on a multilingual reward model) and an **extrinsic measure of student model performance** on multilingual tasks (§2.4, cultural understanding, mathematical reasoning, general chat). We aggregate these measurements into a single metric called POLYGLOT SCORE (PG-SCORE), in order to provide a holistic assessment of a teacher model's data generation capabilities. Our contributions are as follows:

- We close the **evaluation gap** by evaluating 10 teacher models, generating over 1.4M SFT examples and finetuning 240 student models from OLMo 3 7B. We find that Gemma 3 27B consistently ranks within the top three highest PG-SCORE and that the Gemma 3 model family outperforms other families such as Llama 3.1 and IBM Granite (§3.1). Our PG-SCORE rankings are consistent across other base model families (Llama 3.1 8B, Qwen 3 8B, Gemma 3 4B, §3.2).
- We provide **analyses and insights** on the characteristics of a good multilingual teacher model. Our analyses reveal that model scale and benchmark performance, which are common assumptions about a "strong" model, do not significantly predict teacher effectiveness (§4.1). Instead, we find that qualities of the generated data, namely prompt diversity and length coupled with fluent and diverse responses, capture 93.3% of the variance in intrinsic data quality metrics, and their principal components predict student performance with $R^2 = 0.664$ (§4.2).
- Based on these findings, we **recommend a recipe** (§5) for generating multilingual synthetic data. For example, we find that matching the model families of the teacher and student is a useful default when teacher quality is unknown (§3.2), and generating responses to existing prompts or translating from English can yield substantial improvements on less-resourced languages compared to a random mix of data generation methods, though gains vary by teacher model (§3.3). We apply the full recipe to a held-out language, Tagalog, and show that each recommendation yields an observable gain on a language-specific benchmark (§6).

We hope that this work paves the way for developing inclusive and equitable language technologies through high-quality and cost-effective data. We release our code, data, and models to drive research in multilingual synthetic data generation.

2 Evaluating Language Models as Multilingual Teachers

The POLYGLOT SCORE (Figure 1) of a teacher model T for a target language ℓ is based on the intrinsic quality of the synthetic data generated by the teacher (§2.3) and the extrinsic performance of a student model $S_{T,\ell}$ finetuned on this data (§2.4).

2.1 Seed Dataset Construction

In order to bootstrap the synthetic data generation process, we create a seed dataset $\mathcal{D}_{\text{seed},\ell}$ for each target language ℓ. We create $\mathcal{D}_{\text{seed},\ell}$ by aggregating publicly available multilingual instruction-tuning datasets, including the Aya Collection (Aryabumi et al., 2024), WildChat 4.8M (Zhao et al., 2024), EuroBlocks-SFT (Martins et al., 2025), and Magpie-Align (Xu et al., 2025a). In order to simulate scenarios where English prompts are translated into a target language, we also include examples from Tülu 3 SFT (Lambert et al., 2025), Helpsteer3 (chosen responses, Wang et al., 2025), and GSM8K (train split, Cobbe et al., 2021). Detailed seed dataset statistics are in Appendix B.

2.2 Synthetic Data Generation

Given a teacher model T, target language ℓ, and a seed dataset for language ℓ, $\mathcal{D}_{\text{seed},\ell}$, we distill a

synthetic dataset $\mathcal{X}_{T,\ell} = \{(x_i, y_i)\}_{i=1}^{N}$ consisting of N prompt-response pairs (x_i, y_i), then finetune a base model S_{base} on this data to obtain the student model $S_{T,\ell}$:

$$\begin{aligned} \mathcal{X}_{T,\ell} &= T(\mathcal{D}_{\text{seed},\ell}) \\ S_{T,\ell} &= \text{SFT}(S_{\text{base}}, \mathcal{X}_{T,\ell}) \end{aligned} \tag{1}$$

We consider three synthetic data generation methods found in literature (see Appendix A for a brief review of multilingual synthetic data generation methods):

- Generate: we sample k prompt-response pairs from $\mathcal{D}_{\text{seed},\ell}$ as few-shot examples and use T to generate a new pair (x_i, y_i) conditioned on these examples.
- Translate: we forward-translate English prompts from $\mathcal{D}_{\text{seed},\ell}$ to the target language ℓ to obtain x_i, and use T to generate the corresponding response y_i.
- Respond: we take a prompt x_i from $\mathcal{D}_{\text{seed},\ell}$ and use T to generate the response y_i.

2.3 Intrinsic: Multilingual Data Quality & Diversity Metrics

Data is valuable when it is both high-quality and diverse (Raventos et al., 2023; Chen et al., 2024; Zhu et al., 2025).[1] To estimate the value of $\mathcal{X}_{T,\ell}$, we compute a set of lexical and model-based metrics:

- Diversity of prompts and responses (d_x, d_y): a corpus-level statistic that computes the average pairwise cosine distance between the prompt and response embeddings. In practice, we use Llama-Embed-Nemotron-8B (Babakhin et al., 2025), the top-performing model on the MMTEB leaderboard (Enevoldsen et al., 2025), to embed the texts.
- Perplexity (PPL): the perplexity of a base model on the response y_i conditioned on the prompt x_i, measuring the fluency and naturalness of the generated text. Lower perplexity indicates more coherent and linguistically natural responses.
- Reward score of a multilingual reward model (R): the verbalized score (1–5) of a multilingual reward model based on rubrics relating to fluency, naturalness, and instruction-following. In practice, we prompt M-Prometheus 14B (Pombal et al., 2025) as an LM judge to score the quality of the prompt-response pair (Figure 13). We choose M-Prometheus because of its high performance on human-aligned evaluation benchmarks, suggesting that the reward model aligns well with native speakers.

We combine these intrinsic metrics by scaling each metric using z-score normalization and averaging them as shown in Equation 2.

$$\begin{aligned} \text{Intrinsic}_{T,\ell} &= \frac{1}{|M|} \sum_{m \in M} \text{z-score}(m(\mathcal{X}_{T,\ell})) \\ \text{where } M &= \{d_x, d_y, -\log(1 + \text{PPL}), R\} \end{aligned} \tag{2}$$

2.4 Extrinsic: Student Model Performance

We perform supervised finetuning of a base model S_{base} on the synthetic dataset $\mathcal{X}_{T,\ell}$ to obtain a student model $S_{T,\ell}$ (Equation 1). Then, we evaluate $S_{T,\ell}$ on a suite of multilingual tasks to assess how well the student has learned from the teacher. These tasks include:

- Cultural and factual understanding (CULTURE): we evaluate on Global-MMLU Lite (Singh et al., 2025), containing culturally diverse and relevant questions that native speakers localized from English (Hendrycks et al., 2021).
- General chat (CHAT): we evaluate on M-RewardBench (Gureja et al., 2025), which measures the alignment of models with human preferences in conversational settings.
- Mathematical reasoning (MATH): we evaluate on M-GSM (Shi et al., 2023), a multilingual version of the GSM8K dataset (Cobbe et al., 2021) that tests the model's ability to solve mathematical word problems.

We choose these benchmarks because they cover distinct capabilities, are available across all target languages, and are widely adopted in the evaluation suites of recent multilingual LMs (Grattafiori et al., 2024; Gemma Team et al., 2025; Yang et al., 2025, *inter alia*). Instead of reporting the absolute score, we use the *performance gap recovered* metric (PGR, Kim et al., 2025) to measure the improvement of a student over its base model relative to a reference model S_{ref} on the same scale.

$$\begin{aligned} \text{Extrinsic}_{T,\ell} &= \frac{1}{|B|} \sum_{b \in B} \text{PGR}_b(S_{T,\ell}) \\ \text{PGR}_b(S_{T,\ell}) &= \frac{\text{score}_b(S_{T,\ell}) - \text{score}_b(S_{\text{base}})}{\text{score}_b(S_{\text{ref}}) - \text{score}_b(S_{\text{base}})} \\ \text{where } B &= \{\text{CULTURE}, \text{CHAT}, \text{MATH}\} \end{aligned} \tag{3}$$

[1] We use "data quality" to refer to both aspects hereafter.

Teacher Model	Average	Arabic (ar)	Czech (cs)	German (de)	Spanish (es)	Indonesian (id)	Japanese (ja)
Gemma 3 27B Inst.	0.726	**0.145**	**0.360**	1.655	**1.358**	0.214	0.626
Aya Expanse 32B	0.706	−0.058	0.222	1.468	1.129	**1.153**	0.320
Gemma 3 12B Inst.	0.595	−0.464	0.327	1.756	1.228	0.151	0.573
Command A	0.546	−1.360	0.114	1.673	1.102	1.063	**0.683**
Gemma 3 4B Inst.	0.469	−0.488	0.330	1.644	0.929	−0.105	0.504
GPT-4o mini	0.461	−1.117	0.015	**1.766**	0.908	1.003	0.189
IBM Granite 4.0	0.312	−0.072	−0.031	1.000	0.734	−0.079	0.321
IBM Granite Micro	0.304	−0.282	0.290	1.102	0.783	−0.329	0.264
Llama 3.1 70B Inst.	0.140	−0.964	0.109	1.195	0.688	0.182	−0.373
Llama 3.1 8B Inst.	−0.356	−1.693	−0.974	0.891	0.182	0.322	−0.863

Table 1: **Top models with the highest PG-SCORE (average across six languages).** We evaluate teacher models with varying size and model family on 6 typologically diverse languages. For each language, we highlight the best model in **bold** and the second-best model with an underline. Detailed results with standard errors are in Table 14.

Here, S_{ref} is an instruction-tuned model fine-tuned from the same pretrained checkpoint as S_{base}, and both are held fixed across all teachers. For our main experiments, we use OLMo 3 7B as S_{base} and OLMo 3 7B Instruct SFT as S_{ref}.

2.5 POLYGLOT SCORE Computation

To provide straightforward comparisons between teacher models, PG-SCORE reports a single score that combines both extrinsic and intrinsic metrics as shown in Equation 4.

$$\text{PG-SCORE}_{T,\ell} = \text{z-score}(\text{Intr.}_{T,\ell} + \text{Extr.}_{T,\ell}) \quad (4)$$

We combine both intrinsic and extrinsic metrics because they capture complementary aspects of teacher quality. Extrinsic metrics alone may overlook the quality of synthetic data that propagates through the ecosystem, while intrinsic metrics alone do not guarantee that the student model achieves strong downstream performance. The resulting PG-SCORE is z-score normalized, where 0 indicates average teacher effectiveness, and higher scores indicate better synthetic data quality and student performance for that language. We adopt equal weighting as a baseline; we show that teacher rankings are robust to alternative weighting schemes in Appendix G.4.

3 Experiments: Evaluating LMs and PG-SCORE Generalization

In this section, we measure the POLYGLOT SCORE of state-of-the-art LMs (§3.1). Then, we test whether our findings are consistent across other base models (§3.2). Finally, we determine if a certain data generation method is more effective in multilingual settings (§3.3). We conduct additional experiments and ablations in Appendix G.

3.1 Which State-of-the-Art LMs Are Good Multilingual Teachers?

Setup In order to evaluate the effectiveness of different LMs as multilingual teachers, we select 10 state-of-the-art models that vary in scale, architecture, and training data, then evaluate them on 6 typologically diverse languages by generating 10.5k prompt-response pairs for each teacher-language pair where each data generation method (§2.2) is equally represented. We repeat the data generation process three times with different random seeds to account for variability in LM outputs. Then, we finetune a pretrained OLMo 3 7B model (Team Olmo et al., 2026) on each $\mathcal{X}_{T,\ell}$ to obtain $S_{T,\ell}$. Appendix E.1 describes SFT information.

Teacher Models We include Llama 3.1 (8B, 70B, Grattafiori et al., 2024), Gemma 3 (4B, 12B, 27B, Gemma Team et al., 2025), Command A (Team Cohere et al., 2025), Aya Expanse 32B (Dang et al., 2024), and IBM Granite (4.0, Micro, Granite Team, IBM, 2025). In addition, we also include GPT-4o mini (OpenAI et al., 2024) as a representative closed-source model. See Table 8 in Appendix D for detailed model information.

Target Languages We select 6 typologically diverse languages: Arabic (ar), Czech (cs), German (de), Spanish (es), Indonesian (id), and Japanese (ja). These languages are chosen due to their variation in resource availability, script, and family. This language choice is also supported by prior work on informed sampling (Ploeger et al., 2026) that considers typological variety of the chosen languages. See Table 9 in Appendix D for language statistics.

Results Table 1 shows the PG-SCORE of each teacher model across all target languages. The

Teacher Model	Base Model (S_{base}) OLMo 3 7B	Gemma 3 4B	Qwen 3 8B	Llama 3.1 8B
GPT-4o mini	0.551	1.022	1.005	0.621
Llama 3.1 70B Inst.	0.138	0.338	1.039	0.497
Llama 3.1 8B Inst.	−0.160	−0.133	0.365	0.048
Command A	0.459	0.725	0.974	0.737
Aya Expanse 32B	0.854	0.762	1.183	0.793
Gemma 3 27B Inst.	0.672	0.810	1.301	0.800
Gemma 3 12B Inst.	0.481	0.666	1.393	0.804
Gemma 3 4B Inst.	0.350	0.712	0.545	1.062
IBM Granite 4.0	0.283	0.278	0.831	−0.001
IBM Granite Micro	0.164	0.455	1.079	0.396
Qwen 3 8B	0.398	0.615	1.421	0.662
OLMo 3 7B Inst.	0.731	0.295	0.581	0.277

Table 2: **PG-SCORE of each teacher model across different base models (average across Arabic, German, and Indonesian).** We highlight the top, second, and third best teacher models for each base model.

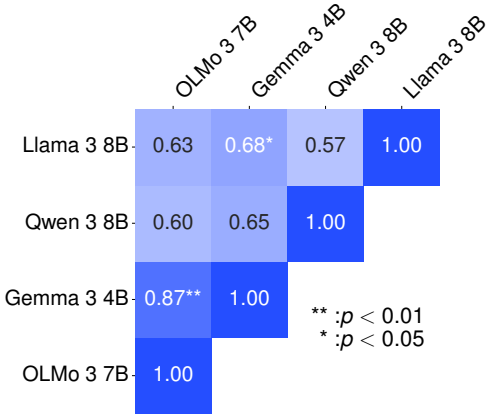


Figure 2: **Consistency of teacher rankings across base models.** Spearman rank correlation ρ between the teacher rankings induced by each pair of base models.

Teacher Model	Arabic (ar) Generate	Translate	Respond	German (de) Generate	Translate	Respond	Indonesian (id) Generate	Translate	Respond
Gemma 3 27B Inst.	0.032	0.276	**0.802**	**2.140**	2.086	1.212	1.189	**1.196**	0.046
Aya Expanse 32B	−0.276	**0.148**	−1.349	**1.473**	1.255	1.451	0.039	0.733	**1.606**
Llama 3.1 70B Inst.	−0.867	−1.025	**−0.215**	**1.391**	0.459	1.187	−0.146	0.089	**0.155**

Table 3: **PG-SCORE across three data generation methods: Generate, Translate, and Respond (§2.2).** For each data generation method, we generate 10k samples per teacher-language pair and finetune a student model on OLMo 3 7B. We show percentage increases in PG-SCORE compared to a baseline (equal representation of the three data generation methods) in Table 15.

results suggest the following:

- **Gemma 3 27B and Aya Expanse 32B are the most effective teachers.** Gemma 3 27B achieves the highest average PG-SCORE (0.726), followed closely by Aya Expanse 32B (0.706), both outperforming larger models such as Llama 3.1 70B Inst. (0.140), suggesting that model scale alone does not determine teacher effectiveness. We also observe that the Gemma 3 family dominates the top ranks, while the Llama 3.1 family underperforms on most languages.
- **Smaller LMs can be effective multilingual teachers.** Gemma 3 12B (0.595) and 4B (0.469) rank among the top-5 teachers, while the Llama 3.1 70B Inst. (0.140) ranks ninth, suggesting that smaller LMs can match or exceed larger LMs in data generation capabilities.
- **Teacher performance varies significantly by language.** German and Spanish consistently show the highest scores across all models, while Arabic proves challenging with most teachers yielding negative scores, suggesting that language-specific factors influence teacher effectiveness. We hypothesize that a language's resource status or presence in pretraining data may contribute to this variability (Appendix G.5).

3.2 Generalization of PG-SCORE Across Different Base Models

Setup Instead of using OLMo 3 7B as the base model (S_{base}) for student finetuning, we use (1) Llama 3.1 8B, (2) Gemma 3 4B PT, and (3) Qwen 3 8B Base (Yang et al., 2025). We recompute S_{base}-dependent metrics such as perplexity and PGR. To reduce computational costs, we focus on three languages: German (high PG-SCORE), Indonesian (mid-range), and Arabic (low PG-SCORE).

Results Table 2 shows the average PG-SCORE of each teacher model across different base models. We observe that the **best teacher models remain consistent across different student base models**, with Gemma 3 27B and Aya Expanse 32B consistently ranking among the top three teachers. Furthermore, the Gemma 3 family continues to outperform other model families. In addition, we find that the model rankings vary slightly de-

pending on the base model used, as Spearman rank correlation ranges from $\rho = 0.57$ (moderate) to $\rho = 0.87$ (strong, Figure 2). We hypothesize that this variation may be due to differences in architecture and pretraining data between base models. Despite this variation, we observe that **matching teacher and student model families is a useful default when teacher quality is unknown**. The matched teacher ranks among the top two on three of the four base models we test (Table 2): Gemma 3 teachers consistently perform well with Gemma 3 student bases, while Qwen 3 8B is the strongest teacher on the Qwen 3 8B base and OLMo 3 7B Instruct ranks second on the OLMo 3 7B base. The exception is Llama 3.1, which is a weak multilingual teacher overall (§3.1), suggesting that family matching helps only when the matched teacher is itself capable in the target languages. This finding is reasonable given that models from the same family likely share similar tokenization schemes, leading to easier transfer from teacher to student. For our core experiment, we use OLMo 3 7B as the base model for finetuning to control the effect of model family alignment when evaluating teacher quality.

3.3 Effect of Synthetic Data Generation Method on PG-SCORE

Setup In order to determine if a data generation method is more effective than others, we generate 10k prompt-response pairs for each method in §2.2 and compare the PG-SCORE of each mix. We recompute intrinsic data quality metrics and finetune OLMo 3 7B to obtain a student model and evaluate the teacher's PG-SCORE. We also compare each mix against a baseline consisting of 10k instances with roughly equal numbers of samples (≈3.3k) from each method. To reduce computational costs, we conduct this experiment on three representative teachers (Gemma 3 27B, Aya Expanse 32B, and Llama 3.1 70B) spanning high to low PG-SCORE, and three languages (German, Indonesian, Arabic) covering diverse resource levels.

Predictor	β	SE	p
log (Param. Size)	0.053	0.080	0.507
Avg. Multilingual Perf.	1.387	2.204	0.529

Table 4: **Results from a mixed-effects regression model on PG-SCORE on an LM's (a) size and (b) avg. multilingual benchmark performance.** The lack of significant correlation suggests that both predictors are not solely sufficient to ensure teacher effectiveness.

PC	Variance Expl.	Cumulative
PC1	42.2%	42.2%
PC2	22.1%	64.3%
PC3	16.5%	80.8%
PC4	12.6%	93.3%
PC5	3.5%	96.8%
PC6	3.2%	100.0%

Table 5: **Variance explained by principal components from intrinsic data quality metrics.** There are four principal components that explain 93.3% (cumulative) of the variance.

Results Table 3 shows the PG-SCORE of each data generation method (see Table 15 for baseline comparisons). We observe that for a **high-resource language like German, the *Generate* method yields the highest PG-SCORE**, while for **less-resourced languages like Arabic and Indonesian, the *Respond* or *Translate* methods are more effective**. We hypothesize that this occurs because the *Generate* method depends on few-shot examples from the seed dataset, which are typically of higher quality in high-resource languages. Overall, our findings suggest that selecting a data generation method can have an impact on teacher effectiveness. In our core experiment, we sample an equal mix of all three methods (3.5k each) to control their effect when evaluating teacher model quality.

4 Analysis: What Makes a Good Multilingual Teacher Model?

We investigate the factors that contribute to effective multilingual teachers. We start by analyzing common assumptions about teacher model performance, such as size and benchmark scores (§4.1), then determine which intrinsic factors drive student performance (§4.2).

4.1 Do stronger models make better teachers?

Setup In order to determine if there is a relationship between a model's size or benchmark performance (i.e., common assumptions to assess a model's "strength") to its effectiveness as a multilingual teacher, we fit a mixed-effects model regressing PG-SCORE on (a) parameter size (N=27, i.e., 9 models × 3 trials, excluding GPT-4o mini whose size is undisclosed), and (b) average multilingual benchmark performance on Global-MMLU

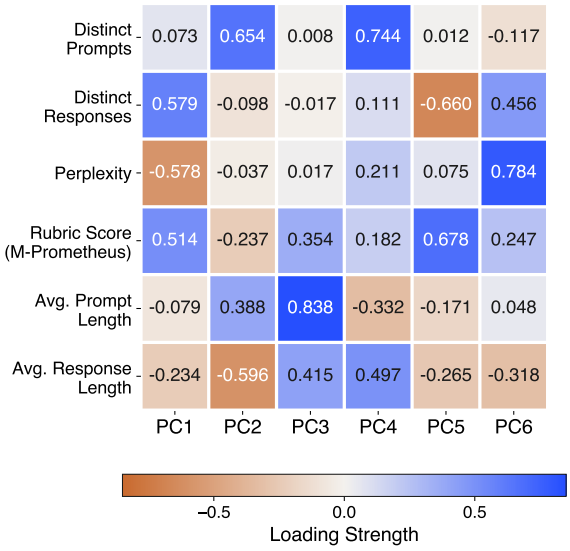


Figure 3: **Loading strength of intrinsic metrics on the principal components (PCs).** PC1 suggests that good teachers produce diverse and high-quality responses, while PC2 focuses on prompt diversity and length. PC3 and PC4, together, indicate the importance of prompts on student performance.

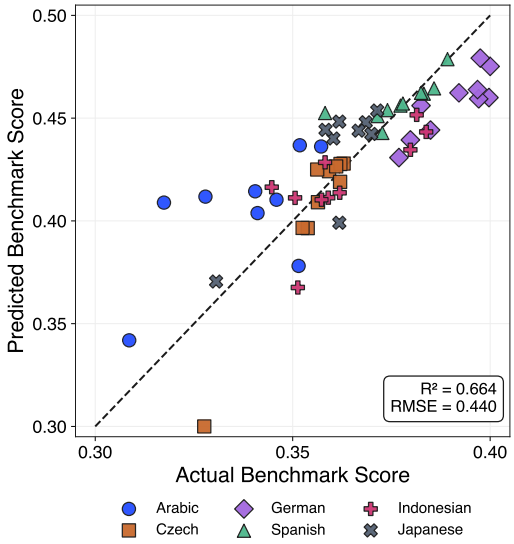


Figure 4: **Fit of a linear regression model on the PCs of the intrinsic metrics to predict student performance.** Intrinsic metrics, via their PCs, can predict extrinsic student performance ($R^2 = 0.664$ and RMSE $= 0.440$) on multilingual benchmarks (§2.4).

Lite, M-GSM, and M-RewardBench (N=180, 10 models $\times$ 6 languages $\times$ 3 trials).

Results Table 4 shows the regression results. We observe that **neither parameter size nor average multilingual benchmark performance significantly predict PG-SCORE** ($p > 0.05$). Specifically, a 1-unit increase in $\log$(Param. Size) corresponds to a non-significant 0.053 increase in PG-SCORE. Although this finding confirms the results of Xu et al. (2025b) and Kim et al. (2025) for English-based tasks, we show that "stronger" models do not necessarily make better multilingual teachers.

4.2 Which intrinsic metrics determine extrinsic student model performance?

Setup In order to identify latent factors from the intrinsic metrics that explain student performance, we perform principal component analysis (PCA) on the intrinsic metrics described in §2.3. Then, we fit a regression model to predict extrinsic student performance based on the principal components (PCs) obtained from PCA: we split 180 data points (10 models $\times$ 6 languages $\times$ 3 trials) into 80% train and 20% test, then train a linear regression model with the PCs as the features and the student performance as the target.

Results Table 5 shows how much of the variance is explained by each principal component while Figure 3 shows the loading strength of each intrinsic metric on the principal components. We observe that the first four PCs explain 93.3% of the variance in the intrinsic data quality metrics. Specifically, PC1 (42.2%) captures characteristics such as **lower response perplexity and high distinctiveness**, PC2 (22.1%) captures variance in characteristics such as **higher prompt diversity and length**, whereas PC3 (16.5%) and PC4 (12.6%) capture variance that reinforces trends on prompt length and diversity. In addition, Figure 4 shows the fit of a linear model on the test set when the PCs learn to predict student performance. We observe that interactions within the intrinsic metrics can predict extrinsic student performance reasonably well, with $R^2 = 0.664$ and RMSE $= 0.440$. This finding suggests that even with a simple linear model, our chosen **intrinsic metrics are predictive of student performance.** In practice, these insights can help practitioners select teacher models based on intrinsic metrics alone, which are cheaper to compute than extrinsic student evaluations.

5 Discussion: Towards a Recipe for Multilingual Synthetic Data Generation

Our results provide actionable insights for selecting and effectively using teacher models in mul-

tilingual synthetic data generation. First, we find that **model scale does not significantly predict teacher effectiveness**: Llama 3.1 70B Instruct, despite being one of the largest models evaluated, ranks at the bottom half in PG-SCORE across all student base models we tested (§3.1, §3.2). Our analyses suggest that what matters instead is the quality of generated data: prompt diversity, response fluency, and length collectively capture over 93% of the variance in intrinsic data quality and predict student performance with $R^2 = 0.664$ (§4.2), offering practitioners a cheaper alternative to full student training runs for screening teachers.

Second, when the optimal teacher is unknown, **matching model families offers a useful default for teacher selection**. Gemma, Qwen, and OLMo teachers paired with students from their own family rank among the top two teachers on that base (Table 2), though this holds only when the matched teacher is itself a capable multilingual model. We hypothesize this finding reflects shared tokenization and similar pretraining distributions, though disentangling these factors remains future work.

Finally, we find that **there are language-dependent considerations for data generation.** For high-resource languages like German, where seed data quality is high, the *Generate* method performs best. For less-resourced languages like Arabic and Indonesian, methods that leverage existing prompts (*Respond*) or transfer from English (*Translate*) can yield substantial gains over a uniform mix of methods, though the magnitude varies by teacher (Table 3). For truly low-resource languages, we recommend combining synthetic data generation with targeted data collection.

6 Case Study: Applying the Synthetic Data Recipe to a Held-Out Language

As an application of our findings and discussion in §5, we present a case study on developing a multilingual synthetic data recipe on a held-out language: Tagalog. It is a mid-resource language (Category 3 in Joshi et al. (2020)'s taxonomy) and the standardized form of Filipino, the national language of the Philippines. We evaluate on FILBENCH (Miranda et al., 2025), a Filipino-centric benchmark that was not used to compute PG-SCORE. In order to measure the contribution of our findings and recommendations, we perform the following ablation experiments as shown in Figure 5. Note that the interventions described below are additive, and we describe the full recipe in Appendix I.1.

We find that each intervention yields an observable gain. Replacing publicly available Tagalog SFT data with synthetic instances from a GPT-4o mini teacher performs similarly ($\Delta = 0.43$ pp), suggesting that there is no significant advantage to using a synthetic data pipeline if the teacher model is not optimal. Swapping in a teacher with a higher PG-SCORE (Aya Expanse 32B, 0.706 vs. 0.461) improves performance, suggesting that the metric is generalizable across an unseen language, while matching the teacher to the family of the Gemma 3 4B base model yields a substantial further improvement. Increasing data scale (10k to 25k instances, $+0.21$ pp) and student model scale (4B to 12B and 27B) provide additional but smaller gains. We report the full ablation, the FILBENCH leaderboard comparison, and a native-speaker inspection of the generated data in Appendix I.

7 Related Work

Synthetic Data Generation for Multilingual SFT In order to offset the high costs of recruiting language experts for data collection, prior work has relied on generating synthetic datasets. This effort resulted in large multilingual datasets such as Bactrian-X (*Translate,* Li et al., 2023), MultiAlpaca (*Generate,* Wei et al., 2023), and xP3 (*Respond,* Muennighoff et al., 2023) that were created through various data generation methods. These works have different data generation recipes, and so we provide a brief survey of these works and their recipes in Appendix A, then classify them across the three strategies (Generate, Translate, Respond; §2.2). Building on these prior efforts, we examine the three core strategies for multilingual synthetic data generation and test each in isolation. This setup enables us to provide practitioners with an empirically grounded recipe for selecting teacher LMs that we hope is applicable across any generation method.

Evaluating and Improving the Synthetic Data Pipeline While prior works have evaluated aspects of the synthetic data pipeline, they typically do so in isolation (i.e., intrinsic *or* extrinsic, but not both) or focus exclusively on English (Zhang et al., 2026a). For instance, Kim et al. (2025) evaluated teacher models solely as a function of *extrinsic* student performance on English tasks (e.g., reasoning and coding), while Cai et al. (2025)'s OpenDataArena focuses on *intrinsic* data quality

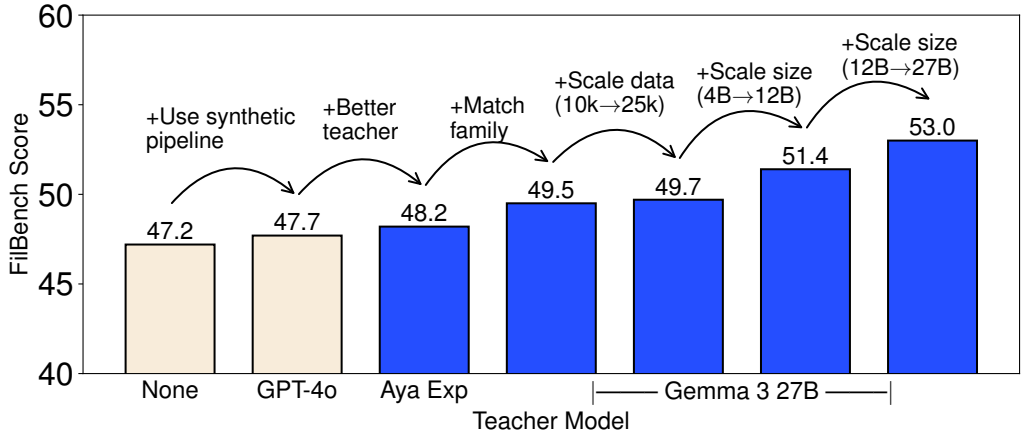


Figure 5: **Student model performance on a held-out language, Tagalog, across several synthetic data interventions.** Given a held-out language (Tagalog) and an evaluation benchmark (FILBENCH), we apply data interventions based on our recommendations for creating a multilingual synthetic data recipe (§5). We provide ablation details and leaderboard results in Appendix I.

(model-based and heuristic) to score models. Signals of multilingual data quality are often a function of corpus-level diversity (Artetxe and Schwenk, 2019; Enevoldsen et al., 2025; Sam et al., 2025) and generation quality (Pombal et al., 2025; Anugraha et al., 2026). On the other hand, multilingual LMs are typically evaluated on general-knowledge and culture-specific benchmarks (Qin et al., 2025; Gemma Team et al., 2025; Salamanca et al., 2026, *inter alia*). These practices informed our choice of intrinsic and extrinsic metrics throughout this work. More importantly, PG-SCORE provides a holistic metric that combines both intrinsic data quality and extrinsic student downstream performance to evaluate teacher models across various generation methods.

8 Conclusion

We present POLYGLOT SCORE, a metric that evaluates LMs as multilingual teachers by combining the intrinsic quality of the synthetic data they generate with the extrinsic performance of the students they train. Across 10 teachers, 6 typologically diverse languages, and 240 student models, we find that model scale alone does not predict teacher effectiveness: Gemma 3 27B and Aya Expanse 32B are the strongest teachers we evaluate, both outperforming Llama 3.1 70B Instruct. What matters instead are properties of the generated data, as prompt diversity, length, and response fluency capture 93.3% of the variance in our intrinsic metrics and predict student performance with $R^2 = 0.664$. We distill these results into a recipe for multilingual synthetic data generation and validate it on a held-out language, Tagalog, where each recommendation yields an observable gain on FILBENCH. We hope our findings guide future work on developing inclusive language technologies through high-quality synthetic data.

Limitations

Our work comes with some limitations and open questions left for future work. For example, our language set encompasses six languages. Although we chose these languages carefully based on (1) whether they can be evaluated on publicly available LM benchmarks and (2) prior theoretical work on principled test language selection (Ploeger et al., 2026), validating our findings across a broader language sample remains important future work. In addition, our *Translate* data generation method assumes access to English prompts that can be meaningfully translated to target languages. This approach inherits limitations from LM-based techniques, such as difficulty localizing culture-specific references and the risk of introducing translationese artifacts.

Ethics Statement

Synthetic data generation risks amplifying biases present in teacher models. If a teacher model underperforms on certain languages or exhibits cultural biases, these weaknesses propagate to student models trained on its outputs. Our finding that teacher effectiveness correlates with CommonCrawl representation (Appendix G.5) suggests that already

underrepresented languages may be further disadvantaged in synthetic data pipelines, potentially widening the performance gap between high- and low-resource languages.

Acknowledgments

LJVM and AK acknowledge the support of the UKRI Frontier Grant EP/Y031350/1 (EQUATE). This work was performed using joint resources provided by the Cambridge Service for Data Driven Discovery (CSD3) EP/T022159/1, Isambard AI National AI Research Resource (AIRR) ST/AIRR/I-A-I/1023, and the Microsoft Research Grant. LJVM would also like to thank Songbo Hu, Chen Cecilia Liu, Millicent Ochieng, and Felermino Ali for helpful and productive discussions on the project. Finally, the authors thank the reviewers and area chair of the May ARR cycle for valuable feedback that further improved this work.

References

Sanchit Ahuja, Kumar Tanmay, Hardik Hansrajbhai Chauhan, Barun Patra, Kriti Aggarwal, Luciano Del Corro, Arindam Mitra, Tejas Indulal Dhamecha, Ahmed Hassan Awadallah, Monojit Choudhury, Vishrav Chaudhary, and Sunayana Sitaram. 2025. sPhinX: Sample Efficient Multilingual Instruction Fine-Tuning Through N-shot Guided Prompting. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 927–946, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

David Anugraha, Shou-Yi Hung, Zilu Tang, En-Shiun Annie Lee, Derry Tanti Wijaya, and Genta Indra Winata. 2026. mR3: Multilingual Rubric-Agnostic Reward Reasoning Models. In *The Fourteenth International Conference on Learning Representations*.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. Aya 23: Open Weight Releases to Further Multilingual Progress. *Preprint*, arXiv:2405.15032.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-Embed-Nemotron-8B: A Universal Text Embedding Model for Multilingual and Cross-Lingual Tasks. *Preprint*, arXiv:2511.07025.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Afina Putri, Emmanuel Dave, Jhonson Lee, Nuur Shadieq, Wawan Cenggoro, Salsabil Maulana Akbar, Muhammad Ihza Mahendra, Dea Annisayanti Putri, Bryan Wilie, Genta Indra Winata, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2024. Cendol: Open instruction-tuned generative large language models for Indonesian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14899–14914, Bangkok, Thailand. Association for Computational Linguistics.

Mengzhang Cai, Xin Gao, Yu Li, Honglin Lin, Zheng Liu, Zhuoshi Pan, Qizhi Pei, Xiaoran Shang, Mengyuan Sun, Zinan Tang, Xiaoyang Wang, Zhanping Zhong, Yun Zhu, Dahua Lin, Conghui He, and Lijun Wu. 2025. OpenDataArena: A Fair and Open Arena for Benchmarking Post-Training Dataset Value. *Preprint*, arXiv:2512.14051.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. 2024. On the Diversity of Synthetic Data and its Impact on Training Large Language Models. *Preprint*, arXiv:2410.15226.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *Preprint*, arXiv:2110.14168.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. *Preprint*, arXiv:2412.04261.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2025. MMTEB: Massive Multilingual Text Embedding Benchmark. In *The Thirteenth International Conference on Learning Representations*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane

Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Granite Team, IBM. 2025. Granite 4.0 Language Models. https://huggingface.co/collections/ibm-granite/granite-40-language-models. Accessed: 2025-12-08.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Srishti Gureja, Lester James Validad Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Triandi Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2025. M-RewardBench: Evaluating Reward Models in Multilingual Settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43–58, Vienna, Austria. Association for Computational Linguistics.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. LightEval: A lightweight framework for LLM evaluation.

Hasan Abed Al Kader Hammoud, Mohamad Bilal Zbib, and Bernard Ghanem. 2026. Hala Technical Report Building Arabic-Centric Instruction & Translation Models at Scale. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*, pages 236–244, Rabat, Morocco. Association for Computational Linguistics.

Daniel Han, Michael Han, and Unsloth Team. 2023. Unsloth.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *Preprint*, arXiv:1612.03651.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Shivani Kapania, Stephanie Ballard, Alex Kessler, and Jennifer Wortman Vaughan. 2025. Examining the Expanding Role of Synthetic Data Throughout the AI Development Pipeline. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 45–60, New York, NY, USA. Association for Computing Machinery.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. 2025. Evaluating language models as synthetic data generators. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6385–6403, Vienna, Austria. Association for Computational Linguistics.

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.

Anoop Kunchukuttan, Raj Dabre, Rudra Murthy, Mohammed Safi Ur Rahman Khan, and Thanmay Jayakumar. 2025. Data and Model Centric Approaches for Expansion of Large Language Models to New languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 12–13, Suzhou, China. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Katherine Lambert and Sasha Luccioni. 2026. Towards Resource-Efficient LLMs: End-to-End Energy Accounting of Distillation Pipelines. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, Seoul, South Korea. PMLR.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha

Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. In *Second Conference on Language Modeling*.

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

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025. Small Models Struggle to Learn from Strong Reasoners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25366–25394, Vienna, Austria. Association for Computational Linguistics.

Ryan Marten, Trung Vu, Charlie Cheng-Jie Ji, Kartik Sharma, Shreyas Pimpalgaonkar, Alex Dimakis, and Maheswaran Sathiamoorthy. 2025. Curator: A Tool for Synthetic Data Creation. `https://github.com/bespokelabsai/curator`.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. EuroLLM-9B: Technical Report. *Preprint*, arXiv:2506.04079.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. EuroLLM: Multilingual Language Models for Europe. *Preprint*, arXiv:2409.16235.

Lester James Validad Miranda, Elyanah Aco, Conner G. Manuel, Jan Christian Blaise Cruz, and Joseph Marvin Imperial. 2025. FilBench: Can LLMs Understand and Generate Filipino? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2496–2529, Suzhou, China. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Raymond Ng, Thanh Ngan Nguyen, Huang Yuli, Tai Ngee Chia, Leong Wai Yi, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulagadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Tan Choon Meng, Brandon Ong, and 11 others. 2025. SEA-LION: Southeast Asian Languages in One Network. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 512–526, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. GPT-4o System Card. *Preprint*, arXiv:2410.21276.

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

Parinthapat Pengpun, Can Udomcharoenchaikit, Weerayut Buaphet, and Peerat Limkonchotiwat. 2024. Seed-free synthetic data generation framework for instruction-tuning LLMs: A case study in Thai. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 445–464, Bangkok, Thailand. Association for Computational Linguistics.

Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux, and Johannes Bjerva. 2026. A Principled Framework for Evaluating on Typologically Diverse Languages. *Computational Linguistics*, pages 1–33.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and Andre Martins. 2025. M-Prometheus: A Suite of Open Multilingual LLM Judges. In *Second Conference on Language Modeling*.

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.

Neel Prabhanjan Rachamalla, Aravind Konakalla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. 2025. Pragyaan: Designing and Curating High-Quality Cultural Post-Training Datasets for Indian Languages. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 285–321, Suzhuo, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Allan Raventos, Mansheej Paul, Feng Chen, and Surya Ganguli. 2023. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Alejandro R. Salamanca, Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaee, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, and 7 others. 2026. Tiny Aya: Bridging Scale and Multilingual Depth. *Preprint*, arXiv:2603.11510.

Dylan Sam, Ayan Chakrabarti, Afshin Rostamizadeh, Srikumar Ramalingam, Gui Citovsky, and Sanjiv Kumar. 2025. Analyzing Similarity Metrics for Data Selection for Language Model Pretraining. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Muhammad Ali Shafique, Kanwal Mehreen, Muhammad Arham, Maaz Amjad, Sabur Butt, and Hamza Farooq. 2025. Alif: Advancing Urdu Large Language Models via Multilingual Synthetic Data Distillation. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 271–284, Suzhuo, China. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. 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)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, and 210 others. 2025. Command A: An Enterprise-Ready Large Language Model. *Preprint*, arXiv:2504.00698.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

Bibek Upadhayay and Vahid Behzadan. 2024. TaCo: Enhancing Cross-Lingual Transfer for Low-Resource Languages in LLMs through Translation-Assisted Chain-of-Thought Processes. In *5th Workshop on practical ML for limited/low resource settings*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Daniel Egert, Ellie Evans, Hoo-Chang Shin, Felipe Soares, Yi Dong, and Oleksii Kuchaiev. 2025. HelpSteer3: Human-Annotated Feedback and Edit Data to Empower Inference-Time Scaling in Open-Ended General-Domain Tasks. *Preprint*, arXiv:2503.04378.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. PolyLM:

An Open Source Polyglot Large Language Model. *Preprint*, arXiv:2307.06018.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2025a. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *The Thirteenth International Conference on Learning Representations*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, and Radha Poovendran. 2025b. Stronger models are not always stronger teachers for instruction tuning. 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)*, pages 4392–4405, Albuquerque, New Mexico. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

Hengyuan Zhang, Shiping Yang, Xiao Liang, Chenming Shang, Yuxuan Jiang, Chaofan Tao, Jing Xiong, Hayden Kwok-Hay So, Ruobing Xie, Angel X. Chang, and Ngai Wong. 2026a. Find Your Optimal Teacher: Personalized Data Synthesis via Router-Guided Multi-Teacher Distillation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14619–14637, San Diego, California, United States. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and Fei Wu. 2026b. Instruction Tuning for Large Language Models: A Survey. *ACM Comput. Surv.*, 58(7).

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations*.

Alan Zhu, Parth Asawa, Jared Quincy Davis, Lingjiao Chen, Boris Hanin, Ion Stoica, Joseph E. Gonzalez, and Matei Zaharia. 2025. BARE: Leveraging Base Language Models for Few-Shot Synthetic Data Generation. *Preprint*, arXiv:2502.01697.

Appendix

A Multilingual Synthetic Data Generation **17**

B Seed Dataset Statistics **17**

C Release of Data and Software Artifacts **17**

D Teacher Model and Target Language Details **17**

E Experimental Details **17**
E.1 Supervised Finetuning 17
E.2 Model Evaluation 17

F Full Results for Intrinsic and Extrinsic Metrics **18**

G Additional Experiments and Ablations **18**
G.1 Effect of Data Scale on Student Model Performance 18
G.2 Generalization Across Model Size 19
G.3 Effect of Translation Method (Prompting an LM vs. Translation Model) 20
G.4 Weighting of Intrinsic and Extrinsic Metrics in PG-SCORE 21
G.5 Effect of language resource levels on PG-SCORE 22

H Disclosure on the Use of LLMs **22**

I Multilingual Synthetic Data Recipe: Case Study on Tagalog **22**
I.1 Setup: Recipe Design and Evaluation 23
I.2 Results: Leaderboard Scores and Ablations 24
I.3 Analysis: Ablation Experiments 25
I.4 Analysis: Instance-Level Qualitative Inspection of Generated Data 26

J Inference Details **26**

Dataset	Language(s)	Generation Method / Description
Bactrian-X (Li et al., 2023)	52 languages - Arabic, Indonesian, Chinese, Malay, Tamil, Tagalog, etc.	*Translate* - used Google Translate API to translate English instructions from Alpaca (52K) and Dolly (15K).
MultiAlpaca (Wei et al., 2023)	18 languages - English, Chinese, Russian, Spanish, German, French, etc.	*Generate, Translate* - used a multilingual self-instruct (Wang et al., 2023) method from English prompt-response pairs to perform translation.
xP3-MT (Muennighoff et al., 2023)	46 languages - Arabic, English, Spanish, Hindi, Chinese, Indonesian, etc.	*Translate, Respond* - used Google Translate API to translate English prompt-response pairs from different sources, in addition to creating template-based prompts that an LM responds to.
Cendol (Cahyawijaya et al., 2024)	18 Indonesian languages - Sundanese, Javanese, Acehnese, Banjarese, Buginese, Gorontalo, etc.	*Translate, Respond* - curated various prompts from past Indonesian NLP tasks, including translations of Dolly.
Seed-Free Thai (Pengpun et al., 2024)	Thai	*Generate* - generated synthetic instruction data without seed examples by using Wikipedia contexts. Identifies fluency, diversity, and cultural context as key properties.
Aya Dataset and Collection (Singh et al., 2024)	114 languages - Arabic, French, Hindi, Indonesian, Japanese, Spanish, Swahili, Turkish, Yoruba, Filipino, etc.	*Translate, Respond* - involves a collection of translated prompts from English, and templated prompts. A sizeable portion of the collection includes native-speaker annotations.
sPhinX (Ahuja et al., 2025)	51 languages - Afrikaans, Arabic, Bengali, Bulgarian, Burmese, Chinese, Croatian, Czech, etc.	*Translate* - selectively translates essential portions of multilingual inputs in order to semantically preserve meaning.
EuroBlocks (Martins et al., 2025, 2024)	31 languages - English, Chinese, Spanish, Italian, French, German, Portuguese, Dutch, Polish, etc.	*Generate, Translate* - prompted Llama 3 or an earlier EuroLLM checkpoint with a document, target language, and category, then asked it to generate an instruction. Also involved translating prompt-response pairs.
SEA-LION Dataset (Ng et al., 2025)	11 languages - English, Chinese, Indonesian, Vietnamese, Malay, Thai, Burmese, Lao, Filipino, Khmer, and Tamil	*Generate, Translate* - for the majority of the datasets, samples were first generated in English using Qwen 32B, and then translated into the target language using Gemma 2 27B.
Urdu-Instruct Dataset (Shafique et al., 2025)	Urdu	*Generate* - uses a modified Self-Instruct from a pool of culturally relevant prompts.
Pragyaan (Rachamalla et al., 2025)	10 Indian languages - Gujarati, Kannada, Marathi, Bengali, Odia, Tamil, Malayalam, Telugu, Punjabi, and Hindi	*Generate, Translate* - perform translation using an LM for a subset of data. Used Self-Instruct from a pool of native prompts for another subset of data.

Table 6: **Short survey of related work on synthetic data generation for multilingual LMs.** For each work, we provide a brief description of its data generation method. We find that most methods fall into one of the three categories described in §2.2, i.e., Generate, Translate, or Respond, which we tested in our experiments.

A Multilingual Synthetic Data Generation

We present an overview of prior works in Table 6 that used synthetic data to train multilingual LMs. In general, we find that most data generation methods fall into one of the three categories described in §2.2, i.e., *Generate*, *Translate*, or *Respond*, which we tested in our experiments. Our survey suggests that our **choice of data generation methods is grounded in prior work** and covers the majority of approaches used in synthetic data generation.

B Seed Dataset Statistics

Table 7 shows the statistics of the seed dataset used for synthetic data generation.

C Release of Data and Software Artifacts

In order to facilitate future research on multilingual synthetic data generation, we introduce the POLYGLOT collection, a collection of synthetic datasets produced by the best teacher model across all target languages, together with the student models finetuned on them. The POLYGLOT collection includes:

- PolyglotTeachers-SFT-Synth-Data: Synthetic datasets for each target language generated by each teacher model using all three data generation methods (§2.2).
- PolyglotTeachers-Multilingual-Instruct: A set of 7B student models finetuned on each synthetic dataset from the OLMo 3 7B base model using the Gemma 3 27B (highest-scoring model) teacher.

Because synthetic data generation can be expensive (Lambert and Luccioni, 2026), especially when accounting for teacher inference costs, we hope that the POLYGLOT Collection can help amortize the cost of obtaining quality multilingual data without re-running these systematic experiments. We publicly release these artifacts on HuggingFace.[2]

D Teacher Model and Target Language Details

In this section, we provide additional details about the teacher models and target languages used in our experiments. Table 8 summarizes the key characteristics of each teacher model. On the other hand, Table 9 provides information about the target languages, including language family, number of speakers, and resource availability.

E Experimental Details

E.1 Supervised Finetuning

Table 10 summarizes the hyperparameters used for finetuning student models. We train models with the Unsloth framework (Han et al., 2023) on a cluster of Grace Hopper GH200 Superchips. Full finetuning (7B) takes around 1.5 hours (wall clock) for 2 epochs on 2 nodes.

Hyperparameter	Value	Hyperparameter	Value
Learning rate	5e-5	Batch size	32
Epochs	2	Grad. Acum. Steps	4
Max seq. length	16,384	Weight decay	0.001
Optimizer	AdamW	Scheduler	Linear

Table 10: Hyperparameters for finetuning a 7B student model from OLMo 3 7B.

E.2 Model Evaluation

We use the Lighteval framework (`v0.13.1dev0`, Habib et al., 2023) for evaluation. Table 11 summarizes the benchmarks used for evaluating student models. We use Global-MMLU *Lite* instead of Global-MMLU because the former contains actual native speaker annotations that localized the benchmark into different cultural contexts.

Benchmark	Formulation	Metric	N-shots
Global-MMLU Lite	MCF	Accuracy	0
M-RewardBench	MCF	Weighted Acc.	0
M-GSM	Generative	Exact-Match	5

Table 11: Evaluation settings for each benchmark (MCF: Multiple-Choice Formulation).

For Global-MMLU Lite and M-RewardBench, we use the Multiple-Choice Formulation (MCF) with character normalization. In addition, we also follow the corpus-level metric in M-RewardBench which uses a weighted accuracy for each data subset and category (Gureja et al., 2025). For M-GSM, we show 5 few-shot examples from the training set in order for the model to properly generate the answer. We run all evaluation experiments for three trials with different random seeds and report the average and standard error.

[2] 🤗: `ljvmiranda921/polyglot-teachers`

Source	Language						
	English (en)	Arabic (ar)	Czech (cs)	German (de)	Spanish (es)	Indonesian (id)	Japanese (ja)
Aya Dataset	-	-	5,000	241	3,854	2,786	6,259
Tülu 3 SFT	10,000	-	-	-	-	-	-
WildChat 4.8M	10,000	4,660	1,266	5,908	5,900	7,983	602
CIDAR	-	6,000	-	-	-	-	-
Cendol v2	-	-	-	-	-	3,000	-
OpenAssistant 2	-	23	4	2,328	8,785	3	306
EuroBlocks SFT	-	-	3,813	12,551	15,641	-	2,893
GSM8K (train)	7,473	-	-	-	-	-	-
Helpsteer3 (chosen)	-	-	-	462	778	156	534
Magpie Pro Filtered	10,000	-	-	-	-	-	-
Total per language	37,473	10,683	10,083	21,490	34,958	13,928	10,594

Table 7: **Seed dataset statistics.** In order to bootstrap our synthetic data generation methods, we use a seed dataset composed of various multilingual instruction-following datasets. We include English samples in order to simulate data generation pipelines where English is translated into a target language. We collect a total of **139,209 seed examples** across 7 languages (including English).

Model Name	Provider	Size (B)	# Langs	License
GPT-4o mini (OpenAI et al., 2024)	OpenAI	–	50+	Proprietary
Llama 3.1 70B Instruct (Grattafiori et al., 2024)	Meta	70	8	Llama 3.1
Llama 3.1 8B Instruct (Grattafiori et al., 2024)	Meta	8	8	Llama 3.1
Command A (Team Cohere et al., 2025)	Cohere	104	23	CC-BY-NC-4.0
Aya Expanse 32B (Dang et al., 2024)	Cohere	32	23	CC-BY-NC-4.0
Gemma 3 27B Instruct (Gemma Team et al., 2025)	Google	27	100+	Gemma
Gemma 3 12B Instruct (Gemma Team et al., 2025)	Google	12	100+	Gemma
Gemma 3 4B Instruct (Gemma Team et al., 2025)	Google	4	100+	Gemma
IBM Granite 4.0 (Granite Team, IBM, 2025)	IBM	3	116	Apache 2.0
IBM Granite Micro (Granite Team, IBM, 2025)	IBM	0.4	116	Apache 2.0

Table 8: **Teacher model details.** We evaluate 10 teacher models across different providers, sizes, multilingual capabilities, and licensing terms. Size is reported in billions of parameters (B) where available. # Langs indicates the number of languages the model was trained on or evaluated for.

F Full Results for Intr. and Extr. Metrics

Table 12 shows all the data quality metrics for each teacher model across all languages. Table 13 shows the full results of student models finetuned on synthetic datasets generated by each teacher model across all target languages.

Percentage Increase Tables We provide additional tables from the main experiments in §3 and §4. Table 15 shows the percentage increase in PG-SCORE when using the best data generation method for each teacher-language pair compared to an equal mix baseline (see §3.3).

G Additional Experiments and Ablations

In this section, we ablate several aspects of our evaluation protocol that may affect a teacher model's PG-SCORE.

G.1 Effect of Data Scale on Student Model Performance

One component of PG-SCORE is the extrinsic student performance metric (§2.4) as measured by PGR. Scaling laws suggest that this performance improves with more data (Kaplan et al., 2020). Then, it is possible to inflate PG-SCORE by simply using more synthetic data. In order to control for this variable, we conduct an experiment to determine how much synthetic data is needed to reliably compute PG-SCORE.

Setup We finetune an OLMo 3 7B base model on n SFT instances where $n \in \{1\text{k}, 5\text{k}, 10\text{k}, 25\text{k}, 50\text{k}\}$. To reduce computational costs, we perform this experiment only on a single teacher model (Gemma 3 27B Instruct) and three target languages that represent diverse scripts and resource availability: Arabic, German, and Indone-

Language	Family	Script	Resource Availability	% in CC
Arabic	Afro-Asiatic	Arabic	5 (High)	0.65%
Czech	Indo-European	Latin	4 (Medium-High)	0.99%
German	Indo-European	Latin	5 (High)	6.01%
Spanish	Indo-European	Latin	5 (High)	4.37%
Indonesian	Austronesian	Latin	3 (Medium)	0.95%
Japanese	Japonic	Japanese	5 (High)	5.20%

Table 9: **Target language details.** We evaluate teacher models across six typologically diverse languages spanning different language families and scripts. Resource availability is based on the classification from Joshi et al. (2020), ranging from 0 (lowest) to 5 (highest). CommonCrawl percentages (Raffel et al., 2020) indicate the proportion of web text available for each language.

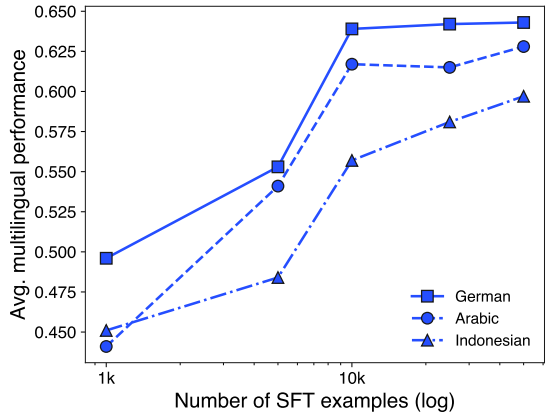


Figure 6: **Effect of synthetic data scale on student model performance.** Performance improves with more synthetic data, but gains diminish beyond 10k examples.

sian. Similar to the main experiments, we represent each data generation method equally when creating the SFT datasets. Then, we recompute the intrinsic metrics and finetune student models and measure their performance across three benchmarks (§2.4).

Results Figure 6 shows the average student model performance as a function of the number of SFT instances. We observe that student performance improves with more synthetic data, but gains diminish beyond 10k examples. This finding suggests that **using 10k synthetic examples per language is sufficient to reliably compute PG-SCORE without inflating the metric** by increasing the number of samples. In our main experiments, we use 10.5k synthetic examples per language (3.5k per data generation method) when computing PG-SCORE. Specifically, we show that 10k synthetic examples from a strong teacher are sufficient to finetune a student model to achieve reasonable performance across multiple benchmarks.

G.2 Generalization Across Model Size

Setup In order to test whether PG-SCORE generalizes beyond 7B parameter size models, we use an OLMo 3 32B base model (S_{base}) and recompute the intrinsic and extrinsic metrics to obtain the PG-SCORE. To reduce computational costs, we train student models across three teachers (Gemma 3 27B Instruct, Aya Expanse 32B, Llama 3.1 70B Instruct) and all 6 target languages.

Results Table 16 shows the PG-SCORE scores for three teacher models when using OLMo 3 32B as the student model. We find that **Gemma 3 27B Instruct remains the highest-scoring teacher in this comparison**, achieving the highest average PG-SCORE of 0.805 across all languages. This result is consistent with our findings using the 7B student model (§3), demonstrating that the superior data quality generated by Gemma 3 27B generalizes across model scales. Aya Expanse 32B achieves a positive average PG-SCORE of 0.227, while Llama 3.1 70B Instruct shows a negative average of -0.267.

Furthermore, the **language-dependent effects observed in the 7B experiments remain consistent at the 32B scale**. German continues to show the highest PG-SCORE values across all three teachers (2.389 for Gemma, 1.979 for Aya, 0.838 for Llama), suggesting that certain languages benefit more from synthetic data regardless of student model size. Similarly, Spanish exhibits strong performance across all teachers, with PG-SCORE values ranging from 1.353 to 1.855. In contrast, Arabic shows the most variable results, with Gemma achieving slightly negative scores (-0.239) while Aya and Llama show substantially lower performance (-0.872 and -1.688, respectively). Overall, these findings demonstrate that PG-SCORE and

Model	Arabic (ar)				Czech (cs)				German (de)			
	d_x	d_y	PPL	R	d_x	d_y	PPL	R	d_x	d_y	PPL	R
GPT-4o mini	0.704	0.869	8.40	3.516	0.643	0.862	3.18	3.716	0.732	0.889	3.65	3.810
Llama 3.1 70B Inst.	0.701	0.875	7.00	2.719	0.654	0.889	3.18	3.327	0.707	0.892	3.22	3.396
Llama 3.1 8B Inst.	0.708	0.779	6.2e4	1.731	0.673	0.799	2.7e4	1.908	0.738	0.873	3.6e3	2.513
Command A	0.690	0.846	5.41	3.996	0.647	0.865	3.24	4.184	0.730	0.889	3.59	4.235
Aya Expanse 32B	0.693	0.888	4.34	3.964	0.650	0.884	3.15	4.133	0.700	0.902	3.44	4.140
Gemma 3 27B Inst.	0.717	0.890	4.40	3.932	0.675	0.885	3.77	4.342	0.731	0.898	3.96	4.260
Gemma 3 12B Inst.	0.721	0.864	4.43	3.774	0.676	0.882	3.88	4.266	0.751	0.899	4.06	4.203
Gemma 3 4B Inst.	0.728	0.869	5.52	3.470	0.682	0.883	3.87	4.127	0.744	0.898	3.96	4.103
IBM Granite 4.0	0.704	0.829	1.9e4	2.463	0.665	0.862	5.29	3.158	0.717	0.885	24.61	3.365
IBM Granite Micro	0.741	0.863	12.45	3.033	0.713	0.874	4.61	3.568	0.726	0.892	4.59	3.704

Model	Spanish (es)				Indonesian (id)				Japanese (ja)			
	d_x	d_y	PPL	R	d_x	d_y	PPL	R	d_x	d_y	PPL	R
GPT-4o mini	0.729	0.887	3.78	3.883	0.728	0.854	5.50	3.656	0.736	0.880	5.81	3.639
Llama 3.1 70B Inst.	0.728	0.892	3.15	3.434	0.727	0.874	4.85	3.293	0.756	0.799	4.52	2.459
Llama 3.1 8B Inst.	0.744	0.898	503.0	2.860	0.738	0.863	1.1e3	2.599	0.759	0.796	5.4e4	1.806
Command A	0.733	0.884	3.77	4.336	0.747	0.857	4.94	3.899	0.739	0.881	4.92	4.174
Aya Expanse 32B	0.724	0.893	3.67	4.181	0.726	0.879	4.43	4.017	0.743	0.883	5.96	3.821
Gemma 3 27B Inst.	0.768	0.903	4.30	4.266	0.740	0.854	5.49	4.057	0.765	0.875	5.90	3.956
Gemma 3 12B Inst.	0.763	0.895	4.14	4.193	0.762	0.851	5.84	3.958	0.756	0.885	5.78	4.017
Gemma 3 4B Inst.	0.754	0.887	4.52	4.021	0.760	0.851	6.46	3.657	0.794	0.875	6.45	3.656
IBM Granite 4.0	0.743	0.882	5.22	3.309	0.729	0.833	16.80	2.437	0.761	0.849	9.79	2.889
IBM Granite Micro	0.729	0.887	4.58	3.779	0.760	0.860	11.92	3.113	0.764	0.877	7.22	3.295

Table 12: **Full intrinsic evaluation results across all languages.** Data quality metrics include the diversity of prompts and responses (d_x and d_y), average perplexity of the base model on the response (PPL), and average reward score based on a multilingual LLM judge (R).

teacher model rankings generalize to the 32B parameter range.

G.3 Effect of Translation Method (Prompting an LM vs. Translation Model)

An alternative to using an LM for translating texts from English to a target language is via a translation model such as NLLB (NLLB Team et al., 2022). In this section, we examine the effect of the translation method on the PG-SCORE of teacher models.

Setup First, we filter and sample 10k English prompt-response pairs from the Tülu 3 SFT dataset.[3] Then, using the NLLB model (`nllb-200-distilled-600M`), we perform two translation methods: (1) *NLLB-Translate-then-Respond:* translate the prompts to each target language and prompt Gemma 3 27B Instruct to generate a response, and (2) *NLLB-Translate-Both*: translate both the prompts and responses from English to the target language. We choose the 600M version due to its computational efficiency and popularity among practitioners, as measured by HuggingFace downloads and community likes.

We compare these methods against our original *Translate* method, i.e., prompting Gemma 3 27B Instruct to directly translate the prompt and generate the response in the target language (*LM-Translate*). Then, we compute the intrinsic data quality metrics and finetune OLMo 3 7B student models on each synthetic dataset to compute PG-SCORE.

Results Figure 7 shows the PG-SCORE and average benchmark performance of the student model for each translation method across Arabic, German, and Indonesian. We find that *LM-Translate* outperforms both NLLB-based approaches, achieving an average PG-SCORE of 1.36 compared to 0.85 for *NLLB-Translate-Both* and 0.80 for *NLLB-Translate-then-Respond*. This pattern holds across all three languages, with the largest gap observed for German (2.09 vs. 1.26/1.68).

Our findings suggest that **prompt naturalness, rather than response quality, is a bottleneck in translation-based pipelines**: having an LM generate responses to NLLB-translated prompts provides no improvement over pure NLLB translation (0.80 vs 0.85), indicating that translated prompts fail to elicit the same quality of responses as LM-

[3] Tülu 3 also contains non-English data. We perform English-language filtering using fastText (Joulin et al., 2016, 2017) and the `staticvectors` library.

Model	Arabic (ar)	Czech (cs)	German (de)	Spanish (es)	Indonesian (id)	Japanese (ja)
GPT-4o mini	−2.086	0.538	3.098	1.395	2.025	0.099
Llama 3.1 70B Inst.	−1.528	0.538	2.265	1.075	0.329	0.013
Llama 3.1 8B Inst.	−0.841	0.525	2.623	0.595	1.425	0.236
Command A	−2.476	0.505	2.759	1.613	1.863	0.841
Aya Expanse 32B	−0.293	0.538	2.491	1.701	1.943	0.221
Gemma 3 27B Inst.	−0.074	0.552	2.635	1.724	0.198	0.677
Gemma 3 12B Inst.	−1.015	0.538	2.700	1.592	−0.017	0.524
Gemma 3 4B Inst.	−1.033	0.538	2.568	1.209	−0.388	0.349
IBM Granite 4.0	1.565	0.538	2.061	1.235	0.614	0.802
IBM Granite Micro	−0.421	0.538	1.842	1.203	−0.659	0.210

Table 13: **Average performance gap recovered (PGR) of a student model across various multilingual benchmarks.** Our multilingual evaluation suite includes Global-MMLU Lite (Singh et al., 2025), M-RewardBench (Gureja et al., 2025), and M-GSM (Shi et al., 2023). The PGR computation is based on Kim et al. (2025) and detailed in §2.4 (Equation 3) where S_{ref} = OLMo 3 7B Instruct SFT and S_{base} = OLMo 3 7B.

Model	Arabic (ar)	Czech (cs)	German (de)	Spanish (es)	Indonesian (id)	Japanese (ja)
Gemma 3 27B Inst.	**0.145 (0.0121)**	**0.360 (0.0004)**	1.655 (0.0141)	**1.358 (0.0141)**	0.214 (0.0167)	0.626 (0.0124)
Aya Expanse 32B	−0.058 (0.0116)	0.222 (0.0004)	1.468 (0.0134)	1.129 (0.0123)	**1.153 (0.0124)**	0.320 (0.0111)
Gemma 3 12B Inst.	−0.464 (0.0119)	0.327 (0.0004)	1.756 (0.0137)	1.228 (0.0140)	0.151 (0.0126)	0.573 (0.0142)
Command A	−1.360 (0.0112)	0.114 (0.0004)	1.673 (0.0139)	1.102 (0.0145)	1.063 (0.0125)	**0.683 (0.0122)**
Gemma 3 4B Inst.	−0.488 (0.0119)	0.330 (0.0004)	1.644 (0.0137)	0.929 (0.0140)	−0.105 (0.0126)	0.504 (0.0113)
GPT-4o mini	−1.117 (0.0117)	0.015 (0.0004)	**1.766 (0.0136)**	0.908 (0.0149)	1.003 (0.0125)	0.189 (0.0117)
IBM Granite 4.0	−0.072 (0.0123)	−0.031 (0.0004)	1.000 (0.0135)	0.734 (0.0151)	−0.079 (0.0125)	0.321 (0.0108)
IBM Granite Micro	−0.282 (0.0121)	0.290 (0.0004)	1.102 (0.0139)	0.783 (0.0133)	−0.329 (0.0126)	0.264 (0.0121)
Llama 3.1 70B Inst.	−0.964 (0.0117)	0.109 (0.0004)	1.195 (0.0146)	0.688 (0.0146)	0.182 (0.0126)	−0.373 (0.0116)
Llama 3.1 8B Inst.	−1.693 (0.0120)	−0.974 (0.0004)	0.891 (0.0148)	0.182 (0.0164)	0.322 (0.0124)	−0.863 (0.0129)

Table 14: **Detailed results from Table 1 with standard errors.** We compute PG-SCORE thrice with different synthetically generated data (each trial uses a different data mix based on a random seed). We report the mean and standard error for each teacher model across all target languages. For each language, we highlight the best model in **bold** and the second-best model with an underline.

translated prompts.

G.4 Weighting of Intrinsic and Extrinsic Metrics in PG-SCORE

Our PG-SCORE formulation uses an assumption-free and equal weighting scheme between the intrinsic ($\mathcal{I}$) and extrinsic ($\mathcal{E}$) metrics. In this section, we test whether these two metrics capture (1) complementary aspects of teacher effectiveness and (2) how model rankings differ if one metric is weighted more than the other.

Setup In order to test whether each metric captures complementary aspects of teacher effectiveness, we compute the Spearman rank correlation (ρ) between the intrinsic and extrinsic metrics across all teacher-language pairs (N=60, 10 models × 6 languages). In addition, in order to test the effect of weighting one metric against the other, we formulate a generalized version of PG-SCORE:

$$\text{PG-SCORE}_{T,\ell} = \alpha\mathcal{I} + (1-\alpha)\mathcal{E} \quad \text{where } 0 \leq \alpha \leq 1 \tag{5}$$

Note that the experiments in §3 and §4 assume $\alpha = 0.5$. We compute the PG-SCORE across $\alpha = \{0.00, 0.25, 0.50, 0.75, 1.00\}$ and then test the resulting model ranks' ρ across all pairs of α. We perform this experiment on all teacher-language pairs where students are finetuned from the OLMo 3 7B base model (N=60, 10 models × 6 languages).

Results Intrinsic and extrinsic metrics show a moderate positive correlation (Spearman $\rho = 0.41$, $p < 0.01$), suggesting that **data quality metrics are predictive of student performance while capturing complementary information**. This finding motivates our combined PG-SCORE computation. In addition, teacher rankings are stable for nearby weighting schemes ($\rho \geq 0.90$ for adjacent α values) as shown in Figure 8. Our finding suggests that **model rankings are robust to small changes in the weighting of intrinsic and extrinsic metrics**. Our equal weighting ($\alpha = 0.5$) balances both perspectives, correlating strongly with extrinsic-focused ($\rho = 0.89$) and reasonably with intrinsic-focused ($\rho = 0.74$) rankings.

Language	Best Method	Teacher Model (T)		
		Gemma 3 27B	Aya Expanse 32B	Llama 3.1 70B
Arabic (ar)	Respond	+453.1%	+355.2%	+77.7%
German (de)	Generate	+29.3%	+0.3%	+16.4%
Indonesian (id)	Translate	+458.9%	+39.3%	−14.8%

Table 15: **Percentage increase in PG-SCORE for best data generation method.** Percentage increase when using the best-performing data generation method compared to an equal mix baseline of all three methods (Generate, Translate, Respond). For less-resourced languages (Arabic and Indonesian), using Translate or Respond methods yields substantial improvements for most teachers, though gains are teacher-dependent.

Teacher Model	Average	Arabic (ar)	Czech (cs)	German (de)	Spanish (es)	Indonesian (id)	Japanese (ja)
Gemma 3 27B Inst.	0.805	−0.239	0.222	2.389	1.855	0.239	0.366
Aya Expanse 32B	0.227	−0.872	−0.038	1.979	1.353	−0.249	−0.809
Llama 3.1 70B Inst.	−0.267	−1.688	−0.807	0.838	1.407	−1.441	0.089

Table 16: **PG-SCORE of three teacher models (S_{base} = OLMo 3 32B).** We show that our findings generalize up to the 32B parameter range on the three teacher models we tested: (1) Gemma 3 27B maintains its position as the most effective teacher, and the (2) language-dependent effects are still apparent with German having the highest PG-SCOREs across most teachers.

G.5 Effect of language resource levels on PG-SCORE

Setup For each language, we consider the following properties drawn from prior work: CommonCrawl (CC) percentage as a proxy for presence in pretraining data (% in CC, Raffel et al., 2020), and linguistic resource availability (score from 0–5, 5 as high-resource, obtained from the LDC Catalog and the ELRA Map, Joshi et al., 2020). We compute the Spearman rank correlation (ρ) between each property and PG-SCORE across all teacher-language pairs (N=60, 10 models $\times$ 6 languages).

Results Figure 9 shows the relationship between a language's percentage in CommonCrawl and PG-SCORE. We observe a **suggestive positive trend between CommonCrawl representation and PG-SCORE** ($\rho = 0.886$, $p < 0.05$). This finding suggests that languages with greater presence in pretraining data enable teacher models to generate higher-quality synthetic data that leads to better student performance. This finding is unsurprising, but it provides empirical evidence of a structural gap that inhibits quality synthetic data generation for long-tail languages. In contrast, we do not find a significant correlation between resource availability and PG-SCORE ($\rho = 0.372$, $p = 0.468$). Our findings suggest that teacher model generation quality depends more heavily on pretraining exposure than linguistic resources. Additionally, the data sources from Joshi et al. (2020) do not reflect the current landscape: recent LMs are trained on either publicly available datasets from HuggingFace or in-house datasets. While our work includes 6 diverse languages, the sample size remains limited; we encourage future work to expand the number of languages to validate these findings.

H Disclosure on the Use of LLMs

In this work, we used generative AI tools such as Claude Opus 4.6 for drafting parts of the paper, including title ideation and proofreading. Additionally, we also used the same model to help with writing software code. We have reviewed all AI-assisted work by proper code review and unit testing to check for correctness. We also revised LLM-written text to improve readability and fit the author's writing style. We take responsibility for the final content of this work, including text/claims/artifacts produced with the aid of generative AI.

I Multilingual Synthetic Data Recipe: Case Study on Tagalog

As an application of our findings and discussion in §5, we present a case study on developing a multilingual synthetic data recipe on a held-out language: Tagalog. It is a mid-resource language (Category 3 in Joshi et al. (2020)'s taxonomy) and the standardized form of Filipino, the national language of the Philippines.

Model Name	Generation Parameters			
	Temperature	Top-p	Top-k	Max Seq Len
GPT-4o mini	0.8	0.9	–	16,384
Llama 3.1 70B Instruct	0.6	0.9	–	131,072
Llama 3.1 8B Instruct	0.6	0.9	–	131,072
Command A	0.3	–	–	128,000
Aya Expanse 32B	0.3	–	–	128,000
Gemma 3 27B Instruct	1.0	0.95	64	8,192
Gemma 3 12B Instruct	1.0	0.95	64	8,192
Gemma 3 4B Instruct	1.0	0.95	64	8,192
IBM Granite 4.0	0.0	–	–	4,096
IBM Granite Micro	0.0	–	–	4,096
Default	0.8	0.9	–	–

Table 17: **Inference settings for each teacher model.** Generation parameters are based on model provider recommendations from HuggingFace and/or official documentation. The Default row indicates parameters used when model-specific recommendations are unavailable. The "–" symbol indicates the parameter was not specified in the official recommendations.

I.1 Setup: Recipe Design and Evaluation

Data We collect Filipino seed data from various publicly available SFT datasets such as WildChat 4.8M and the Aya Collection. In addition, we also include English data from the Tülu 3 SFT dataset for the *Translate* method. Table 18 shows the statistics of the seed dataset used for Tagalog synthetic data generation. Then, we implement the following data interventions based on our findings:

- **Teacher Model:** we use Gemma 3 27B Instruct as the teacher model, as it was the best-performing model across most target languages we evaluated (§3).
- **Data Generation Method:** we use the *Translate* and *Respond* methods, as they were the best-performing methods for mid-resource languages like Indonesian (§3.3). In addition, we add a small sample of prompt-response pairs synthesized via the *Generate* method.
- **Synthetic Data Scale:** we generate 10k synthetic examples using the selected teacher and data generation method, as we found that this scale is sufficient to achieve strong student performance (Appendix G.1). However, we also test on finetuning a model with 25k synthetic examples to see if more data improves performance.
- **Student Base Model:** we finetune using the Gemma 3 4B model, as we find that family-matched teacher-student pairs yield higher PG-SCORE (§3.2).

Source	Num. Instances
TaCo Alpaca	10,000
Aya Collection	1,241
WildChat 4.8M	997
WildChat 1M	250

Table 18: **Tagalog seed dataset statistics.** In order to bootstrap the synthetic data generation recipe for Tagalog, we curate a seed dataset containing a mix of Tagalog and English prompts from various sources. The majority of the seed dataset is from the TaCo paper (Upadhayay and Behzadan, 2024).

For the purposes of this work, we designate the model finetuned on Gemma 3 4B using our synthetic recipe as 10K-Polyglot-TL, where "10K" indicates the number of SFT instances used during finetuning.

Evaluation We evaluate on FILBENCH (Miranda et al., 2025), a benchmark for LMs that includes Filipino-centric multiple-choice and generative tasks. It measures an LM's performance across four categories: classical NLP, cultural knowledge, reading comprehension, and generation, alongside an aggregated FILBENCH score.

We also compare against two data mix baselines:

1. **10K-Public**: we sample 10k Tagalog prompt-response pairs from the seed dataset. This base-

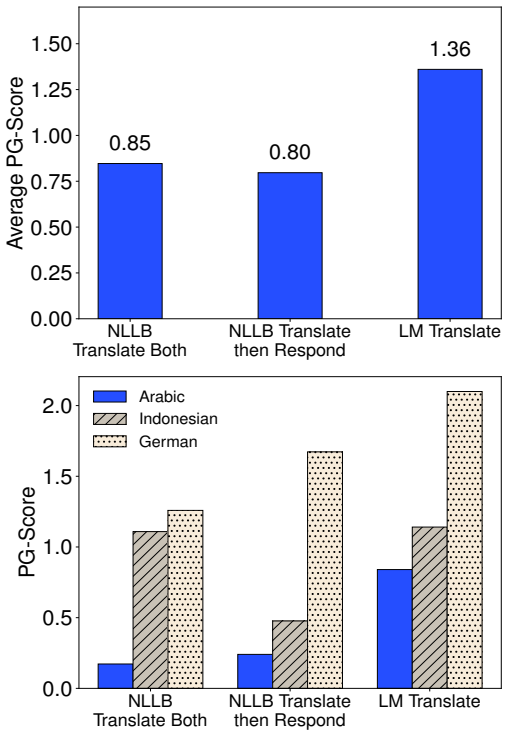


Figure 7: **Effect of translation method on PG-SCORE.** We compare three methods: LM translates prompt `EN-to-XX` and responds (*LM-Translate*), NLLB translates prompt `EN-to-XX` and LM responds (*NLLB-Translate-then-Respond*), and NLLB translates both prompt and response (*NLLB-Translate-Both*).

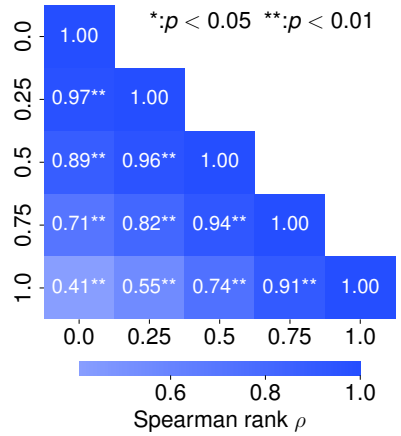


Figure 8: **Effect of weighting intrinsic and extrinsic metrics in PG-SCORE.** Model rankings remain relatively stable across neighboring weightings of intrinsic and extrinsic metrics.

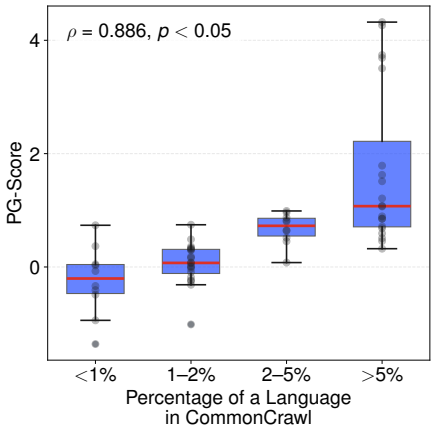


Figure 9: **Relationship between a language's percentage in CommonCrawl and PG-SCORE.** We observe a suggestive positive trend ($\rho = 0.886$, $p < 0.05$) between CommonCrawl representation and PG-SCORE across the six languages tested.

line aims to simulate a non-synthetic data approach to training multilingual LMs.

2. **10K-GPT-4oM**: we synthesize 10k instances using an off-the-shelf teacher model (GPT-4o mini). This baseline simulates a typical data generation approach of choosing a teacher in an ad hoc manner due to its perceived strength (size or benchmark performance) or ease of use.

For all methods, we finetune a Gemma 3 4B base model using the same training settings indicated in Appendix E.1.

I.2 Results: Leaderboard Scores and Ablations

Table 19 shows the FILBENCH score of our optimal synthetic recipe compared to other models in the same parameter range. We find that 10K-Polyglot-TL outperforms both 10K-GPT-4oM (+1.85 pp) and 10K-Public (+2.28 pp). These results suggest that (1) synthetic data generation is a viable approach for building models for less-resourced languages, and (2) selecting teacher models based on PG-SCORE is effective, as larger models do not always produce better training data (§3).

In addition, comparing 10K-Polyglot-TL to other models in the FILBENCH leaderboard[4] shows that our model is competitive against Qwen 3 4B and Llama 3.1 8B Instruct. We highlight that **our 4B models are competitive against other models with larger parameter sizes**, suggesting that a multilingual synthetic data recipe based on our PG-SCORE findings is data-efficient. We also find that increasing the number of SFT instances (10k to 25k) led to a performance increase of 0.21 pp. While we previously found that gains diminish beyond 10k instances (see Appendix G.1), the con-

[4]Official FILBENCH leaderboard: https://hf.co/spaces/filbench/filbench-leaderboard

Model	FILBENCH Score
GPT-4o (2024-08-06)	74.27
Gemma 3 27B Inst.	55.17
Gemma 3 12B Inst.	54.04
25K-Polyglot-TL 4B	49.73
10K-Polyglot-TL 4B	49.52
Qwen 3 4B	48.42
10K-GPT-4oM	47.67
Llama 3.1 8B Inst.	47.38
Ministral 8B Inst.	47.33
10K-Public	47.24
Pangea 7B	43.98
SeaLLMs 3 1.5B	43.20

Table 19: **Model performance on a held-out language (Tagalog) as evaluated on FILBENCH** (Miranda et al., 2025). We compare our optimal synthetic recipe against baseline approaches and other models in the same parameter range.

tinued gains from scaling to 25k instances on FILBENCH suggest that saturation points may depend on task diversity. FILBENCH covers a broader range of NLP tasks (e.g., named-entity recognition) compared to our experimental benchmarks in §3 and Appendix G, indicating that **practitioners working with diverse task distributions may benefit from exploring larger synthetic datasets beyond the 10k threshold**.

I.3 Analysis: Ablation Experiments

In order to measure the contribution of our findings and recommendations in §5, we perform the following ablation experiments as shown in Figure 5. Note that the interventions described below are additive.

Curation of publicly available data vs. synthetic data generation We compare student models trained on (1) publicly available Tagalog SFT data and (2) synthetic SFT instances generated by a GPT-4o mini teacher (note that these are also the same baselines in Appendix I.2). We find that the performance of these two baselines is similar ($\Delta = 0.43$ pp), suggesting that there is no significant advantage to using a synthetic data pipeline if the teacher model is not optimal. We also hypothesize that some publicly-accessible datasets in Tagalog were semi-synthetic (e.g., TaCo uses a synthetic pipeline akin to the *Translate* method, but using chain-of-thought to improve the quality of translations), making it difficult to perform a fair comparison.

Using a teacher with a higher PG-SCORE We then swap the GPT-4o mini teacher with Aya Expanse 32B, a teacher with a higher PG-SCORE based on our main findings (0.706 vs. 0.461, cf. §3, Table 1). We observe a slight performance improvement in this intervention, suggesting that the PG-SCORE metric is generalizable across an unseen language.

Matching teacher and student model families One of our key findings and recommendations is to match the model families of the teacher and the student (§3.2). We use a Gemma 3 27B Instruct teacher model to match the family of the Gemma 3 4B base model. This intervention yields a substantial performance improvement, consistent with our finding that family matching is a useful default for teacher selection (§3.2). We attribute this to shared tokenization schemes and architectural similarities that facilitate better knowledge transfer from teacher to student.

Increase data scale We increase the number of synthetic instances from 10k to 25k to assess whether additional data continues to improve performance. We observe a modest gain of 0.21 pp, which is smaller than the improvements from teacher model selection and model family matching. This finding aligns with our earlier observation that gains diminish beyond 10k examples (Appendix G.1), though the continued improvement on FILBENCH's diverse task distribution suggests that saturation points may be task-dependent.

Increase model scale Finally, we explore whether scaling the student model from 4B to 12B and 27B parameters provides additional performance gains. We find that the larger student model achieves higher performance, demonstrating that our synthetic data recipe benefits from increased model capacity. This result is consistent with our generalization experiments (Appendix G.2), where we showed that PG-SCORE generalizes across different model sizes while maintaining the relative ranking of teacher models. However, we note that the performance of our best models is still behind Gemma 3 27B Instruct and Gemma 3 12B Instruct (Table 19). Given that observation, we still argue that our synthetic pipeline, which uses 25k instances trained only via SFT, can be considered

data- and resource-efficient compared to the post-training interventions done in Gemma 3, which involved instruction-tuning and reinforcement learning objectives (Gemma Team et al., 2025).

I.4 Analysis: Instance-Level Qualitative Inspection of Generated Data

Setup We also inspect the quality of generated Tagalog data across teacher models with different PG-SCOREs. To do so, we sample 50 instances from a dataset generated by the following models (in increasing PG-SCORE): GPT-4o mini, Aya Expanse, and Gemma 3 27B.We evaluate these instances in terms of the following criteria:

- *Fluency*: whether the generated text reads naturally in Tagalog, with grammatical correctness and idiomatic word choice rather than literal translations or code-switched constructions that a native speaker would find awkward.
- *Cultural accuracy*: whether the content reflects appropriate Filipino cultural context, including correct references to local entities, customs, and norms, as opposed to generic or Western-centric framings that have been superficially translated into Tagalog.

The scoring is from 1–5, using the rubric:

1. **Unacceptable**: ungrammatical, code-switched, or reads as a literal translation (*fluency*); content is absent, generic, or Western-centric (*cultural accuracy*).
2. **Poor**: frequent grammatical or idiomatic errors (*fluency*); cultural references are mostly missing or incorrect (*cultural accuracy*).
3. **Adequate**: comprehensible but noticeably awkward (*fluency*); some appropriate grounding mixed with generic or Western framings (*cultural accuracy*).
4. **Good**: mostly natural with only minor lapses (*fluency*); cultural context is largely accurate with small inaccuracies (*cultural accuracy*).
5. **Excellent**: fully fluent and idiomatic, indistinguishable from native writing (*fluency*); consistently and accurately grounded in Filipino cultural context (*cultural accuracy*).

These annotations were performed by a native speaker who is one of the authors of this work. To reduce bias, the annotator only saw a prompt-response pair and its generation method without knowledge of which teacher model it was distilled from. Finally, we also include some observations by the annotator during the annotation process.

Teacher Model	Fluency	Cul. Acc.
GPT-4o mini	2.800	2.740
Aya Expanse	2.480	3.320
Gemma 3 27B	4.100	4.220

Table 20: **Native-speaker annotation of teacher-generated outputs.** We show the average across 50 instances of the native-speaker ratings on Tagalog synthetic data.

Results Table 20 shows the annotation results for the 50 instances sampled from each teacher model. We find that for Tagalog, teacher models with a higher PG-SCORE also produce higher-quality Tagalog data as judged by a native speaker. Gemma 3 27B, which has the highest PG-SCORE$_{\text{tl}}$ (1.049), achieves the strongest ratings on both criteria. This pattern is most consistent on cultural accuracy, which increases monotonically with PG-SCORE, suggesting that PG-SCORE captures aspects of cultural grounding that are perceptible to native speakers.

However, we observe that fluency does not follow the same trend: Aya Expanse receives a lower fluency score than GPT-4o mini (2.480 vs. 2.800) despite a higher PG-SCORE$_{\text{tl}}$. We also note the substantial gap between Gemma 3 27B and the other two teachers on both dimensions ($\Delta = 1.3$–1.6 on fluency, $\Delta = 0.9$–1.5 on cultural accuracy), consistent with our earlier finding that matching the teacher model family with the student is a useful default for teacher selection (§3.2).

J Inference Details

Prompt templates Figure 10 to Figure 12 show the prompt templates used for each data generation method. In addition, Figure 13 shows the prompt template used for the LLM-as-a-judge method to evaluate text quality.

Inference settings We use vLLM (Kwon et al., 2023) and Curator (Marten et al., 2025) for inference. For each teacher model, we check whether the model provider recommended settings for usage. If not, then we set a default configuration (temperature=0.8, top_p=0.9). Table 17 summarizes the inference settings we used for each teacher model.

Generate: sample k prompt-response pairs from $\mathcal{D}_{\text{seed},\ell}$ and use it as in-context examples

As a multilingual data generator, your task is to generate a new example (`'prompt'` and `'response'`) for a dataset demonstrating how AI agents can fulfill general instructions for `{lang_name}`.

To do this, you will want to generate two pieces of information:
1) A "prompt" specifying a task to be completed or a question to be answered (what, where, when, how, who, why). The task should be very challenging yet solvable.
2) A "response" representing a valid completion of that task in natural language. If the "response" does not satisfy the "prompt", then you have failed at your job. Do not provide unnecessary details, beyond what is explicitly needed to satisfy the instruction you generated.

Hard constraint: The generated task MUST belong to exactly one of the following categories (pick one at random and do NOT mention the category).
1. Logical reasoning / error analysis
2. Math or quantitative reasoning with explanation
3. Classification or labeling
4. Dialogue or role-play
5. Translation or paraphrasing with constraints
6. Procedural instructions (step-by-step)
7. Grammar correction or linguistic analysis
8. Short-form creative output ($\leq$50 words)
9. Knowledge recall with verification or correction
10. Cultural or pragmatic judgment

Add diversity to your generations by varying the types of tasks you create, the styles and tones of the responses, and the complexity of the language used. This will help ensure a rich and varied dataset. For example, you might create tasks that involve answering knowledge-based questions, answering math questions, providing explanations, generating creative content, or performing translations.

Please provide a JSON dictionary response that includes the new `'prompt'` and its corresponding `'response'`. Use the `'prompt'` and `'response'` keys in the dictionary.
Do not generate any other text in your response (for example, do not start your message with any greetings, and never ask for clarification or apologize for struggling with the task).
Try your best to ensure that the input and response you generate are distinct from the provided examples while maintaining a diverse, detailed, precise, comprehensive, and high-quality response.
It is important to generate responses that are contextually relevant and culturally appropriate for `{lang_name}`.

Here are some examples to guide your generation. The best way to use these examples is to identify the patterns and structures they follow, rather than copying them directly:

```
{% for example in examples[:k] %}
Prompt: {{example["prompt"]}}
Response: {{example["response"]}}
{% endfor %}
```

New Example:

Figure 10: Prompt template for the *Generate* data generation method.

Translate: forward-translate English prompts from $\mathcal{D}_{\text{seed},\ell}$ and use teacher T to generate the response y_i

As a multilingual data generator, your task is to translate the given prompt from English into `{lang_name}` and generate the appropriate response in the same language.
Important: you must return both the translated prompt (into `{lang_name}`) and the response. Ensure that both the translated prompt and the response are coherent, culturally appropriate, and demonstrate a deep understanding of the language nuances.

Do not generate any other text in your response (for example, do not start your message with any greetings, and never ask for clarification or apologize for struggling with the task). Do not return the original English prompt. Remember, you must translate the prompt first and return it.
Here is the prompt you need to translate and respond to:

`{prompt}`

Figure 11: Prompt template for the *Translate* data generation method.

Respond: take prompts from $\mathcal{D}_{\text{seed},\ell}$ and use teacher T to generate the response y_i

As a multilingual data generator, you will be presented a user request or instruction in the `{lang_name}` language. Your task is to generate an appropriate response for the given request. Ensure that your response is coherent, culturally appropriate, and demonstrates a deep understanding of the language nuances Do not generate any other text in your response (for example, do not start your message with any greetings, and never ask for clarification or apologize for struggling with the task). Here is the prompt you need to respond to:

`{prompt}`

Figure 12: Prompt template for the *Respond* data generation method.

LLM-as-a-judge: evaluating text quality using the multilingual rubric language model

Task Description:
An instruction (might include an Input inside it) in `{language}`, a response to evaluate, and a score rubric representing a evaluation criteria are given.
1. Write a detailed feedback that assess the quality of the response strictly based on the given score rubric, not evaluating in general.
2. After writing a feedback, write a score that is an integer between 1 and 5. You should refer to the score rubric.
3. The output should contain the score and feedback only.
4. Please do not generate any other opening, closing, and explanations.

The instruction to evaluate:
`{{instruction}}`

Response to evaluate:
`{{response}}`

Score Rubrics:
[Is the model proficient in language `{lang_name}`, including its cultural nuance and grammatical usage, and responds in a helpful and harmless manner according to the instruction?]
Score 1: The response contains severe grammatical errors, lacks cultural appropriateness, or is unhelpful/harmful. The language proficiency is very poor.
Score 2: The response has noticeable grammatical errors and limited cultural awareness. It partially addresses the instruction but with significant gaps in language proficiency or helpfulness.
Score 3: The response demonstrates adequate language proficiency with some minor grammatical errors. It shows reasonable cultural awareness and addresses the instruction in a helpful manner, though improvements are possible.
Score 4: The response exhibits strong language proficiency with minimal grammatical errors and good cultural nuance. It addresses the instruction in a helpful and harmless way with only minor room for improvement.
Score 5: The response demonstrates excellent language proficiency with proper grammar, appropriate cultural nuance, and idiomatic usage. It fully addresses the instruction in a helpful and harmless manner.

Feedback:

Figure 13: We evaluate text quality of synthesized texts using a multilingual rubric model called M-Prometheus (Pombal et al., 2025). We choose M-Prometheus due to its strong performance on multilingual and human-aligned benchmarks.