

Beyond Aggregate Loss: Training Dynamics and Language-Specific Scaling Laws for Multilingual Language Models

Renren Jin and Deyi Xiong*

TJUNLP Lab, School of Computer Science and Technology, Tianjin University, China
{rrjin, dyxiong}@tju.edu.cn

Abstract

Scaling laws typically characterize the performance of large language models (LLMs) through aggregate loss, potentially obscuring substantial variation across languages in multilingual pretraining. We study this variation through controlled experiments on a diverse set of languages spanning two language families, training 0.3B- and 0.5B-parameter models on up to 137B tokens. Specifically, we compare continued pretraining from a strong pretrained model with training the same architecture from scratch, examine the effects of expanding the multilingual corpus, analyze cross-family transfer through language-specific loss trajectories, and fit separate scaling laws for individual languages. Our results show that continued pretraining from a strong pretrained model leads to faster convergence and consistently stronger downstream performance than training from scratch. In contrast, reductions in training loss and more stable gradient ℓ_2 norms do not necessarily translate into downstream benchmark gains. We further demonstrate that training on a single language family can reduce loss for languages in another family, providing evidence of cross-family transfer. Finally, we find that data-scaling behavior is relatively consistent across languages, whereas the benefits of model scaling vary substantially. A sampling-weighted combination of the language-specific scaling laws closely reconstructs the aggregate training loss. These findings show that aggregate metrics provide an incomplete account of multilingual pretraining dynamics and highlight the importance of language-specific analyses for evaluating and scaling multilingual models.

1 Introduction

Scaling laws have become a central tool for predicting how the performance of LLMs varies with model size, training data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022). However, most

existing scaling analyses summarize performance using a single aggregate loss. This abstraction is less informative in multilingual pretraining, where languages differ substantially in corpus size, linguistic structure, and representation in the training data. These differences raise practical questions that an aggregate curve cannot adequately address: whether low-resource languages benefit more from continued pretraining of a strong pretrained LLM than from training a model from scratch; whether incorporating data from another language family improves downstream performance; and whether model and data scaling affect all languages in comparable ways. Such questions are becoming increasingly important as multilingual corpora continue to expand. FineWeb-2 (Penedo et al., 2025), for example, provides training data for more than 1,000 languages, making it possible to train models on a substantially broader range of languages than was previously feasible.

In this work, we conduct a controlled study of multilingual pretraining across seven languages from the Afro-Asiatic and Austro-Asiatic families. We train models with approximately 0.3B and 0.5B parameters on up to 137B tokens from FineWeb-2 and examine four complementary settings. First, we compare continued pretraining from Qwen2.5-0.5B (Yang et al., 2024) with training the same architecture from random initialization. Second, we compare training exclusively on Afro-Asiatic data with training on an expanded corpus that additionally includes Austro-Asiatic languages. Third, we track language-specific losses throughout training to assess knowledge transfer across language families. Finally, we fit a separate scaling law for each language and use the resulting language-specific laws to reconstruct the aggregate training loss.

Our contributions can be summarized as follows

- We show that continued pretraining from a strong LLM yields faster convergence and consistently

* Corresponding author.

outperforms training from scratch across eight downstream benchmarks, even after the randomly initialized model has been trained on tens of billions of tokens.

- We demonstrate that reductions in training loss and more stable gradient norms do not necessarily translate into meaningful benchmark gains. A continuous Brier Score analysis further shows that the underlying probability estimates improve only marginally when accuracy remains close to random guessing.
- We provide evidence of cross-family transfer, showing that training only on Afro-Asiatic languages reduces loss on Austro-Asiatic languages. We further find that data-scaling exponents are comparatively stable across languages, whereas model-scaling exponents vary substantially, and that the fitted language-specific laws accurately reconstruct aggregate loss trajectories.

2 Related Work

Neural scaling laws describe predictable power-law relationships between language-model loss and model size, training data, or compute (Kaplan et al., 2020). Hoffmann et al. (2022) revise the compute-optimal allocation of parameters and tokens and introduce an additive formulation that separates model- and data-scaling effects. Subsequent studies extend this framework beyond abundant, homogeneous data. In data-constrained regimes, repeated exposure to the same tokens yields diminishing returns, motivating scaling laws that explicitly model data reuse (Muennighoff et al., 2023). Data mixing laws instead model performance as a function of the proportions assigned to different data domains and use small-scale training runs to select mixtures for larger budgets (Ye et al., 2025). Collectively, these studies show that scaling behavior is governed not only by the total number of parameters and tokens, but also by the availability, reuse, and composition of the training data.

Early work on multilingual scaling focused primarily on neural machine translation (Fernandes et al., 2023). More recent studies examine scaling laws for multilingual language models directly. He et al. (2025) fit loss as a function of model size, data volume, and language-family sampling ratios across 23 languages from five families. Their formulation assumes that each family’s loss is determined by its own sampling ratio and uses this simplification to derive compute-efficient multilin-

gual data mixtures. ATLAS (Longpre et al., 2025) substantially broadens the empirical scope, spanning models from 10M to 8B parameters, more than 400 training languages, and 48 evaluation languages. It introduces transfer-aware scaling laws, measures pairwise language benefit and interference, and models the capacity cost associated with the curse of multilinguality. Cao et al. (2026) further frame multilingual pretraining as a cooperative game and use language-specific contributions to parameterize cross-lingual transfer for performance prediction and training-data mixture optimization.

3 Experimental Setup

For the training dataset, we use the FineWeb-2 corpus (Penedo et al., 2025) to train the LLMs.¹ The dataset contains pretraining data covering more than 1,000 languages. During training, the context length is set to 8,192 tokens. To prevent self-attention across different documents within the same sequence, which can negatively affect downstream performance, we employ intra-document causal masking such that each token is allowed to attend only to preceding tokens within the same document (Zhao et al., 2024; Team, 2024).

We train the LLMs using Megatron-LM (Shoeybi et al., 2019).² We set the learning rate to 10^{-5} and use a cosine decay schedule that reduces it to 10% of its initial value by the end of training. We use 200 warm-up steps and a batch size of 16,777,216 tokens (2048×8192). Optimization is performed using AdamW (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay coefficient of 0.1. We apply gradient clipping with a maximum norm of 1.

4 Experimental Results

Our experiments analyze multilingual pretraining from several complementary perspectives, moving from comparisons of training strategies to language-specific analyses. We first isolate the effect of initialization by comparing continued pretraining with training from scratch. We then assess whether optimization gains from a broader multilingual corpus, as reflected in training loss and gradient ℓ_2 norm, translate into downstream improvements. Next, we examine language-specific

¹We use version 2.0.0 of the dataset, corresponding to the initial release of FineWeb-2. The dataset is available at <https://huggingface.co/datasets/HuggingFaceFW/fineweb-2>.

²<https://github.com/NVIDIA/Megatron-LM>

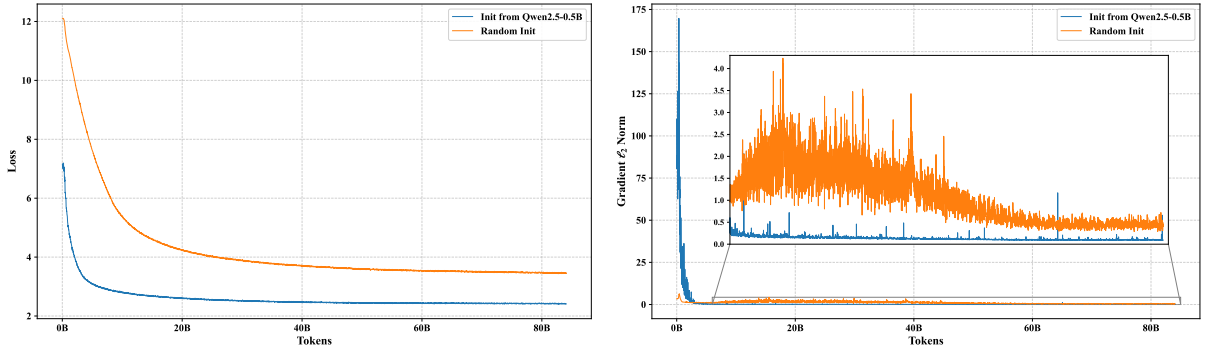


Figure 1: Training dynamics under different initialization settings. **Left:** training loss for a randomly initialized LLM and an LLM initialized from Qwen2.5-0.5B. **Right:** gradient ℓ_2 -norm for the same models.

Language	Language Code	#Tokens	#Documents	Language Family
Egyptian Arabic	arz	928,212,704	1,410,134	Afro-Asiatic
Najdi Arabic	ars	1,249,402,189	1,639,585	
Moroccan Arabic	ary	3,617,006,055	6,111,598	
Standard Arabic	arb	62,627,851,015	57,752,149	
Hebrew	heb	15,629,300,425	13,639,095	
Khmer	khm	3,582,904,670	1,467,637	Austro-Asiatic
Vietnamese	vie	49,694,903,737	40,741,340	

Table 1: Token and document counts for selected low-resource languages in the FineWeb-2 corpus, grouped by language family. Token counts are computed after applying the Qwen2.5-0.5B tokenizer to the corresponding text in each language subset.

loss trajectories to investigate transfer across language families. Finally, we fit language-specific scaling laws to quantify the effects of model and data scale and evaluate whether these laws recover the aggregate training dynamics.

4.1 Continued Pretraining versus Training from Scratch

For a strong pretrained LLM, continued pretraining on target-language data that are under-represented in its original pretraining corpus provides clear advantages over training a model from scratch on the same data.

To assess how model initialization affects both training dynamics and downstream performance, we compare two LLMs with identical architectures but different initialization schemes: (1) a model initialized from Qwen2.5-0.5B (Yang et al., 2024), and (2) a model with the same architecture as Qwen2.5-0.5B but randomly initialized and therefore trained from scratch.

We construct the training corpus from the subset of FineWeb-2 (Penedo et al., 2025) corresponding to the Afro-Asiatic language family, focusing on five low-resource languages: (1) Egyptian Arabic,

(2) Najdi Arabic, (3) Moroccan Arabic, (4) Standard Arabic, and (5) Hebrew. This corpus comprises approximately 84B tokens (84,051,772,388) across 80M documents (80,552,561). Table 1 reports the token and document counts for each language. We used the lm-evaluation-harness toolkit³ to assess the downstream performance of the trained LLMs. Because the toolkit provides limited support for other languages considered in our experiments, our downstream evaluation focused primarily on Arabic and related languages. Specifically, we evaluated the trained LLMs on eight benchmarks: (1) Belebele (Bandarkar et al., 2024), (2) ArabicMMLU (Koto et al., 2024), (3) Arabic EXAMS (Hardalov et al., 2020), (4) Arabic Leaderboard (Elfilali et al., 2024; Almazrouei et al., 2023; Huang et al., 2024; Fourier et al., 2023), (5) XNLI (Conneau et al., 2018), (6) XStoryCloze (Lin et al., 2022), (7) MMLU-ProX (Xuan et al., 2025), and (8) XQuAD (Artetxe et al., 2020).

As shown in Figure 1, compared with training from scratch, continual pretraining on the training data from the five low-resource languages from the Afro-Asiatic language family using an LLM initial-

³<https://github.com/eleutherai/lm-evaluation-harness>

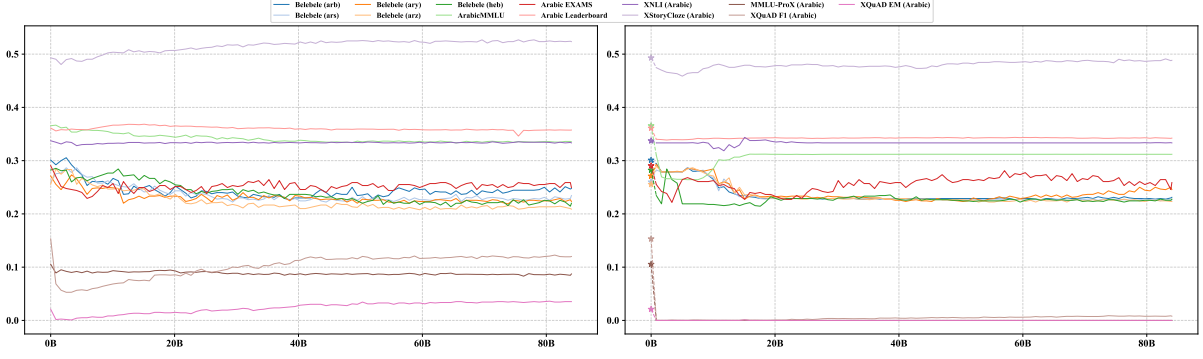


Figure 2: Benchmark performance of LLMs under two initialization settings. **Left**: models initialized from Qwen2.5-0.5B. **Right**: models with the same architecture but randomly initialized. The star marker indicates the performance of Qwen2.5-0.5B on these benchmarks. Language codes in parentheses specify the evaluation language: “arb” denotes Modern Standard Arabic, “ars” denotes Najdi Arabic, “ary” denotes Moroccan Arabic, “arz” denotes Egyptian Arabic, and “heb” denotes Hebrew. “XQuAD F1” and “XQuAD EM” report performance on the XQuAD benchmark measured by F1 score and exact match, respectively.

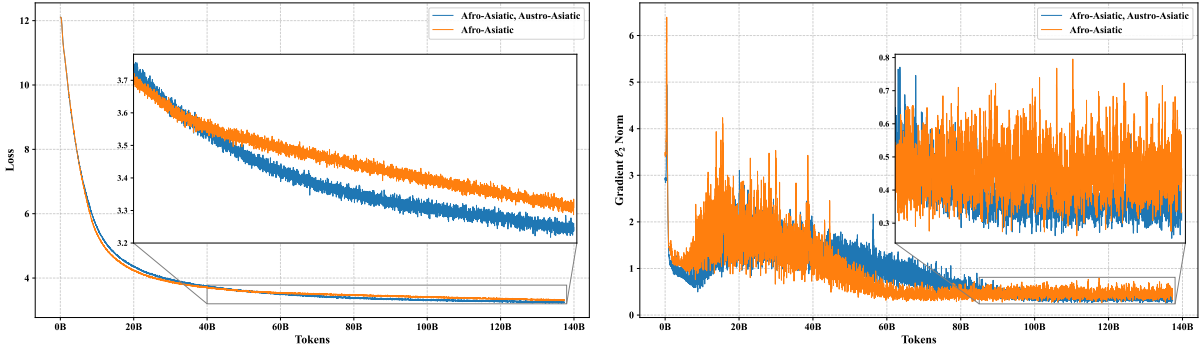


Figure 3: Training dynamics of LLMs trained from scratch on either Afro-Asiatic data alone or a combined corpus of Afro-Asiatic and Austro-Asiatic data. **Left**: training loss curves under the two training settings. **Right**: gradient ℓ_2 norm curves under the same settings.

ized from Qwen2.5-0.5B achieves faster loss convergence and converges to a lower loss value. Moreover, although the continually pretrained model exhibits a larger gradient ℓ_2 norm than the model trained from scratch during the early stage of training, its gradient ℓ_2 norm rapidly decreases to a low level after a short training period. In contrast, the model trained from scratch initially has a smaller gradient ℓ_2 norm, but it remains consistently higher than that of the continually pretrained model during the middle and later stages of training.

As shown in Figure 2, the continually pretrained model initialized from Qwen2.5-0.5B consistently outperforms the model trained from scratch across all evaluated benchmarks throughout training. This persistent performance gap indicates that initialization from a strong pretrained model not only accelerates optimization, as reflected by faster loss convergence and a more rapid reduction in gradient norm, but also translates into stronger down-

stream performance on low-resource language benchmarks. Moreover, the right panel of Figure 2 shows that, even after training on tens of billions of tokens, the model trained from scratch still does not surpass Qwen2.5-0.5B on these benchmarks.

4.2 Optimization Gains without Downstream Improvement

Pretraining LLMs on larger corpora can reduce training loss and stabilize the gradient ℓ_2 norm, but these optimization gains do not necessarily translate into improved performance on downstream benchmarks.

To extend our experiments to a broader range of language families, we augmented the Afro-Asiatic training corpus with additional data from the Austro-Asiatic family. Specifically, we trained the model on a combined corpus comprising Afro-Asiatic and Austro-Asiatic data, which contains

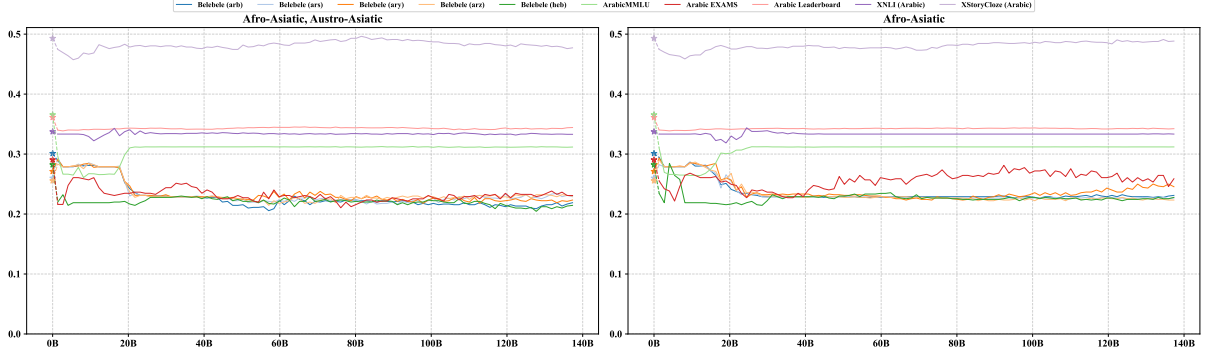


Figure 4: Benchmark performance of LLMs trained from scratch under two data settings: training on the combined Afro-Asiatic and Austro-Asiatic corpus (left) and training on the Afro-Asiatic corpus alone (right). The star marker indicates the performance of Qwen2.5-0.5B on these benchmarks.

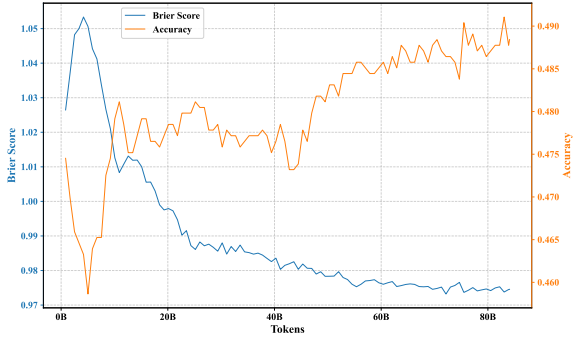


Figure 5: Changes in Brier Score and accuracy on the XStoryCloze benchmark over the course of training.

approximately 137B tokens. To ensure a fair comparison with this combined corpus setting, we further continued training the model trained only on Afro-Asiatic data, increasing its total training budget from approximately 84B to 137B tokens.

As shown in Figure 3, the training loss decreased substantially during training, while the gradient ℓ_2 norm also declined before stabilizing in the later stages. However, these optimization improvements led to only marginal changes in downstream benchmark performance, as shown in Figure 4. Because accuracy is a discrete metric that captures only whether the final prediction is correct, it may fail to reflect gradual changes in predicted probabilities during training. Motivated by prior work showing that evaluation metrics can substantially affect observed model performance (Schaeffer et al., 2023), we additionally use the Brier Score as a continuous metric. Unlike accuracy, the Brier Score measures the squared difference between predicted probabilities and the true label distribution, capturing changes in output probabilities even when the final prediction remains unchanged.

Specifically, we compute the Brier Score on XStoryCloze, a binary multiple-choice benchmark for narrative completion. Let y_{correct} and $y_{\text{incorrect}}$ denote the correct and incorrect options, respectively. For each test example i , the model assigns probabilities $p_i(y_{\text{correct}})$ and $p_i(y_{\text{incorrect}})$ to these two options. Given N test examples, the Brier Score (BS) is defined as follows:

$$\text{BS} = \frac{1}{N} \sum_{i=1}^N [(p_i(y_{\text{correct}}) - 1)^2 + (p_i(y_{\text{incorrect}}) - 0)^2]. \quad (1)$$

A lower Brier Score indicates more accurate and better calibrated probabilistic predictions.

Figure 5 shows the evolution of the Brier Score and accuracy during training. Although the Brier Score gradually decreases and varies more smoothly than accuracy, the magnitude of this change remains small, while benchmark accuracy stays close to 0.5, comparable to random guessing. These results further suggest that pretraining on the FineWeb-2 corpus does not lead to meaningful improvements on the evaluated benchmarks.

4.3 Cross-Family Transfer in Multilingual Pretraining

Even when training is restricted to a single language family, losses on languages from other families can still decrease, suggesting the presence of cross-family transfer.

Because the evaluated benchmarks appear to be relatively insensitive to changes in model behavior during training, we further analyze training dynamics by randomly sampling 1,000 documents for each language from the corpus and computing the loss of each saved checkpoint on these documents.

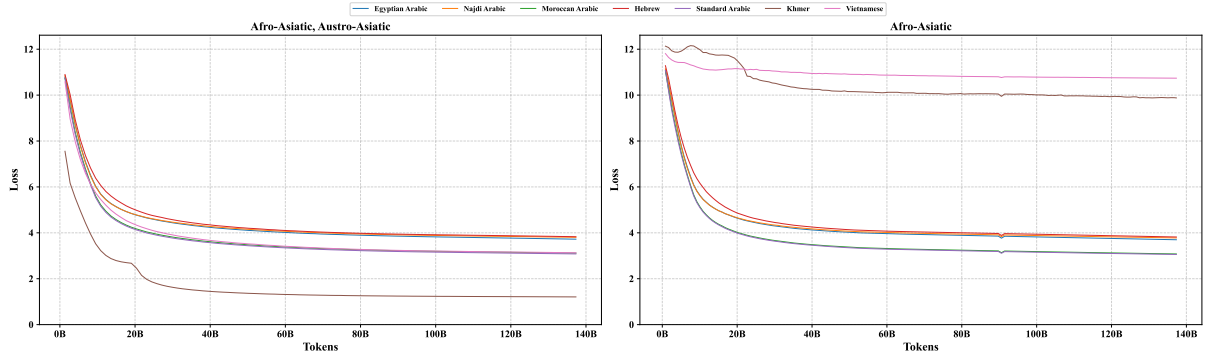


Figure 6: Language-specific losses on sampled documents for LLMs trained on the combined Afro-Asiatic and Austro-Asiatic corpus (left) and on the Afro-Asiatic corpus alone (right).

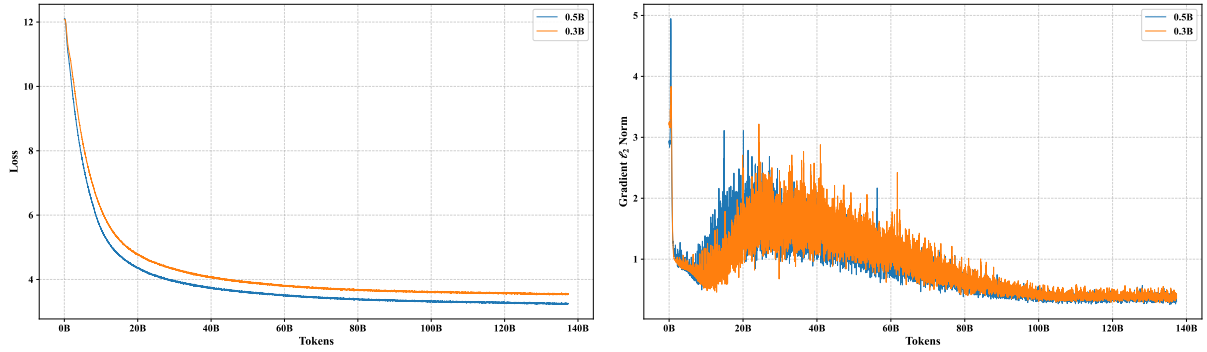


Figure 7: Training dynamics of the 0.3B- and 0.5B-parameter models trained on the combined Afro-Asiatic and Austro-Asiatic corpus. The left panel shows the training-loss trajectories, and the right panel shows the corresponding gradient ℓ_2 norm trajectories over the course of training.

As shown in Figure 6, when the model is trained on data from both the Afro-Asiatic and Austro-Asiatic language families, the loss on the sampled documents decreases for languages in both families. Notably, even when the model is trained only on Afro-Asiatic languages, the loss on Austro-Asiatic languages also decreases. We hypothesize that, despite the absence of Austro-Asiatic data in this setting, the model learns language-general patterns from Afro-Asiatic languages that partially transfer to Austro-Asiatic languages. Such cross-family transfer may explain the moderate loss reductions observed for those languages.

We also observe abrupt loss reductions at several checkpoints during training. For example, when the model is trained on both Afro-Asiatic and Austro-Asiatic languages, the loss on Khmer declines more sharply after approximately 19B training tokens. Similarly, when training is restricted to Afro-Asiatic languages, sudden loss reductions occur across multiple languages at around 90B tokens. This pattern is consistent with transient loss fluctuations reported in LLM training (Molybog et al.,

2023; Nishida et al., 2024; Takase et al., 2023). Although the precise cause remains unclear, one plausible explanation is that, given the large batch size of 16,777,216 tokens and the use of 1,000 randomly sampled documents for loss evaluation, some evaluation documents may have appeared in training batches shortly before the corresponding checkpoints. These abrupt reductions may therefore reflect local training effects, such as transient fitting or sample-level memorization, rather than sustained improvements in generalization.

4.4 Language-Specific Scaling Laws and Aggregate Loss Prediction

Language-specific scaling laws show that increasing training data consistently reduces loss across languages, whereas gains from larger models vary substantially; nevertheless, the fitted laws accurately recover aggregate training-loss trends across model scales.

To examine training dynamics across model scales, we additionally trained a smaller model

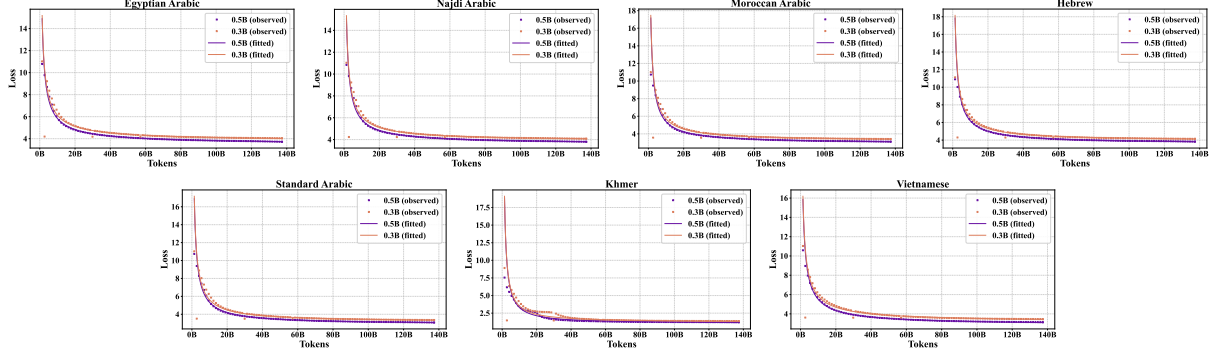


Figure 8: Language-specific training-loss trajectories measured on 1,000 randomly sampled documents for models trained on the combined Afro-Asiatic and Austro-Asiatic corpus. The x markers indicate the observed training losses, while the solid lines represent the losses predicted by the fitted scaling laws.

Parameters	Layers	Q Heads	KV Heads	Hidden Size	FFN Hidden Size
493,803,392 (0.5B)	24	14	2	896	4864
267,547,392 (0.3B)	16	12	12	768	3072

Table 2: Architectural configurations of the 0.5B- and 0.3B-parameter models.

with approximately 0.3B parameters on the combined Afro-Asiatic and Austro-Asiatic corpus. The architectural hyperparameters of this model are summarized in Table 2. Figure 7 compares the training loss and gradient ℓ_2 norm trajectories of the 0.3B- and 0.5B-parameter models, while Figure 8 reports the loss measured on 1,000 randomly sampled documents throughout training.

As shown in the left panel of Figure 7 and Figure 8, the 0.3B-parameter model consistently exhibits higher loss than the 0.5B-parameter model throughout training, as expected from its smaller capacity. It also exhibits the abrupt loss drops discussed in Section 4.3. In contrast, the right panel of Figure 7 shows that the gradient ℓ_2 -norm trajectories of the two models follow similar patterns throughout training.

To characterize language-specific scaling behavior during training, we analyze how the training loss for each language varies as a function of model size and training data size. Let N denote the number of model parameters, D denote the number of training tokens, and $\mathcal{L}(N, D)$ denote the corresponding training loss. Following prior work on neural scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022), we model the relationship among language-specific training loss, model size, and data size as follows:

$$\mathcal{L}(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta} \quad (2)$$

where (E, A, B, α, β) are parameters to be esti-

mated, and E represents the irreducible entropy of natural text. We estimate these parameters by minimizing the Huber loss between the predicted log loss, $\log \hat{\mathcal{L}}(N, D)$, and the observed log loss, $\log \mathcal{L}(N, D)$, using the L-BFGS algorithm:

$$\min_{E, A, B, \alpha, \beta} \text{Huber}_\delta \left(\log \hat{\mathcal{L}}(N, D) - \log \mathcal{L}(N, D) \right) \quad (3)$$

where $\text{Huber}_\delta(x)$ is defined as

$$\text{Huber}_\delta(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq \delta, \\ \delta \cdot (|x| - \frac{1}{2}\delta) & \text{otherwise,} \end{cases} \quad (4)$$

Following Hoffmann et al. (2022), we set δ to 0.001.

Table 3 reports the fitted scaling laws for each language. The estimated parameters reveal several systematic patterns. First, the data-scaling exponent β is relatively stable across languages, ranging from 0.777 to 1.026, suggesting that additional training tokens yield broadly consistent loss reductions across the multilingual corpus. Second, the model-scaling exponent α shows substantially greater cross-lingual variation. Egyptian Arabic and Najdi Arabic have the largest α values, indicating that their losses are especially sensitive to increases in model capacity. In contrast, Moroccan Arabic, Standard Arabic, and Hebrew have much smaller α values, suggesting more limited gains from model scaling alone. Third, the fitted entropy term E also varies considerably: Khmer, Moroccan Arabic, and Standard Arabic have lower values, whereas Egyptian Arabic and Najdi Arabic have higher values, indicating language-specific differences in asymptotic loss. Overall, these results suggest that data scaling is comparatively consistent across languages, while the benefits of increasing model size are more language dependent. We further observe substantial variation among languages

Language Family	Language	Scaling Laws
Afro-Asiatic	Egyptian Arabic	$\mathcal{L}(N, D) = 3.104 + \frac{669927555.63}{N^{1.073}} + \frac{149873290.534}{D^{0.779}}$
	Najdi Arabic	$\mathcal{L}(N, D) = 3.278 + \frac{8329410405.246}{N^{1.211}} + \frac{279628650.024}{D^{0.808}}$
	Moroccan Arabic	$\mathcal{L}(N, D) = 1.175 + \frac{379.877}{N^{0.271}} + \frac{1093830819.922}{D^{0.863}}$
	Standard Arabic	$\mathcal{L}(N, D) = 1.359 + \frac{471.323}{N^{0.289}} + \frac{1020485784.673}{D^{0.861}}$
	Hebrew	$\mathcal{L}(N, D) = 1.919 + \frac{525.337}{N^{0.289}} + \frac{725375445.333}{D^{0.844}}$
Austro-Asiatic	Khmer	$\mathcal{L}(N, D) = 0.702 + \frac{4575825.218}{N^{0.826}} + \frac{42365446931.761}{D^{1.026}}$
	Vietnamese	$\mathcal{L}(N, D) = 2.176 + \frac{2705218.39}{N^{0.769}} + \frac{164450830.132}{D^{0.777}}$

Table 3: Fitted language-specific scaling laws for modeling training loss as a function of model size and training data size across the Afro-Asiatic and Austro-Asiatic languages.

within the same family, with no clear family-level regularities in scaling behavior with respect to either model size or data volume. Conversely, languages from different families can exhibit similar scaling patterns. For example, Egyptian Arabic, an Afro-Asiatic language, has a data-scaling exponent β of 0.779, comparable to that of Vietnamese, an Austro-Asiatic language. These findings suggest that scaling behavior with respect to model size and data volume may be only weakly associated with language-family membership.

To assess the quality of the fits, we visualize the fitted scaling-law curves and compare them with the observed training losses in Figure 8. The observed losses are generally well aligned with the fitted curves, suggesting that the scaling-law formulation and the estimated parameters capture the empirical training dynamics effectively.

Furthermore, by combining the language-specific sampling ratios with the fitted scaling laws, we can estimate how the aggregate training loss evolves as a function of the number of training tokens across model scales. Let p_i denote the sampling probability of language i , and let $\hat{\mathcal{L}}_i(N, D)$ denote the predicted training loss for that language for a model with N parameters after training on D tokens. The predicted aggregate training loss, $\hat{\mathcal{L}}(N, D)$, is then defined as the sampling-weighted average of the language-specific losses:

$$\hat{\mathcal{L}}(N, D) = \sum_i p_i \hat{\mathcal{L}}_i(N, D) \quad (5)$$

Using the language-specific scaling laws reported in Table 3, we compute the weighted average loss

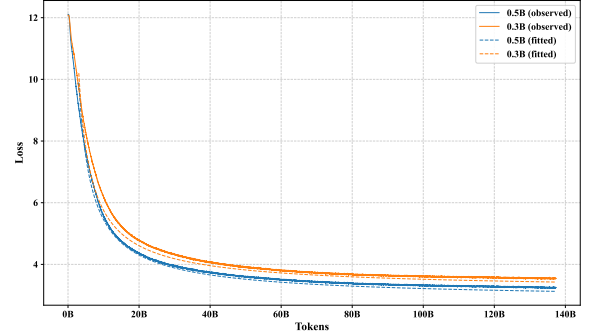


Figure 9: Predicted and observed aggregate training losses for the 0.5B- and 0.3B-parameter models. To reduce the influence of early-stage outliers in the fitted losses, the fitted loss curves are shown beginning at approximately 3 billion training tokens.

based on the empirical sampling probability of each language in the multilingual training corpus. This aggregate estimate approximates the model’s overall observed training loss. Because training examples are sampled randomly from the corpus, we estimate each language’s sampling probability as its proportion of the total training tokens. As illustrated in Figure 9, the predicted aggregate loss closely tracks the observed training loss.

To further evaluate the quality of the fitted scaling laws, we compute the mean absolute difference between the predicted and observed losses. The resulting differences are 0.097 and 0.131 for the 0.5B- and 0.3B-parameter models, respectively. This close agreement indicates that the language-specific scaling laws accurately capture the empirical training dynamics and provide reliable estimates of the overall training loss.

5 Conclusion

We have investigated multilingual language-model pretraining from the perspectives of training dynamics, downstream evaluation, cross-lingual transfer, and language-specific scaling laws. Our results demonstrate that initialization from a pretrained model provides clear advantages over training from scratch. At the same time, improvements in optimization metrics, including training loss and gradient ℓ_2 norm, do not necessarily translate into stronger downstream performance. We further observe evidence of transfer across language families, together with substantial cross-lingual variation in scaling behavior. Taken together, these findings underscore the importance of language-specific analysis for understanding and designing multilingual models, and they motivate further investigation across broader model scales, language distributions, and training-data mixtures.

Acknowledgments

The present research was supported by the National Key Research and Development Program of China (Grant No. 2023YFE0116400). We would like to thank the anonymous reviewers for their insightful comments.

Limitations

Due to computational constraints, we conducted experiments with 0.3B- and 0.5B-parameter models, while extending the training data to nearly 137B tokens. Experiments with larger models would provide further validation of our findings. Nevertheless, our empirical language-specific analysis and experimental framework can be extended to larger-scale settings and may inspire future research on multilingual scaling laws.

References

- Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammedi, Julien Launay, and Badreddine Noune. 2023. [Alghafa evaluation benchmark for arabic language models](#). In *Proceedings of ArabicNLP 2023, Singapore (Hybrid), December 7, 2023*, pages 244–275. Association for Computational Linguistics.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. [On the cross-lingual transferability of monolingual representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4623–4637. Association for Computational Linguistics.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabza. 2024. [The belebele benchmark: a parallel reading comprehension dataset in 122 language variants](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 749–775. Association for Computational Linguistics.
- Xuyang Cao, Qianying Liu, Chuan Xiao, Yusuke Oda, Pontus Stenetorp, Daisuke Kawahara, Makoto Onizuka, Sadao Kurohashi, and Shuyuan Zheng. 2026. [Shapleylaw: A game-theoretic approach to multilingual scaling laws](#). *CoRR*, abs/2603.17945.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2475–2485. Association for Computational Linguistics.
- Ali Elfilali, Hamza Alobeidli, Cl  mentine Fourier, Basma El Amel Boussaha, Ruxandra Cojocaru, Nathan Habib, and Hakim Hacid. 2024. Open arabic llm leaderboard. <https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard>.
- Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. 2023. [Scaling laws for multilingual neural machine translation](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10053–10071. PMLR.
- Cl  mentine Fourier, Nathan Habib, Thomas Wolf, and Lewis Tunstall. 2023. [Lighteval: A lightweight framework for llm evaluation](#).
- Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. [EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5427–5444. Association for Computational Linguistics.
- Yifei He, Alon Benhaim, Barun Patra, Praneetha Vadamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, and Xia Song. 2025. [Scaling laws for multilingual language models](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*,

- volume ACL 2025 of *Findings of ACL*, pages 4257–4273. 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](#). *CoRR*, abs/2203.15556.
- Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Mosen Alharthi, Bang An, Juncal He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. [Acept, localizing large language models in arabic](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 8139–8163. 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](#). *CoRR*, abs/2001.08361.
- Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. [Arabicmmlu: Assessing massive multitask language understanding in arabic](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 5622–5640. Association for Computational Linguistics.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, and 2 others. 2022. [Few-shot learning with multilingual generative language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9019–9052. Association for Computational Linguistics.
- Shayne Longpre, Sneha Kudugunta, Niklas Muennighoff, I-Hung Hsu, Isaac Caswell, Alex Pentland, Sercan Ö. Arik, Chen-Yu Lee, and Sayna Ebrahimi. 2025. [ATLAS: adaptive transfer scaling laws for multilingual pretraining, finetuning, and decoding the curse of multilinguality](#). *CoRR*, abs/2510.22037.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Igor Molybog, Peter Albert, Moya Chen, Zachary DeVito, David Esiobu, Naman Goyal, Punit Singh Koura, Sharan Narang, Andrew Poulton, Ruan Silva, Binh Tang, Diana Liskovich, Puxin Xu, Yuchen Zhang, Melanie Kambadur, Stephen Roller, and Susan Zhang. 2023. [A theory on adam instability in large-scale machine learning](#). *CoRR*, abs/2304.09871.
- Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A. Raffel. 2023. [Scaling data-constrained language models](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Kosuke Nishida, Kyosuke Nishida, and Kuniko Saito. 2024. [Initialization of large language models via reparameterization to mitigate loss spikes](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 22699–22714. Association for Computational Linguistics.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolcec, Bettina Messmer, Negar Foroutan, Amir Hosein Kargaran, Colin Raffel, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2025. [Fineweb2: One pipeline to scale them all - adapting pre-training data processing to every language](#). *CoRR*, abs/2506.20920.
- Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. [Are emergent abilities of large language models a mirage?](#) In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. [Megatron-lm: Training multi-billion parameter language models using model parallelism](#). *CoRR*, abs/1909.08053.
- Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. 2023. [Spike no more: Stabilizing the pre-training of large language models](#). *CoRR*, abs/2312.16903.
- Llama Team. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu,

- Yuekang Li, Xiaofei Xie, and 13 others. 2025. [Mmlu-prox: A multilingual benchmark for advanced large language model evaluation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 1513–1532. Association for Computational Linguistics.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. 2025. [Data mixing laws: Optimizing data mixtures by predicting language modeling performance](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Yu Zhao, Yuanbin Qu, Konrad Staniszewski, Szymon Tworowski, Wei Liu, Piotr Milos, Yuxiang Wu, and Pasquale Minervini. 2024. [Analysing the impact of sequence composition on language model pre-training](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7897–7912. Association for Computational Linguistics.