Scaling Language Models on Machine-Translated Data: Effects of Source Text Complexity on Generalization to Native Text

Anonymous ACL submission

Abstract

Pretraining on Machine-Translated text appears to be a viable alternative for pretraining language models in low-resource languages. Yet, we still lack a clear picture of how well language models scale on such noisy corpora and which properties of the source text matter. We fill this gap with a controlled study in Indonesian and Tamil. Starting from one English corpus, we build two MT datasets—Natural-MT and a Simplified-MT variant generated with an LLM—and pretrain GPT-2 models of three sizes (124M, 355M, 774M). Our results show: (1) loss on held-out native text continues to fall with model size, indicating that extra capacity learns transferable patterns despite translation noise; (2) models trained on Natural-MT consistently outperform those trained on Simplified-MT, implying that the linguistic richness of the source text survives translation and aids generalization; (3) a brief continual-pretraining phase on a modest native corpus pushes performance beyond a native-only baseline; (4) when downstream task data are also MT, MT-pretrained checkpoints match native-pretrained ones on sentiment analysis, NLI, and causal reasoning, though native exposure remains crucial for toxicity detection. Together, these findings suggest a practical recipe for data-poor languages: translate diverse English text, scale models, and devote any native data to a short adaptation phase.

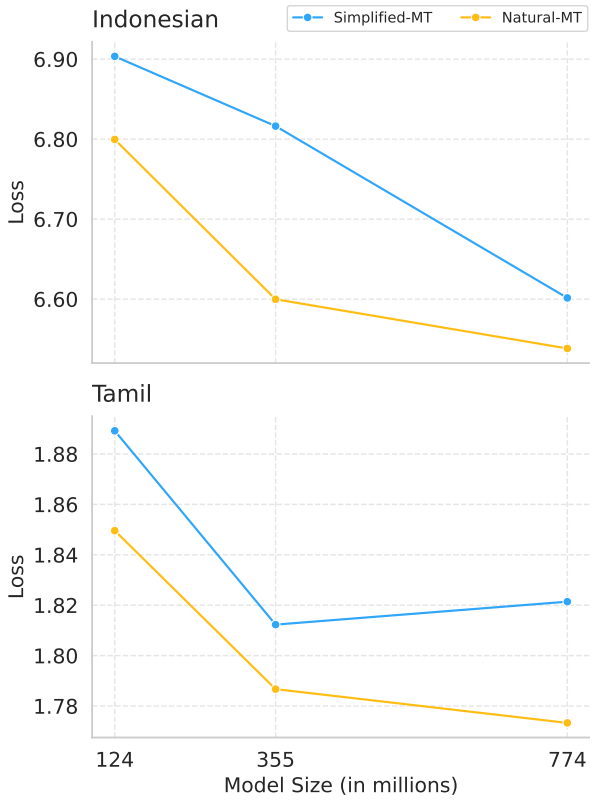

Figure 1: **(Top)** Loss vs. model size for Indonesian. **(Bottom)** Loss vs. model size for Tamil. Loss is evaluated on native text in their respective languages. In both settings, adding more parameters improves loss and models trained on Natural-MT consistently outperform those trained on Simplified-MT.

1 Introduction

Language technologies have advanced rapidly, with Large Language Models (LLMs) achieving strong performance across an array of tasks (Brown et al., 2020; Team et al., 2024; Qwen et al., 2025; Grattafiori et al., 2024). Scaling-law studies show that performance improves almost predictably with larger parameter counts and more training tokens (Kaplan et al., 2020; Hoffmann et al., 2022; Wei et al., 2022). Yet the corpora required to reap these benefits exist only for a handful of data-rich languages such as English, leaving most of the world's languages under-served (Üstün et al., 2024). A workaround is to machine-translate English text into the target language, instantly producing billions of words but also introducing "translationese"—literal phrasing, source-language bias, and cultural mismatches that diverge from native prose (Jalota et al., 2023).

Although MT has known limitations, its accessibility makes it a valuable source for supplemental

data or even as primary training data. Recent works explore the utility of machine-translated pretraining data in some Indic languages (Doshi et al., 2024) and Arabic (Alcoba Inciarte et al., 2024), both showing encouraging results of achieving downstream performance comparable with models pretrained on native texts. While these works produced encouraging results on MT-pretraining, key questions remain about its scalability and generalization.

This shift from native to synthetic data raises our first, fundamental question. **When pretraining on noisy MT data, does increasing model size improve generalization to native language—or does it lead to overfitting on translation artifacts?** We answer this by measuring how native held-out loss changes as we increase the model size under a fixed MT corpus.

To further examine the noise in MT data, we explore whether simplifying source text before translation can yield MT data with fewer errors. This assumption is grounded in the intuition that simpler language is easier to translate. However, this may also reduce linguistic diversity, yielding less expressive training data. **What is the net effect of lowering source-side complexity on generalization from MT pretraining to native text?** We isolate this factor using parallel Natural-MT and Simplified-MT corpora.

Beyond data characteristics lie training-scheme choices. Continual pretraining (CPT) is widely used for domain adaptation, but almost always starts from a clean checkpoint. **Can models pretrained on MT data serve as useful initialization for continual native-only pretraining, or do translationese artifacts hinder learning by forcing the model to unlearn unnatural patterns?** We study the drop in native loss and separately probe linguistic capabilities after a fixed CPT budget.

Lastly, fine-tuning resources for truly low-resource languages aren't always native. In practice, one takes an English task dataset, machine-translates it, and trains on it as a substitute. **Given such translation-based fine-tuning data, do MT-pretrained models ultimately outperform native-pretrained ones on native evaluation sets, or does the advantage disappear?** We test evaluate this by fine-tuning on MT task data from English and testing on four native natural language understanding (NLU) tasks.

Our contributions are as follows:

(1) We release machine-translated corpora in Indonesian and Tamil by translating both Natural and Simplified English sources.

(2) We provide insights into the generalization potential of MT data as pretraining data by observing the scaling behavior across three GPT-2 sizes (124M, 355M, 774M parameters).

(3) We quantify how source complexity, pretraining origin, and CPT affect loss on held-out native texts and linguistic probing accuracy, discovering that simplification usually hurts and MT-pretraining $\rightarrow$ native-CPT has the potential to yield much lower loss than native-pretraining alone.

(4) We evaluate downstream transfer on four NLU tasks whose training sets are machine-translated, revealing that pretraining origin matters little except in specific tasks like toxicity detection, where native exposure remains critical.

This systematic study closes key evidence gaps about *scaling, text complexity, and continual pretraining* when machine-translated data is the only realistic option.

2 Related Work

Performance gap in low-resource languages. Recent breakthroughs in LLMs have been concentrated in high-resource languages like English, where vast amounts of high-quality data are readily available (Joshi et al., 2020). In contrast, low-resource languages continue to lag significantly behind, largely due to limited training data and benchmarks. This disparity has motivated several community-driven efforts aimed at closing the gap, including initiatives like Masakhane for African languages (Orife et al., 2020), SEA-CROWD for Southeast Asian languages (Lovenia et al., 2024), and developments of multilingual open-source LLMs such as BLOOM (Workshop et al., 2023) and Aya (Üstün et al., 2024). These efforts underscore the importance of inclusive data and model development to make LLMs more accessible across languages.

Continual pretraining of language models. The core idea of continual pretraining (CPT) is to

take advantage of general patterns learned from previous pretraining regime and adapt it to the new domain. Depending on the context, domain can mean different things. For example, adapting language models from one language to another (Cahyawijaya et al., 2023; Yong et al., 2023; Joshi et al., 2025) or it can be adapting language models trained on general knowledge to specialized domains like computer science publications (Gururangan et al., 2020). In our work, we frame CPT as adapting representations learned from translationese domain to native language domain. Our setup shares similarities with Doshi et al. (2024) but our focal point is to observe loss improvements on native texts and linguistic capabilities.

Machine-Translated Data for Pretraining. Pretraining on machine-translated (MT) data has been explored in several languages, including Arabic (Alcoba Inciarte et al., 2024) and Indic languages such as Hindi, Marathi, and Gujarati (Doshi et al., 2024). These works primarily examine whether MT data can effectively bootstrap language models and match the performance of models trained on native text. In contrast, our work focuses on the properties of MT data itself, aiming to understand the conditions under which it supports effective pretraining.

3 Target Languages and Machine Translation Models

For the source language, we chose English due to its high-resourceness. For target languages, we decided based on several criteria: (1) language is not yet studied in the context of MT-pretraining (2) overall data in that language is relatively scarce, (3) availability of open-source MT model, (4) availability of high-quality human-created NLU benchmarks, and (5) presence of a diagnostic benchmark for linguistic knowledge, similar to BLiMP (Warstadt et al., 2020). All are essential for better understanding MT-pretraining's generalization potential to native text beyond language modeling performance.

For MT models, we use OPUS-MT (Tiedemann et al., 2023) English → Indonesian[1] and English → Tamil[2], achieving BLEU score of 38.7 and 4.6 on flores101-devset, respectively (opu). We use OPUS-MT due to its open-source nature[3], small model size, and fast inference.

4 Data Setup

Corpus	Words	Types	TTR	Entropy
Natural	3.72B	12.70M	0.34%	10.77
Simplified	3.45B	9.56M	0.28%	10.34

Table 1: Source-side corpus statistics. Words are space-separated words, Types are unique word count, TTR is Type-Token Ratio, and Entropy refers to Unigram Entropy. Lower TTR means lower lexical diversity. Lower Entropy means lower complexity.

Source Language Data. The source English data was curated from three permissively licensed corpora[4]: Dolma v1.6 (Soldaini et al., 2024), FineWeb-Edu (Penedo et al., 2024) and Wiki-40B (Guo et al., 2020). The combined dataset, herein referred to as Natural Corpus, contains 3.98 billion tokens, comprises of 40% Dolma (web, social media, books, academic), 10 % Wiki-40B (wiki), and 50% FineWeb-Edu (web). More details about the sampling can be found in Anonymous (2025).

Indonesian Native Corpus. We use Indo4B (Wilie et al., 2020) since it's one of the largest and most widely adopted pretraining dataset in Indonesian.

Tamil Native Corpus. We randomly sampled 5B tokens from Tamil subset of IndicMonoDoc (Doshi et al., 2024) since it's one of the largest, document-level pretraining dataset in Tamil.

Simplified Data. We use Llama 3.1 8B (Grattafiori et al., 2024) to transform Natural Corpus into simplified texts, referred to as Simplified Corpus. For efficient inference, we use the INT8 quantized version[5] of the model and vLLM (Kwon et al., 2023) as our LLM serving system. More details about the filtering and prompt can be found in Anonymous (2025). The resulting data will be referred to as Simplified Corpus. Table 1 summarizes corpus statistics and surface-level text complexity metrics, with Simplified Corpus showing consistently lower Types, TTR, and Unigram Entropy than Natural Corpus, suggesting overall lower text

[1] opus-2019-12-18 version accessed at https://huggingface.co/Helsinki-NLP/opus-mt-en-id

[2] opus-2020-07-26 version accessed at https://huggingface.co/Helsinki-NLP/opus-mt-en-dra

[3] CC-BY 4.0

[4] Dolma and FineWeb-Edu (ODC-BY), Wiki-40B (Creative Commons)

[5] https://huggingface.co/neuralmagic/Meta-Llama-3.1-8B-Instruct-quantized.w8a8

complexity. Here's an example of what the simplified texts look like:

> **Original**: Maintaining a relaxed state of mind allows you to approach challenges with clarity and calm, making it easier to find balanced solutions.
>
> **Simplified**: Staying calm helps you face challenges more clearly and find better solutions.

Machine-Translated Data. Before MT, we split Native Corpus into sentences and apply pre-MT filtering: drop any document that contains a sentence exceeding a given token-count threshold. This is done simply for efficiency purposes. More details in Appendix A.

After MT, we apply post-MT filtering: calculate sentence length ratio (in tokens) of translation to source text then drop documents in which any translation exceeds a sentence-length ratio of 2.

After filtering, sentences are reconstructed back to documents. To control for core text content of the corpus, we ensure that all documents in Natural Corpus and Simplified Corpus are parallel. The final translated corpus will be referred to as Natural-MT Corpus if source data is Natural Corpus and Simplified-MT Corpus if source data is Simplified Corpus.

4.1 Evaluation and Fine-tuning Data

Our evaluation touches on three aspects: (1) out-of-distribution generalization to native text, (2) native-language proficiency, and (3) native-language downstream performance.

Aspect (1): Out-of-distribution generalization to native text. We set a held-out validation set comprising 200 million tokens from the native corpus of each language.

Aspect (2): Native-language proficiency. We use the syntax subset of LINDSEA (Leong et al., 2023), which probes phenomena such as morphology, negation, argument structure, and filler-gap dependencies.

Aspect (3): Native-language downstream performance. We evaluate on the Indonesian and Tamil subsets of SEA-HELM (Susanto et al., 2025) for four NLU tasks: sentiment analysis, toxicity detection, natural-language inference, and causal reasoning.

4.2 Fine-tuning Data

To investigate the generalization potential of machine-translated fine-tuning data, we train on machine-translated English task-specific datasets. The list of datasets is summarized in Table 2. All task datasets, except for causal reasoning, will go through balanced-label sampling → pre-MT filtering → MT → post-MT filtering. For more details, refer to Appendix A.

Task	Train Data	Labels
SA	Amazon (Hou et al., 2024) Yelp (Zhang et al., 2015)	negative (50K) positive (50K)
TD	HateSpeech (Davidson et al., 2017)	hate (0.6K) clean (2.4K) rough (10.3K)
NLI	WANLI (Liu et al., 2022)	contradiction (11.2K) entailment (10.9K) neutral (11K)
CR	B-COPA (Kavumba et al., 2019)	–

Table 2: Overview of fine-tuning tasks, data sources, label splits, and example counts (in thousands). SA = Sentiment Analysis, TD = Toxicity Detection, NLI = Natural Language Inference, CR = Causal Reasoning.

5 Experimental Setup

5.1 Model Architecture

Size	Layers	d_{model}	Heads	MLP	Params
Small	12	768	12	3 072	124 M
Medium	24	1 024	16	4 096	355 M
Large	36	1 280	20	5 120	774 M

Table 3: Model configurations.

All models are plain GPT-2 decoders trained from scratch. We train a 50,257-token byte-pair encoding (Sennrich et al., 2016) separately for Indonesian and Tamil on their respective native corpora. Two additional special tokens are introduced: a `[PAD]` token serving as both padding and end-of-sequence, and a `[SEP]` token used only for sequence-pair classification heads. During pretraining we attach a language-model head; downstream experiments swap in a linear classification head.

5.2 Pretraining Configurations

Models. For each language, we train nine models:

$$(3 \text{ corpora: Natural-MT, Simplified-MT, Native}) \times (3 \text{ sizes: Small, Medium, Large})$$

Optimization. We optimize left-to-right language modeling with a 1 024-token window and an effective batch size of 384. AdamW is used with the default $\beta_1/\beta_2/\varepsilon$ (0.9/0.999/1e–8) and weight-decay 0.01. A 100M-token sweep over {5e–5, 1e–4, 5e–4} showed 5e–4 to be consistently strongest; this value is fixed with 5 % warm-up and linear decay. Mixed precision (autocast + GradScaler) and gradient clipping (1.0) are enabled; Large models use gradient checkpointing to work around memory constraints.

Hardware and runtime. Small/Medium models train on 8 × P100 (16GB); Large models on 8 × P40 (24GB). Wall-clock times range from 19h (Indonesian Simplified-MT, Small) to 12d 11h (Tamil Simplified-MT, Large).

5.3 Continual Pretraining (CPT)

Continual pretraining is applied to the small and medium models that were first trained on the Natural-MT and Simplified-MT corpora. Each run restarts from its final MT checkpoint and continues on the Native corpus: 1B tokens for Indonesian and 2.5B tokens for Tamil, which correspond to roughly half of the respective MT token counts. All optimization hyperparameters of the first stage are retained— AdamW with default moments, an effective batch size of 384 (1 024-token sequences), mixed precision, and gradient clipping— except that the peak learning rate is reduced by an order of magnitude to 5×10^{-5} while maintaining a 5 % warm-up and linear decay. This second stage therefore adapts MT-initialized representations to genuinely native data without altering the overall training dynamics.

5.4 Fine-tuning & Evaluation

Supervised tasks. Each pretrained checkpoint is fine-tuned on four downstream classification tasks—*causal reasoning*, *sentiment analysis*, *natural-language inference* (NLI), and *toxicity detection* (for Indonesian only). Training data consist solely of machine-translated (translationese) instances; a held-out portion of each task's translationese set serves as a development split for hyperparameter selection. After grid search, the best configuration for each random seed is evaluated on the native SEA-HELM test set.

Classification head and optimization. Fine-tuning attaches a simple classification head on top of the decoder: a single linear layer mapping the hidden dimension to the number of class labels. The model pools by taking the logits at the final non-padding token of each sequence; cross-entropy loss is computed on those pooled logits. All decoder parameters and the output layer are updated jointly. We sweep over learning rates $\{1\times10^{-4}, 5\times10^{-5}, 2\times10^{-5}, 1\times10^{-5}, 5\times10^{-6}\}$ and task-dependent epoch budgets (1–3 epochs for causal and toxicity, 1–2 for NLI, and exactly one epoch for sentiment). The maximum sequence length remains 1024 tokens; batch sizes are tuned to the longest example in each corpus (see Table 6 in the Appendix). All runs follow the pretraining schedule of 5 % warm-up followed by linear decay and employ no early stopping.

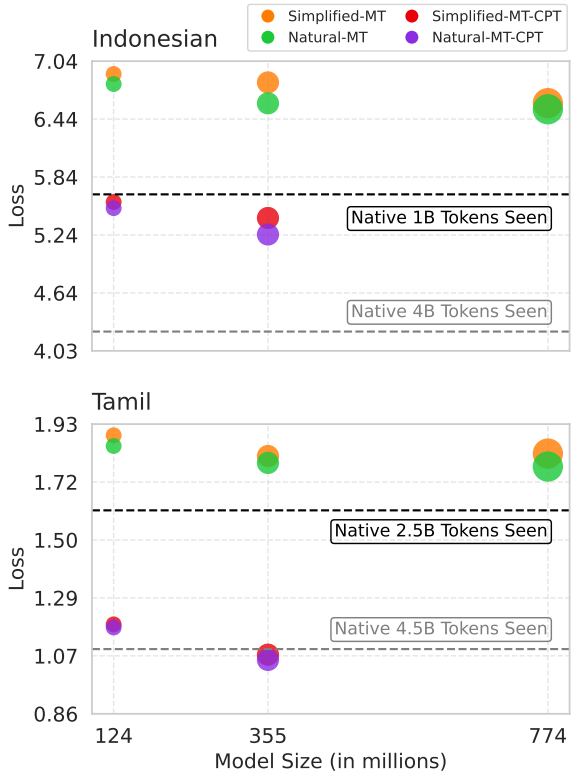

Figure 2: **(Top)** Loss vs. model size for Indonesian. **(Bottom)** Loss vs. model size for Tamil. Loss is evaluated on native text in their respective languages. In both cases, continual pretraining of Natural-MT and Simplified-MT on partial native corpus (1B tokens for Indonesian, 2.5B for Tamil) significantly reduces loss, surpassing the native model with equal native token exposure. Dashed lines show the lowest loss achieved by the largest native model.

Metric and model selection. Systems are ranked by **balanced accuracy** on the dev split, where balanced accuracy is the unweighted mean of per-class recall (macro-averaged). Final scores are

Size	Pretraining data	Indonesian		Tamil	
		Acc.	Δ	Acc.	Δ
Small	Native	**55.8**		71.5	
	Natural-MT	47.6		66.2	
	Natural-MT → Native-CPT	52.9	+5.3	69.1	+2.9
	Simplified-MT	46.6		61.3	
	Simplified-MT → Native-CPT	52.4	+5.8	**72.1**	+10.8
Medium	Native	52.4		62.8	
	Natural-MT	50.5		65.5	
	Natural-MT → Native-CPT	**53.7**	+3.2	72.8	+7.3
	Simplified-MT	49.5		65.1	
	Simplified-MT → Native-CPT	52.1	+2.6	**76.0**	+10.9
Large	Native	**57.4**		**68.9**	
	Natural-MT	49.7		62.8	
	Simplified-MT	49.7		62.8	

Table 4: Accuracy on the LINDSEA Syntax subset (higher is better; random chance is 50 %). Native pretraining produces the strongest Indonesian model (57.4%), whereas CPT lifts MT models to the top for Tamil (76.0% for Medium Simplified-MT → Native). In Indonesian, MT models score close to or below random, but CPT raises them by 4–6 percentage points, partially closing the gap to native. Tamil results are uniformly higher: even MT-only models exceed 60%, and CPT adds another 6–10 percentage points. Scaling remains non-monotonic—Simplified-MT CPT Medium surpasses all Large models in Tamil.

reported as means—and, where noted, standard deviations—over three independent random seeds.

Zero-shot syntactic probing. To gauge the linguistic knowledge encoded by the pretrained representations, we also evaluate every checkpoint on the Syntax subset of LINDSEA. The subset is converted to BLiMP-style minimal pairs; a model is correct when it assigns a higher log-probability to the grammatical member of the pair. Accuracy is averaged across all syntactic phenomena.

Compute budget. Fine-tuning uses the same hardware as pretraining: 8 × P100-16 GB for small and medium models, and 8 × P40-24 GB for large. A complete grid search for a single model across all tasks finishes in roughly 5h (small), 11h (medium), and 20h (large).

6 Results and Discussion

6.1 Scaling on Noisy Supervision

MT data often contains artifacts like translation errors and unnatural phrasing. We hypothesized that increasing model capacity would lead to overfitting on this noise, causing the model to memorize artifacts rather than learn transferable patterns. However, on a fixed MT data setup, our experiment shows that Natural-MT benefits from model scaling despite the inherent noise of MT data (see Figure 1). The lack of overfitting suggests that the model isn't merely learning surface-level MT artifacts. Rather, it's acquiring representations that transfer to native data.

6.2 Source-Side Complexity Effects

We hypothesized that source text complexity would influence how learnable the resulting MT data is, though the direction of that influence was unclear. To test this, we compare models trained on Natural-MT versus Simplified-MT. Across all sizes, Natural-MT consistently achieves lower native validation loss than Simplified-MT (see Figure 1), confirming that richer source text yields more learnable MT data.

The largest advantage appears in the Tamil small model on LINDSEA—a gap of 6.2 percentage points—while for larger models and for Indonesian the difference shrinks to under two points. A similar pattern holds on downstream NLU tasks: accuracy gaps rarely exceed two percentage points, with the sole exception of the Indonesian medium NLI system, where Natural-MT leads by 4.9 points. That substantial perplexity improvements translate into only modest accuracy gains suggests that downstream performance depends on factors beyond native-like fluency.

We hypothesize that the simplification step produces English that is less natural, resulting in MT data with weaker signals for morphology, syntax,

Model size	Pretraining data	Indonesian				Tamil		
		Causal Reasoning	Sentiment Analysis	NLI	Toxicity Detection	Causal Reasoning	Sentiment Analysis	NLI
Small	Native	**54.5**	63.4	53.7	**52.6**	50.8	87.1	42.8
	Natural-MT	51.6	61.9	56.9	42.5	48.8	88.4	42.3
	Simplified-MT	51.2	61.3	56.2	44.5	**51.3**	88.8	40.7
	Natural-MT → Native-CPT	51.2	**63.5**	57.4	47.6	50.9	88.9	**43.5**
	Simplified-MT → Native-CPT	49.4	62.9	**58.2**	49.6	50.0	**89.0**	43.0
Medium	Native	51.5	62.7	57.7	**53.0**	50.8	84.8	41.1
	Natural-MT	49.6	62.6	**60.7**	44.1	**53.7**	90.3	43.8
	Simplified-MT	47.7	61.6	55.8	44.6	51.9	90.6	44.8
	Natural-MT → Native-CPT	51.9	**64.2**	59.7	49.5	50.9	**91.2**	**45.1**
	Simplified-MT → Native-CPT	**53.4**	62.6	57.2	48.3	50.7	90.5	**45.1**
Large	Native	51.5	**63.7**	56.6	**54.7**	**51.9**	86.2	43.4
	Natural-MT	**54.8**	62.6	61.6	45.2	50.9	**90.6**	**43.6**
	Simplified-MT	52.7	61.5	**63.2**	46.2	49.0	90.0	43.3

Table 5: Average balanced accuracy on the SEA-HELM test sets after fine-tuning each model on translationese over three random seeds (best runs). Across languages and sizes, several patterns stand out. (1) **Toxicity detection** clearly favours native-pretraining; MT-pretrained models lag by 3–11 percentage points, despite identical fine-tuning data. (2) For **NLI** and **sentiment**, the gap between MT-pretrained models and native-pretrained models is narrow; continual pretraining (CPT) usually nudges performance to the top of each size block. (3) **Causal reasoning** (binary choice) remains the hardest task: the best Indonesian model reaches only 54.8%, and Tamil peaks at 53.7%, close to chance for a two-way decision. (4) **Model size is not strictly monotonic**: medium-sized models match or surpass large ones on three of the seven task–language pairs.

Standard deviations over the three random seeds are reported in Table 7 in the appendix.

and discourse. Together, these observations show that preserving linguistic variety before translation consistently improves learnability, although the magnitude of this benefit varies with language and model capacity.

6.3 MT Pretrain → Native CPT

If the model is trained on noisy MT data and learned to model translation artifacts, will it struggle to adapt to native language? Our experiments show that continually pretraining Natural-MT and Simplified-MT models on a subset of native texts (1B tokens for Indonesian, 2.5B for Tamil) significantly improves loss on native texts, even surpassing the native model under equal native token exposure (see Figure 2). This also boosted syntactic accuracy on LINDSEA by up to six percentage points in Indonesian and ten in Tamil. Our findings suggest that despite known translation artifacts, MT-pretraining learns robust, adaptable representations, making it an effective initialization for continual pretraining on native text.

6.4 Translationese Fine-Tuning Outcomes

Fine-tuning exclusively on translationese task data still transfers well to native evaluation sets (Table 5). In fact, MT-pretrained models perform comparably with native-pretrained models—balanced-accuracy gaps are usually within two percentage points—highlighting MT-pretraining as a viable alternative for bootstrapping models in low-resource scenarios. Three consistent observations emerge:

1. **Task sensitivity.** Indonesian toxicity detection clearly prefers native pretraining, with gaps of 3–11 percentage points.

2. **General parity elsewhere.** On sentiment analysis and NLI, MT-pretrained models lie within one to two percentage points of native models, and MT-pretrained → native-CPT variants often top each size block.

3. **Capacity trade-offs.** Although validation loss still falls with size (§6.1), balanced-accuracy gains taper: medium models match or exceed large ones on four of seven task–language pairs.

This tapering may reflect limitations of the translationese fine-tuning data—once the dataset lacks sufficient size or diversity, extra model capacity cannot be fully utilized. Nevertheless, translationese fine-tuning endows both MT-pretrained and native-pretrained models with robust knowledge that generalizes to native test inputs. Outside of toxicity detection, relying on MT data in either pretraining or fine-tuning does not seem to incur a systematic drawback.

7 Conclusion

This study asked whether language models can be scaled effectively on machine–translated (MT) corpora, how the linguistic complexity of the source text shapes that outcome, and whether MT–pretrained checkpoints remain useful once

native data or translationese fine-tuning becomes available. Through controlled experiments in Indonesian and Tamil, four clear messages emerge:

1. **Scaling on MT data works.** Larger GPT-2 models (124M → 774M) trained purely on Natural-MT corpora achieve lower perplexity on held-out native text than their smaller counterparts—evidence that added capacity captures transferable patterns rather than merely memorizing translation artifacts. This suggests that despite its noise, MT data contains transferable features that larger models can effectively learn.

2. **Keep the source text rich.** Simplifying the English input before translation consistently degrades downstream generalization. Linguistic variety in the source provides signals that survive translation and remain valuable for learning.

3. **MT checkpoints are excellent springboards.** Continual pretraining on even a fraction of native text (1B tokens in Indonesian, 2.5B in Tamil) lowers native validation loss beyond what a native-only model reaches with the same budget and lifts syntactic accuracy by up to ten points.

4. **Translationese fine-tuning is usually sufficient.** When task data are also MT, models pretrained on MT match or surpass native-pretrained peers on sentiment, NLI, and causal reasoning; only toxicity detection retains a clear preference for native exposure.

Future work should probe even larger capacities, extend the analysis to additional language families, and explore adaptive source simplification strategies that balance MT quality with linguistic breadth.

Limitations

Our study has several limitations. First, text simplification using LLMs may introduce hallucinations, so Simplified-MT may deviate semantically from Natural-MT. Second, translating massive amounts of text from one language to another may carry over biases from the source language to the target language. Pretraining language models with such artifacts may reinforce those biases. Third, our fixed MT dataset and three GPT-2 model sizes (124M, 355M, 774M) limit the scope; varying both dataset and model size could yield more generalizable insights. Fourth, we only examined English as the source language, so findings may not hold when translating from other languages. Fifth, results are influenced by MT model quality; despite notable BLEU score differences, we cannot isolate the impact of translation quality due to linguistic confounds. Finally, while language and culture are intertwined, this work focuses solely on language translation without addressing cultural knowledge transfer.

References

OPUS: the open parallel corpus. https://opus.nlpl.eu/dashboard/. [Online; accessed 2025-05-10].

Alcides Alcoba Inciarte, Sang Yun Kwon, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2024. On the utility of pretraining language models on synthetic data. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 265–282, Bangkok, Thailand. Association for Computational Linguistics.

Anonymous. 2025. Paper under review.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023. InstructAlign: High-and-low resource language alignment via continual crosslingual instruction tuning. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 55–78, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Preprint*, arXiv:1703.04009.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pretraining language models using translationese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5843–5862, Miami, Florida, USA. Association for Computational Linguistics.

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.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *Preprint*, arXiv:2403.03952.

Rricha Jalota, Koel Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2023. Translating away translationese without parallel data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7086–7100, Singapore. Association for Computational Linguistics.

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.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2025. Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus: A case study for Hindi LLMs. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 50–57, Abu Dhabi. Association for Computational Linguistics.

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.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When choosing plausible alternatives, clever hans can be clever. *Preprint*, arXiv:1911.00225.

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. *Preprint*, arXiv:2309.06180.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *Preprint*, arXiv:2309.06085.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. *Preprint*, arXiv:2201.05955.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James Validad Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius Hadiwijaya, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, and 6 others. 2020. Masakhane – machine translation for africa. *Preprint*, arXiv:2003.11529.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2406.17557.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *Preprint*, arXiv:1508.07909.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xian Bin Yong, Weiqi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. Sea-helm: Southeast asian holistic evaluation of language models. *Preprint*, arXiv:2502.14301.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

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

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, and 375 others. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A Pre-MT Filtering, Batch Sizes, and Standard-Deviation Tables

The pre-MT filtering criteria avoids the need to translate excessively long sentences, which significantly reduces the overall translation time. For Indonesian, we only select sentences with token counts between 3 to 250, while for Tamil we select between 4 and 150 tokens.

Lang.	Task	Batch size
Indonesian	Causal reasoning	50
	Sentiment analysis	12
	NLI	10
	Toxicity detection	2
Tamil	Causal reasoning	10
	Sentiment analysis	2
	NLI	2

Table 6: Batch sizes used during downstream fine-tuning.

Model size	Pretraining data	Indonesian				Tamil		
		Causal Reasoning	Sentiment Analysis	NLI	Toxicity Detection	Causal Reasoning	Sentiment Analysis	NLI
Small	Native	**54.5 ± 2.8**	63.4 ± 0.4	53.7 ± 0.3	**52.6 ± 0.4**	50.8 ± 0.8	87.1 ± 0.7	42.8 ± 1.4
	Natural-MT	51.6 ± 0.9	61.9 ± 1.0	56.9 ± 1.8	42.5 ± 0.8	48.8 ± 3.3	88.4 ± 0.6	42.3 ± 0.5
	Simplified-MT	51.2 ± 1.9	61.3 ± 0.5	56.2 ± 1.2	44.5 ± 3.5	**51.3 ± 3.3**	88.8 ± 0.4	40.7 ± 0.7
	Natural-MT → Native-CPT	51.2 ± 3.1	**63.5 ± 0.5**	57.4 ± 0.8	47.6 ± 2.9	50.9 ± 0.2	88.9 ± 0.3	**43.5 ± 0.7**
	Simplified-MT → Native-CPT	49.4 ± 1.3	62.9 ± 0.7	**58.2 ± 0.4**	49.6 ± 1.0	50.0 ± 1.7	**89.0 ± 0.6**	43.0 ± 0.5
Medium	Native	51.5 ± 3.8	62.7 ± 0.2	57.7 ± 1.8	**53.0 ± 0.7**	50.8 ± 3.0	84.8 ± 0.2	41.1 ± 0.9
	Natural-MT	49.6 ± 2.8	62.6 ± 0.5	**60.7 ± 0.9**	44.1 ± 1.1	**53.7 ± 2.2**	90.3 ± 0.2	43.8 ± 0.2
	Simplified-MT	47.7 ± 2.2	61.6 ± 0.8	55.8 ± 0.4	44.6 ± 1.5	51.9 ± 3.1	90.6 ± 0.1	44.8 ± 0.9
	Natural-MT → Native-CPT	51.9 ± 3.6	**64.2 ± 0.5**	59.7 ± 0.7	49.5 ± 0.7	50.9 ± 1.5	**91.2 ± 0.5**	**45.1 ± 0.8**
	Simplified-MT → Native-CPT	**53.4 ± 1.6**	62.6 ± 0.7	57.2 ± 0.3	48.3 ± 1.6	50.7 ± 3.1	90.5 ± 0.2	**45.1 ± 0.3**
Large	Native	51.5 ± 3.7	**63.7 ± 0.5**	56.6 ± 1.1	**54.7 ± 1.9**	**51.9 ± 1.5**	86.2 ± 0.9	43.4 ± 0.8
	Natural-MT	**54.8 ± 1.6**	62.6 ± 0.3	61.6 ± 1.6	45.2 ± 1.3	50.9 ± 4.7	**90.6 ± 0.2**	**43.6 ± 1.4**
	Simplified-MT	52.7 ± 3.0	61.5 ± 0.3	**63.2 ± 1.0**	46.2 ± 0.5	49.0 ± 0.9	90.0 ± 0.4	43.3 ± 0.7

Table 7: Balanced accuracy **mean ± standard deviation** on the SEA-HELM native test sets, computed over three random seeds. For most cells the standard deviation is below 2, or even 1 percentage point, confirming that the trends discussed in Table 5 are statistically robust. The few wider spreads (≈ 2–4 percentage points) are mostly confined to the most challenging task of **causal reasoning**. Even with these broader error bands, the qualitative picture is unchanged: native-pretraining dominates toxicity, MT-CPT delivers the strongest NLI and sentiment models, causal reasoning hovers near chance, and medium-sized models occasionally surpass their large counterparts.