CROSS-LINGUAL DATA SCALING FOR LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many target languages have limited pretraining text even as English corpora continue to grow. We study a controlled cross-lingual scaling question: when the available target-language corpus is fixed, how does target-language performance change as English data increase? In English–Chinese experiments, proportional mixing dilutes Chinese exposure and can worsen Chinese perplexity. Under staged training, a single cosine schedule couples the Chinese-stage learning rate to the length of English training; holding the Chinese-stage schedule fixed instead produces positive trends over the evaluated scales. Token-category ablations are consistent with shared lexical content contributing to this transfer, although they do not establish a general mechanism. Repeating the fixed target-language corpus preserves its effective training share, with diminishing returns at high repetition. We instantiate these interventions in ScaleX and ScaleX-DR and evaluate them across English data scales. In multilingual experiments that scale English while holding Chinese, Turkish, Hungarian, and Bengali corpora fixed, ScaleX-DR attains lower perplexity and higher downstream accuracy at the largest than at the smallest English scale for all four target languages. A separate 2.5B-parameter English–Chinese comparison at 500B and 1T tokens improves both languages. These results characterize how schedule and target-language exposure shape cross-lingual scaling in the evaluated settings.

1 INTRODUCTION

Large language models exhibit predictable gains from increasing model size, compute, and training data (Kaplan et al., 2020; Hoffmann et al., 2022). Data, however, are distributed unevenly across languages. English continues to accumulate web-scale text, whereas many languages have much smaller corpora and slower data growth (Villalobos et al., 2022; Conneau et al., 2020; Xue et al., 2021). When unique target-language data are exhausted, additional compute either reduces that language's share in a growing multilingual mixture or repeats its existing data. Both choices can produce diminishing returns (Muennighoff et al., 2023; Xue et al., 2023).

We study a specific scaling regime: the unique target-language corpus is fixed while the available high-resource-language corpus grows. We call this *cross-lingual data scaling*. The question is not whether multilingual pretraining transfers knowledge at one budget, but whether target-language performance improves consistently across a sequence of larger high-resource budgets. Under proportional mixing, the target share shrinks with scale, and we observe that Chinese performance degrades even though the total training budget increases.

We investigate two complementary ways to change this trend. First, we study training dynamics. An English-to-Chinese curriculum can use English pretraining as initialization, but a single cosine schedule couples the Chinese-stage learning rate to the length of the English stage. Holding the Chinese-stage schedule fixed makes the target trend positive over the measured range. Second, we study the data mixture. Translation and code-switching add Chinese signal but only mitigate degradation in our setup; controlled repetition improves the target trend until repeated exposure saturates.

These findings lead to ScaleX, which combines a constant-rate high-resource stage with a fixed cosine target-language stage, and ScaleX-DR, which additionally repeats target data. Our contributions are:

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

- We formulate cross-lingual data scaling as a fixed-target, growing-high-resource problem and identify target dilution under proportional mixing.
- We show that the learning-rate trajectory experienced by the target stage can determine the sign of the cross-scale trend.
- We compare synthetic data, repetition, balanced repetition, and loss weighting, identifying a finite useful regime for repeated target exposure.
- We evaluate ScaleX and ScaleX-DR at two Chinese-data budgets, test multilingual generalization on Chinese, Turkish, Hungarian, and Bengali, and report English-retention costs.

Chinese serves as a controlled target because it permits stable data quality and evaluation while its unique corpus is artificially capped. This design isolates the fixed-target scaling question in the bilingual study, while the multilingual evaluation separately tests the trend on naturally smaller target corpora.

2 Related Work

Multilingual data allocation and scaling. Multilingual pretraining balances language coverage, data imbalance, and finite model capacity (Conneau et al., 2020; Xue et al., 2021). UniMax limits repeated epochs while allocating examples across languages (Chung et al., 2023); multilingual scaling laws model interactions among sampling proportions and language families (He et al., 2025); and ATLAS jointly studies model size, data, language count, and cross-lingual transfer (Longpre et al., 2026). Recent large multilingual studies further show that interference depends on token allocation, language coverage, and capacity rather than on language count alone (Chang et al., 2024; Foroutan et al., 2026). M^3 compares multilingual, multi-epoch, and two-stage recipes for compute-optimal training with limited data (Akimoto et al., 2024). These works optimize or predict multilingual performance across broad design spaces. We instead hold unique target data fixed and measure how its performance changes as high-resource data alone grow.

Cross-lingual adaptation. Language-order curricula and continual pretraining use high-resource data before target-language adaptation, often with replay or learning-rate restart (Choi et al., 2023; Zheng et al., 2024; Li et al., 2025b). Under constrained target-language data, auxiliary-language data, translation, and upsampling provide complementary ways to improve bilingual pretraining (Seto et al., 2025). Our focus is a scale-dependent confound: when the first stage grows, a single global schedule changes the optimization trajectory of an otherwise fixed target stage.

Synthetic data and repeated exposure. Translation and code-switching increase target-language signal (Wang et al., 2025a;b; Yoo et al., 2025). Upsampling and loss weighting also emphasize underrepresented data, but through different compute and optimization pathways (Li et al., 2025a). Repeated epochs remain useful only over a finite range (Muennighoff et al., 2023; Xue et al., 2023). We place these alternatives on the same fixed-target scaling axis to compare how data transformation, processed exposure, and objective weighting affect the cross-scale trend.

3 Preliminary Study

3.1 Experimental setup

We train randomly initialized 1.3B-parameter LLaMA-style models using Megatron-LM (Touvron et al., 2023; Shoeybi et al., 2019). Unless stated otherwise, sequence length is 4,096, global batch size is 256, and the learning rate ranges from 3×10^{-4} to 3×10^{-5} without warmup. Chinese PPL is averaged over five domains. Chinese downstream evaluation uses CEVAL, AGIEval, CMMLU, and Math23K; English evaluation uses five-domain PPL and GSM8K, BBH, MMLU, and HumanEval (Huang et al., 2023; Li et al., 2024; Zhong et al., 2024; Cobbe et al., 2021; Suzgun et al., 2023; Hendrycks et al., 2021; Chen et al., 2021). Each plotted point is an independently trained model. The main curves contain one run per condition, so we report point estimates without error bars or significance tests. Appendix A reports the schedule controls and additional experimental details.

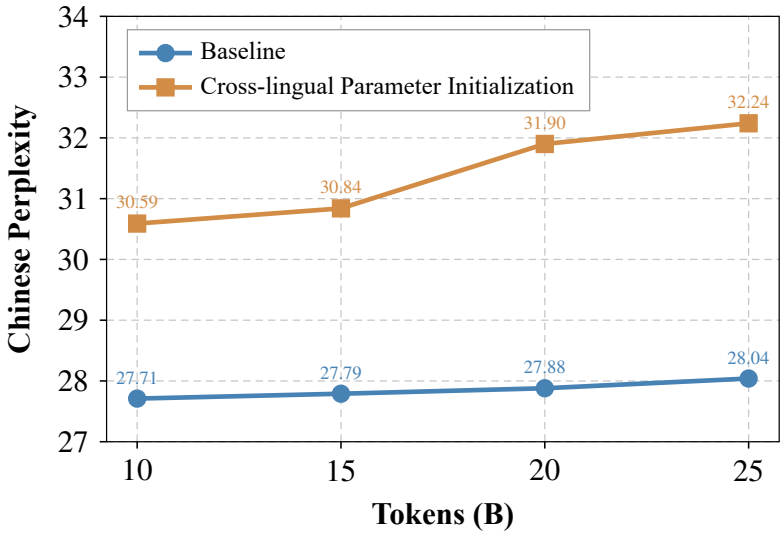


Figure 1: Cross-lingual parameter initialization versus proportional mixing. The fixed 5B Chinese corpus is evaluated at 10, 15, 20, and 25B total tokens.

We distinguish unique target data from processed target exposure after repetition. Translation and code-switching replace English tokens, balanced repetition adds processed data in both languages, and loss weighting changes the objective without duplicating input examples. These comparisons test related but non-identical interventions.

3.2 Cross-lingual parameter initialization

We first fix 5B Chinese tokens and vary total data over 10, 15, 20, and 25B tokens. Under proportional mixing, the shrinking Chinese share produces a mild degradation from 27.71 to 28.04 PPL. We then train on English followed by Chinese. As Figure 1 shows, this naive two-stage alternative degrades more sharply and therefore does not by itself solve cross-lingual scaling.

3.3 Learning-rate schedule

The naive curriculum uses one cosine schedule over the full run. Increasing English data lengthens the first stage, so the fixed Chinese stage starts progressively later in the decay and receives a different learning-rate trajectory at every scale. We compare three schedules (Figure 2): Cos applies one cosine over the full run; Cos-Cos restarts a fixed cosine for the Chinese stage while matching the first-stage portion of Cos; and Const-Cos uses a constant English stage followed by the same fixed Chinese cosine. Figure 3 shows that controlling the target-stage schedule changes the trend. Over 10–25B, Cos worsens from 30.59 to 32.24 PPL, whereas Cos-Cos improves from 26.57 to 24.62 and Const-Cos from 26.39 to 24.44.

This comparison isolates the second-stage restart locally through Cos versus Cos-Cos. Const-Cos additionally changes the first-stage schedule and should be interpreted as a complete protocol rather than a single-component ablation.

3.4 Shared lexical categories

Fixing the target-stage schedule permits improvement, but does not identify what information is transferred between stages. We partition surface tokens into Chinese characters, Latin-alphabet tokens, and other symbols such as punctuation and digits. To avoid heavily perturbing the Chinese target corpus, this diagnostic reverses language order: Chinese is used first and English second. Table 1 reports the controlled configurations. English improves with first-stage scale when a category used by English is shared across stages. This supports an association between shared lexical cate-

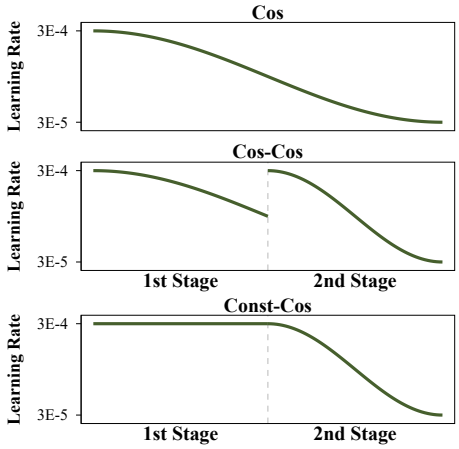


Figure 2: Learning-rate schedules.

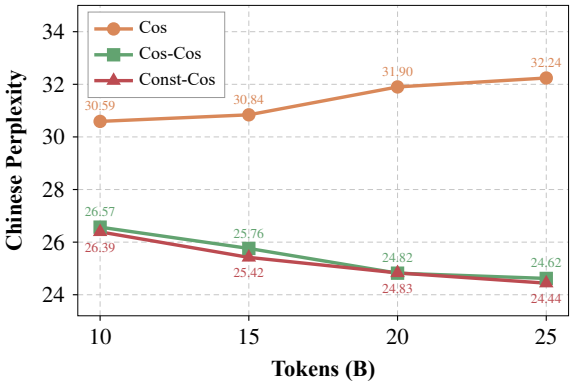


Figure 3: Chinese PPL under Cos, Cos-Cos, and Const-Cos.

Table 1: Controlled shared-token experiment. Checkmarks denote token categories shared across stages.

#	Shared Token		Scalable	English Perplexity		
	English	Other		4B	40B	80B
1			✗	**12.54**	12.63	13.02
2	✓		✓	12.26	11.50	**11.34**
3		✓	✓	12.42	11.74	**11.49**
4	✓	✓	✓	12.16	11.34	**11.16**

gories and transfer, but does not prove a unique mechanism because token filtering changes the data distribution. Appendix F details the construction and reports all configurations.

3.5 Data transformation and repeated exposure

Translation and code-switching inject Chinese signal while preserving total tokens. Figure 4 shows that both alleviate degradation relative to proportional mixing, but neither produces a positive trend over 20–160B. At 160B, Chinese PPL is 27.67 with translation and 30.26 with code-switching, versus 32.79 for the baseline. This result applies to the evaluated generation and replacement pipelines; it does not isolate synthetic quality, domain mismatch, or injection amount as a unique cause. Appendices D and E describe the two pipelines and their selected operating points.

Repeating the fixed Chinese corpus improves perplexity at low repetition counts but eventually reverses, while balanced repetition continues to improve over the tested interval (Figure 5). This motivates moderate repetition together with high-resource scaling. It does not imply that duplicated tokens are equivalent to unique data or improve indefinitely.

4 ScaleX

The preliminary study identifies two complementary controls. ScaleX uses a constant learning rate for English pretraining and a fixed cosine schedule for the Chinese stage. ScaleX-DR additionally repeats Chinese data during the second stage to increase effective target exposure (Figure 6). The target-stage schedule is held constant as English data grow. ScaleX operationalizes these controls

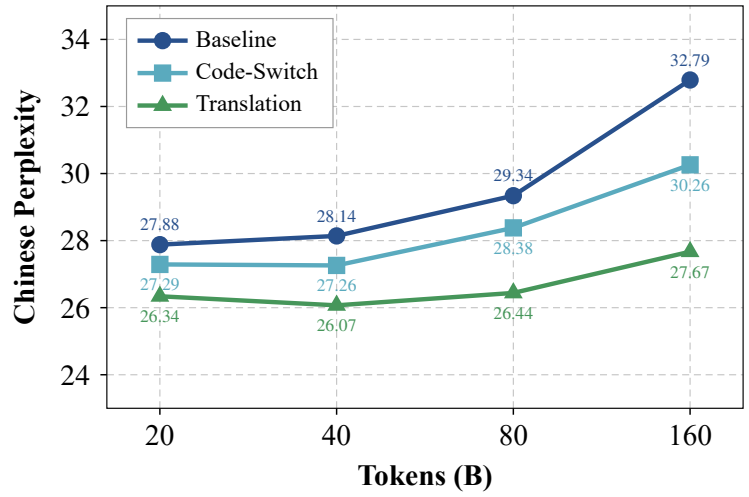


Figure 4: Translation and code-switching mitigate, but do not reverse, Chinese degradation as English data grow.

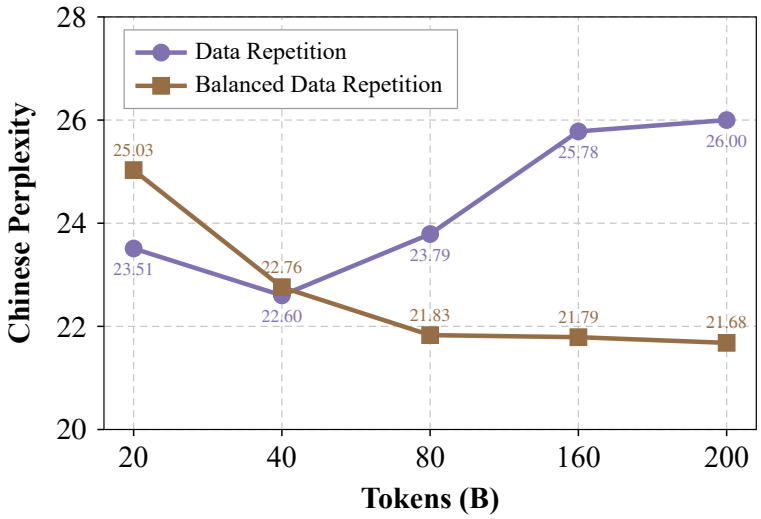


Figure 5: Data repetition versus balanced repetition. Unique Chinese data are fixed; processed exposure changes.

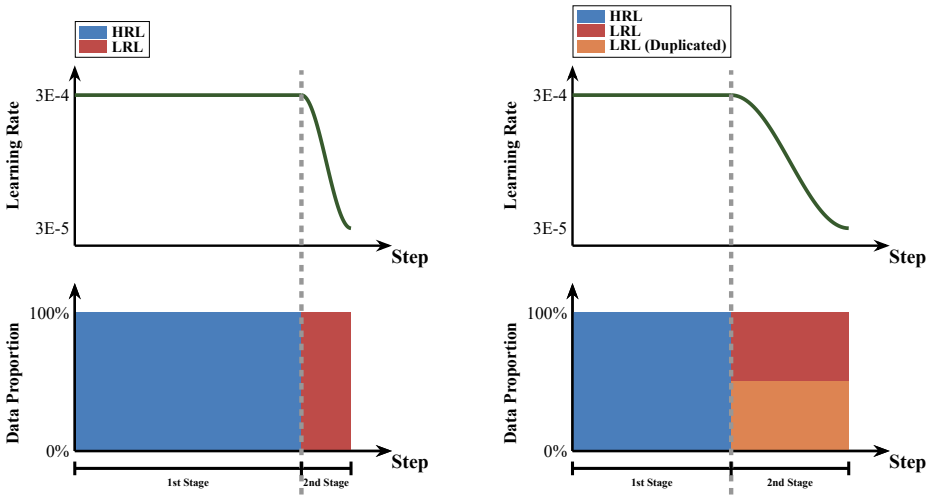
as a two-stage pretraining strategy in which the optimization path and processed target exposure remain comparable across the high-resource scaling axis.

5 Experiments

5.1 Main bilingual results

We evaluate 5B- and 20B-Chinese settings. In the 5B setting, total budgets are 20, 40, 80, and 160B tokens; in the 20B setting they are 80, 120, 200, and 320B. Figure 7 reports Chinese PPL and downstream accuracy.

At 5B Chinese, proportional mixing worsens Chinese PPL from 27.88 to 32.79. ScaleX improves from 24.83 to 22.03, and ScaleX-DR from 23.25 to 20.71; Chinese downstream accuracy under ScaleX-DR rises from 22.55 to 29.14. At 20B Chinese, the baseline worsens from 22.56 to 25.83,



(a) ScaleX (b) ScaleX-DR

Figure 6: ScaleX and ScaleX-DR. ScaleX fixes the target-stage schedule across English-data scales; ScaleX-DR additionally increases target exposure.

while ScaleX improves from 19.40 to 18.43 and ScaleX-DR from 18.50 to 17.73. Thus, the positive target trend replicates at both controlled Chinese-data budgets.

Loss weighting improves over proportional mixing at individual budgets but does not reverse the PPL trend: 25.61 at 40B rises to 27.44 at 160B in the 5B setting, and 20.99 rises to 22.12 over 80–320B in the 20B setting. Its accuracy gains are also smaller than those of balanced repetition, which improves both PPL and accuracy across the measured ranges. Changing the objective weight is therefore not equivalent to controlling the data sequence and target-stage exposure.

5.2 Multilingual generalization

The bilingual study uses Chinese as a controlled target; we next test whether the cross-scale trend extends beyond that setting. Figure 8 evaluates four fixed target corpora from distinct language families: Chinese (Sino-Tibetan; 3.988B tokens), Turkish (Turkic; 1.124B), Hungarian (Uralic; 0.755B), and Bengali (Indo-European; 0.167B), while English grows from 50B to 450B tokens. At the largest versus the smallest English scale, ScaleX-DR attains lower target-language PPL and higher downstream accuracy for all four languages. The result provides a multilingual generalization check across the tested families and imbalance levels; it is not a claim of universal transfer to arbitrary languages.

5.3 Target gains and English retention

The target improvements are not Pareto improvements. Table 2 reports matched Chinese and English PPL at representative endpoints. In both data settings, ScaleX and ScaleX-DR improve Chinese PPL but retain less English capability than proportional mixing. Additional English data still improves English within each staged method, but the final adapted models do not dominate the baseline on English. The method is therefore appropriate when target-language adaptation is the primary objective and the retention cost is acceptable.

5.4 Large-scale case study

To test a more practical retention setting, we use a 2.5B-parameter model with an 80/20 Chinese–English mixture in the second stage. We treat this ratio as a retention-oriented operating point rather than an optimized value. Table 3 reports 500B- and 1T-token runs. These two budgets provide a large-scale case study; they do not establish a continuous scaling law without a matched

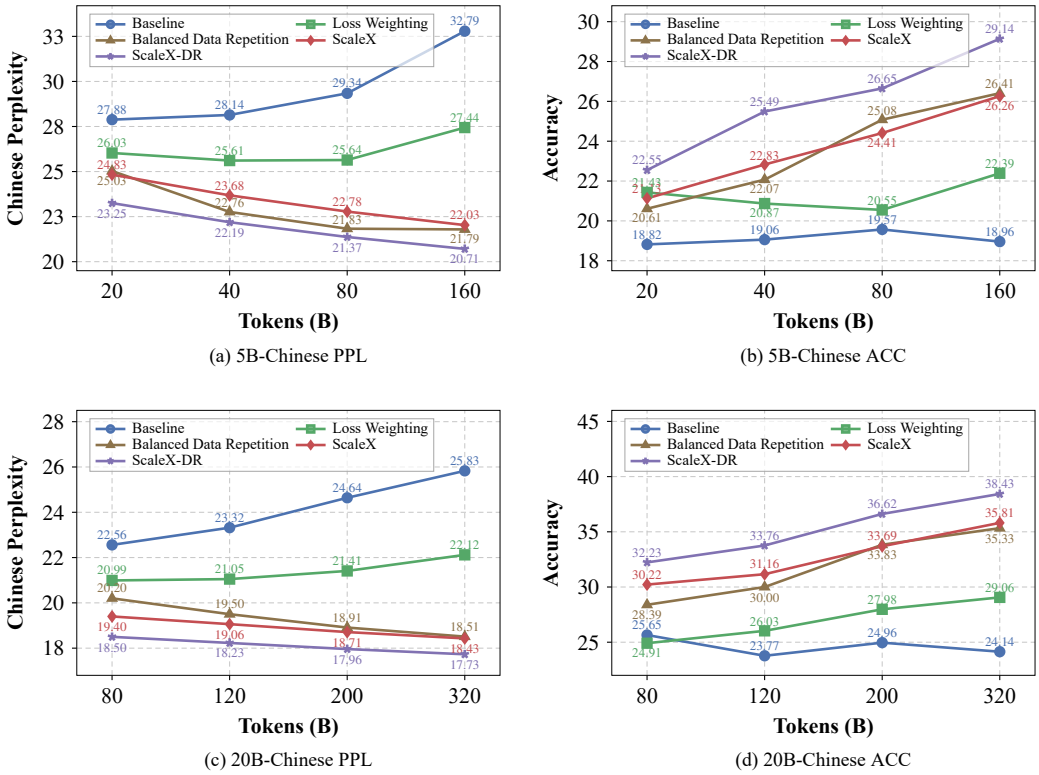


Figure 7: Main bilingual results at 5B and 20B fixed Chinese-data budgets. ScaleX and ScaleX-DR improve Chinese PPL and accuracy as English data grow.

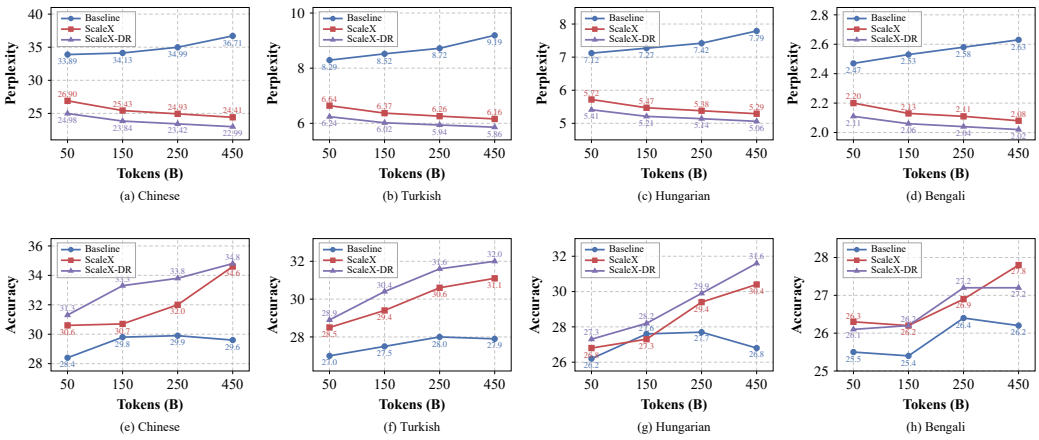


Figure 8: Multilingual generalization as English data scale while each target-language corpus remains fixed. The top row reports PPL and the bottom row reports downstream accuracy for Chinese, Turkish, Hungarian, and Bengali.

proportional-mixing baseline. Appendix C describes this mixed-stage setting and its connection to the controlled two-stage experiments.

6 Discussion

The experiments support a simple interpretation. Under proportional mixing, scaling English changes both total compute and the amount of Chinese signal. Under naive staging, scaling English

Table 2: Target-language gains and English-retention cost at representative endpoints. Lower PPL is better.

Fixed Chinese	Method	Chinese PPL	Chinese Acc.	English PPL	English Acc.
5B / 160B	Baseline	32.79	18.96	10.11	23.85
	ScaleX	22.03	26.26	15.33	21.00
	ScaleX-DR	**20.71**	**29.14**	17.49	20.85
20B / 200B	Baseline	24.64	24.96	10.08	29.37
	ScaleX	18.71	33.69	18.02	22.18
	ScaleX-DR	**17.96**	**36.62**	19.58	22.70

Table 3: ScaleX-DR large-scale mixed-stage case study.

Language	Metric	Tokens	
		500B	1T
Chinese	PPL	15.45	**15.20**
	ACC	45.35	**48.41**
English	PPL	10.18	**9.89**
	ACC	37.47	**42.39**

also changes the learning rate seen by the Chinese stage. Cross-scale comparisons become meaningful only after these quantities are controlled. Once the target-stage schedule is fixed, a longer English stage provides a better initialization over the tested range; moderate repetition strengthens target adaptation, but eventually saturates.

The four-language experiment shows that the positive cross-scale trend is not specific to Chinese, including under severe data imbalance. It remains narrower than a universal multilingual claim: four target languages do not span the diversity of scripts, tokenization regimes, domains, or language families. Strict staging makes the optimization effect visible, whereas production systems may prefer replay, periodic target batches, or mixed second stages to retain more English capability. The 80/20 large-scale configuration illustrates one compromise but does not determine an optimal ratio. In practice, the choice should reflect the desired point on the target-performance versus retention frontier.

7 Limitations

The main curves are single runs, so trend variance and statistical significance are unknown; this is especially relevant where the absolute Bengali change is small. Chinese provides a controlled fixed-data target, while Turkish, Hungarian, and Bengali broaden the evaluation but still cover only four target languages in total. Most experiments use 1.3B models, and the 2.5B-parameter case study contains only two budgets. Strict two-stage training differs from interleaved production pipelines. The shared-token control reverses language order and does not establish a causal representation mechanism. Translation, repetition, and loss weighting follow different compute pathways, and the 80/20 large-scale mixture is not optimized. Broader language and model-scale validation is therefore required before generalizing the method.

8 Conclusion

When target-language data are fixed, scaling high-resource data changes both the target share under proportional mixing and the target-stage learning-rate trajectory under staged training. Fixing that schedule makes the Chinese trend positive over the evaluated interval, with moderate repetition

adding gains before saturation. ScaleX and ScaleX-DR reproduce the trend at two Chinese-data budgets and across four multilingual targets, but trade target gains for English retention. Cross-lingual data scaling is therefore a resource-allocation problem rather than a free transfer effect.

Reproducibility Statement

The main text states model size, sequence length, batch size, learning-rate range, data budgets, evaluation suites, and the absence of seed-level error bars. The appendix reports the schedule controls, exposure definitions, synthetic-data construction, large-scale retention setting, and shared-token control. Reported results use measured values, and unavailable metrics are left unreported.

AI Use Statement

Generative AI tools were used to improve the readability and organization of the manuscript and to assist with literature retrieval and discovery. The authors verified all cited sources and manually reviewed and revised all AI-assisted text. These tools were not used to generate experimental measurements or results. The authors take responsibility for the final content of this work.

References

Kosuke Akimoto, Taiki Miyagawa, and Masafumi Oyamada. M3 scaling law: Optimizing multi-epoch, multi-lingual, and multi-stage training for low-resource language models. *arXiv preprint arXiv:2410.12325*, 2024.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of EMNLP*, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Dami Choi, Derrick Xin, Hamid Dadkhahi, Justin Gilmer, Ankush Garg, Orhan Firat, Chih-Kuan Yeh, Andrew M. Dai, and Behrooz Ghorbani. Order matters in the presence of dataset imbalance for multilingual learning. In *Advances in Neural Information Processing Systems*, 2023.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *International Conference on Learning Representations*, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, et al. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, 2020.

Negar Foroutan, Paul Teiletche, Ayush Kumar Tarun, and Antoine Bosselut. Revisiting multilingual data mixtures in language model pretraining. *arXiv preprint arXiv:2510.25947*, 2026.

Yifei He, Alon Benhaim, Barun Patra, Praneetha Vaddamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, and Xia Song. Scaling laws for multilingual language models. In *Findings of ACL*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, et al. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Haonan Li, Yixuan Zhang, Fajri Koto, et al. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of ACL*, 2024.

Tianjian Li, Haoran Xu, Weiting Tan, Kenton Murray, and Daniel Khashabi. Upsample or upweight? balanced training on heavily imbalanced datasets. In *Proceedings of NAACL-HLT*, 2025a.

Zihao Li, Shaoxiong Ji, Hengyu Luo, and Jorg Tiedemann. Rethinking multilingual continual pretraining: Data mixing for adapting llms across languages and resources. In *Conference on Language Modeling*, 2025b.

Shayne Longpre, Sneha Kudugunta, Niklas Muennighoff, I-Hung Hsu, Isaac Caswell, Alex Pentland, Sercan O. Arik, Chen-Yu Lee, and Sayna Ebrahimi. Atlas: Adaptive transfer scaling laws for multilingual pretraining, finetuning, and decoding the curse of multilinguality. In *International Conference on Learning Representations*, 2026.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, et al. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*, 2023.

Skyler Seto, Maartje Ter Hoeve, Richard He Bai, Natalie Schluter, and David Grangier. Training bilingual LMs with data constraints in the targeted language. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19096–19122. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.977.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Mirac Suzgun, Nathan Scales, Nathanael Scharli, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of ACL*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Pablo Villalobos, Jaime Sevilla, Lennart Besiroglu, Marius Hobbhahn, and Anson Ho. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*, 2022.

Jindong Wang, Yifan Lu, Maurice Weber, et al. Multilingual language model pretraining using machine-translated data. *arXiv preprint arXiv:2502.13252*, 2025a.

Zhijun Wang, Jiahuan Li, Hao Zhou, et al. Investigating and scaling up code-switching for multilingual language model pre-training. In *Findings of ACL*, 2025b.

Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. To repeat or not to repeat: Insights from scaling llm under token-crisis. In *Advances in Neural Information Processing Systems*, 2023.

Linting Xue, Noah Constant, Adam Roberts, et al. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL-HLT*, 2021.

Haneul Yoo, Chanjun Park, Seongjin Yun, Alice Oh, and Heuiseok Lee. Code-switching curriculum learning for multilingual transfer in llms. In *Findings of ACL*, 2025.

Wenting Zheng, Wenxuan Pan, Xu Xu, Libo Qin, Linlin Yue, and Ming Zhou. Breaking language barriers: Cross-lingual continual pre-training at scale. In *Proceedings of EMNLP*, 2024.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, et al. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of NAACL*, 2024.

A TRAINING-SCHEDULE CONTROLS

All main experiments use independently initialized models. Neither stage uses learning-rate warmup. In Cos-Cos, the first-stage learning-rate curve matches the corresponding portion of Cos, while the second stage restarts a fixed cosine schedule. The comparison between Cos and Cos-Cos therefore isolates the target-stage restart while retaining the first-stage schedule. ScaleX additionally changes the first stage to a constant learning rate and evaluates the complete protocol across high-resource data scales.

The main text reports averages for compactness. PPL and downstream metrics are evaluated on held-out sets. Each scale uses a separately initialized model rather than a checkpoint continued from the preceding scale. This construction makes the high-resource data budget the controlled scaling variable while keeping the target-stage schedule fixed.

B DATA EXPOSURE AND INTERVENTION DEFINITIONS

We distinguish unique target-language data from processed exposure. Repetition factors increase the number of times existing target documents are processed; they do not add unique documents. Translation and code-switching replace English tokens while preserving the total token budget, balanced repetition adds processed exposure in both languages, and loss weighting changes the training objective without duplicating input examples. Accordingly, we interpret these comparisons as intervention-level diagnostics because they differ in token composition, processed exposure, and optimization pathway.

C LARGE-SCALE RETENTION SETTING

The strict English-to-Chinese setup isolates schedule coupling by exposing a fixed target stage to the learning-rate trajectory induced by progressively longer English stages. The 2.5B-parameter experiments evaluate a retention-oriented mixed-stage setting at 500B and 1T tokens. Their second stage uses an 80/20 Chinese–English mixture to preserve English exposure while scaling the overall training budget. Because the mixture ratio is fixed and no matched proportional-mixing baseline is available, we interpret these results as evidence of large-scale feasibility rather than as an estimate of an optimal mixture.

D TRANSLATION DATA CONSTRUCTION

Translation uses a segmentation–translation–merge pipeline. Long documents are segmented into sentences, translated with Qwen2.5-Instruct through vLLM, and reassembled in their original order. We select Qwen2.5-7B-Instruct as a quality–cost operating point. Table 4 reports the original model-selection measurements. Larger models require substantially more generation time, while every translated variant remains worse than native Chinese on this PPL diagnostic.

Table 4: Translation-model data quality and generation cost.

Model	Perplexity	GPU Hours
Qwen2.5-1.5B	34.45	810
Qwen2.5-3B	35.76	979
Qwen2.5-7B	30.46	1,696
Qwen2.5-14B	28.31	2,640
Qwen2.5-32B	28.47	6,636
Native Chinese	11.58	–

E CODE-SWITCHING DATA CONSTRUCTION

Code-switching follows SynCS: 20% of English samples are selected and a fraction of their content is replaced with Chinese at token or sentence level using a fine-tuned Qwen2.5-3B-Instruct model. We use a 30% replacement ratio because it gives the highest reported downstream accuracy in the measured sweep, although the lowest PPL occurs at 90%. Table 5 therefore motivates an operating point rather than a universal optimum.

Table 5: Code-switching replacement-ratio ablation.

Ratio	Perplexity	Accuracy
10%	22.3	24.7
30%	22.1	**25.8**
50%	22.0	24.4
70%	22.1	23.7
90%	**21.9**	25.2

F SHARED-TOKEN CONSTRUCTION

Tokens are divided into Chinese characters, Latin-alphabet tokens, and other tokens such as punctuation, digits, and special symbols. Table 6 reports the resulting sample composition. Most Chinese samples contain all three categories; deleting English fragments from the Chinese target corpus would therefore perturb the target data. The controlled experiment reverses language roles and filters the first-stage Chinese data instead.

Table 6: Proportion of samples containing each token category.

Zh	En	Other	Proportion	
			Chinese	English
✓	✓	✓	98.20%	2.80%
✓	✗	✓	1.80%	0.00%
✗	✓	✓	0.00%	97.20%

Table 7 reports all six configurations under the reversed-language, token-filtered setting described above.

Table 7: Full shared-token control.

#	Shared Token			Scalable	English Perplexity		
	Zh	En	Ot		4B	40B	80B
1				✗	12.54	12.63	13.02
2	✓			✗	12.59	12.55	13.01
3		✓		✓	12.26	11.50	11.34
4			✓	✓	12.42	11.74	11.49
5		✓	✓	✓	12.16	11.34	11.16
6	✓	✓	✓	✓	12.18	11.36	11.17