On the Limits of Model Merging for Multilinguality in Pre-Training

Seth Aycock **Fedor Vitiugin** **Aleksandr Umnov**
Christof Monz **Khalil Sima'an**
University of Amsterdam University of Turku Booking.com
s.aycock@uva.nl

Abstract

Endowing models with consistent multilingual performance can be achieved by *mixing* pre-training data, or post-training approaches such as language-specific model *merging*. In this work, we test whether merging can be applied to monolingually pre-trained models. We conduct a controlled study on the efficacy of mixed, merged, and monolingual pre-training setups. We find that while monolingual pre-training results in strong in-language performance, merging any combination of monolingual models leads to performance collapse due to interference. Our analysis suggests representational similarity is a prerequisite for model merging. We therefore conclude that the flexibility of merging in fine-tuning does not extend trivially to language-specific pre-training.

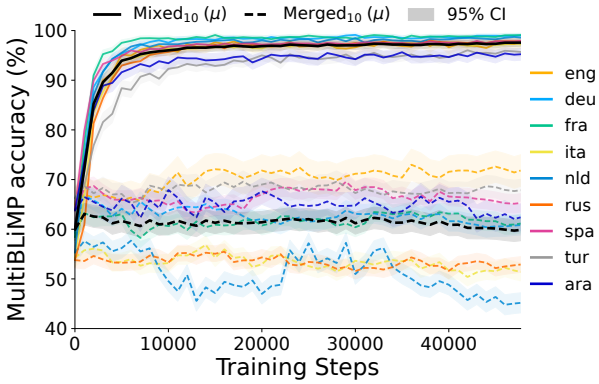

Figure 1: Average and per-language MultiBLiMP accuracies over training for a mixed data model (Mixed$_{10}$) and a linearly merged model (Merged$_{10}$) combining 10 monolingual models at each training step. Scores near 100% are expected for grammatically competent models; 50% indicates random chance. Equal data mixing in pre-training gives consistent multilingual performance, while merging leads to near-random performance.

1 Introduction

Multilinguality is a key desideratum in training large language models (LLMs), but consistent capabilities across languages are difficult to achieve (Moskvina et al., 2026), due to both data choices (Shani et al., 2026) and modelling choices (Chang et al., 2024). Common approaches involve mixed pre-training for early language exposure (Foroutan et al., 2025; Longpre et al., 2026), or fine-tuning a pre-trained model on language-specific data (Aggarwal et al., 2024; Salamanca et al., 2026). These methods are performant but somewhat inflexible, requiring further adaptation to modify language coverage.

Model merging has emerged as a cheap, post-hoc, and flexible method for improving models' multilingual or multi-task capabilities (Ilharco et al., 2023; Bandarkar and Peng, 2025). Standardly, task or language-specific experts are *fine-tuned* from a shared *pre-trained* model, then merged (Chronopoulou et al., 2024; Parović et al., 2024; Yang et al., 2024; Cohere et al., 2025; Zeng et al., 2025), optionally with parameter interference mitigating methods applied (Yu et al., 2024). In fine-tuning settings with only a few languages or tasks, this setup can outperform data mixing (Aakanksha et al., 2024; Yang et al., 2025). However, the picture for language-specific merging is less clear: recent work suggests multilingual merging for fine-tuned models can suffer from weight-space incompatibilities (Gain et al., 2026) and sometimes underperform mixed fine-tuning (Glocker et al., 2025). Further, while fine-tuning is cheaper than pre-training, the choice of base model constrains capabilities for all downstream evaluations. This raises the question: can merging, and its potential benefits, be extended to language-specific *pre-training*?

We investigate this research question through controlled experiments: we pre-train a mixed multilingual model and test against comparable open-source monolingual models, evaluating various merging methods, and testing all models across 10 languages over 5 benchmarks. We observe that

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 159–169

merging monolingual pre-trained models leads to a collapse in models' capabilities due to interference. Conversely, mixed data training results in consistent but modest multilingual performance. Our analysis suggests that independently pre-trained models' representations diverge too far for merging to succeed. We conclude that the flexibility of model merging does not extend trivially to independently pre-trained models despite homogenous architectures, and that some alignment is required before language-specific adaptation.

2 Experimental Methodology

Model Pre-Training We use the open-source HPLT 2.15B monolingual decoder-only models (OpenEuroLLM, 2025), which were each trained on 100B tokens of HPLT language-specific data (de Gibert et al., 2024; Burchell et al., 2025). HPLT v2 (Arefyev et al., 2025) is a large-scale open multilingual corpus constructed through web-crawling and language filtering. HPLT models use the Gemma-3 tokenizer and follow the Llama architecture (Touvron et al., 2023) with 24 layers, 32 attention heads, and a sequence length of 2048. We pre-train a mixed data model with the same architecture (Mixed_{10}) on 100B tokens from 10 languages (10B tokens per language), to compare monolingual, mixed, and merged pre-training strategies. Pre-training was run with Megatron-LM (Shoeybi et al., 2020), using 16 nodes with AMD MI250x GPUs for 3,000 GPU hours on the LUMI supercomputer, for an estimated carbon footprint of 59 kg CO_2 per model.

While these models are relatively small and require fewer pre-training resources than closed-source alternatives, achieving state-of-the-art performance is not the primary objective of this work. Instead, we target controllability and scientific transparency. This design provides a foundation for reproducible experiments comparing different multilingual pre-training strategies, letting us systematically isolate and evaluate cross-lingual and monolingual performance across tasks[1].

Model Merging We use linear weight averaging (Wortsman et al., 2022) to equally merge the 10 monolingual HPLT models giving Merged_{10} models; and we merge all 45 bilingual combinations of HPLT models for analysis in Section 4. We also apply the interference-mitigation method, TIES (Yadav et al., 2023); here, we calculate a *task vector* for each trained model by subtracting the base models' parameters, which here is the shared random initialisation. After this, TIES prunes lower magnitude parameters given a threshold, and resolves sign conflicts, then linearly adds the resulting task vectors to the base model. We also test DARE-TIES (Yu et al., 2024), which randomly drops parameters and rescales before applying TIES. We note these methods are designed to work on fine-tunes of a base model where task vectors are of small magnitudes; however our pre-trained models see 100B tokens and we therefore expect both large magnitude differences in the task vectors calculated from the shared initialisation, and varied weight-spaces, which may compromise these methods. We merge models using Mergekit (Goddard et al., 2024).

Baselines We test similarly-sized baselines: EuroLLM (1.7B) (Martins et al., 2024) as an open-data multilingual pre-trained model, and Gemma-2 (2B) (Riviere et al., 2024) and Tiny Aya Base (3.35B) (Salamanca et al., 2026), with open-weights but private data mixes. These models were trained on 20-60x more tokens than HPLT models.

Evaluation We test models on 5 multilingual benchmarks: MultiBLiMP (Jumelet et al., 2026) testing formal language competence on linguistic minimal pairs; Belebele (Bandarkar et al., 2024) to test reading comprehension; multilingual HellaSwag (Lai et al., 2023) for functional language understanding; X-CSQA (Lin et al., 2021) testing common-sense reasoning; and FLORES-200 translation from eng–xxx (Costa-jussà et al., 2024) to test cross-lingual generation. We select 10 diverse, high-resource languages across the benchmarks: Arabic (ara), German (deu), English (eng), French (fra), Italian (ita), Dutch (nld), Russian (rus), Spanish (spa), Turkish (tur), and Mandarin Chinese (zho)[2]. Belebele, HellaSwag, X-CSQA, and MultiBLiMP are evaluated with token-normalised accuracy, and run in a cloze-formulation measuring log-likelihoods; and we evaluate FLORES-200 with ChrF++ (Popović, 2017). We calculate 95% confidence intervals as $1.96 \times \text{SE}$, estimating standard error (SE) from 1000 resamples over per-item scores.

3 Results

Monolingual models are strong oracles Results in Table 1 show the HPLT_1 monolingual models

[1] We make our models and code openly available.

[2] We note that MultiBLiMP lacks zho, X-CSQA lacks tur, and for FLORES in eng, we use fra–eng.

Model	Params/B	Tokens/B	MultiBLiMP		Belebele		HellaSwag		X-CSQA		FLORES	
			μ	CV%	μ	CV%	μ	CV%	μ	CV%	μ	CV%
Multilingual baselines												
EuroLLM-1.7B$_{35}$	1.7	4000	95.9	3.62	36.5	7.34	45.9	13.44	31.6	15.65	37.0	33.89
Gemma-2-2B	2.6	2000	94.8	3.46	48.6	11.16	52.2	17.45	43.8	21.86	**42.2**	27.72
Tiny-Aya-Base$_{70}$	3.35	6000	97.2	2.14	**54.3**	8.96	**56.8**	13.15	**47.8**	14.48	39.2	35.65
Monolingual experts												
HPLT$_1$ (μ)	2.15	100*	**98.7**	0.78	39.3	7.74	48.6	14.30	39.0	18.45	37.3	25.99
Mixed pre-training												
Mixed$_{10}$	2.15	100	97.6	1.39	34.9	13.50	37.5	17.97	32.3	21.70	30.0	33.28
Merged$_{10}$ experts												
Linear	2.15	1000	59.9	14.23	23.9	6.43	25.1	6.18	20.3	7.93	2.1	17.24
TIES	2.15	1000	60.0	10.79	24.4	6.11	25.3	6.02	20.4	6.02	12.3	48.11
DARE-TIES	2.15	1000	60.8	10.63	24.4	4.40	25.6	4.53	20.4	6.83	1.7	24.67

Table 1: Mean results over 10 languages across tasks (ChrF++ for FLORES, token-normalised accuracy % for other tasks). μ = mean performance across languages; CV = coefficient of variation (%) measuring cross-lingual consistency. Model parameters and pre-training tokens indicated in billions (B); *HPLT$_1$ monolingual experts have seen 100B tokens *each*; supported model languages, if available, in subscript. **Bold** = best result per column. Mixed pre-training gives consistent performance while merging leads to catastrophic interference and near-random results.

perform best among models on test when averaged across languages, and even outperform the stronger baselines on MultiBLiMP despite substantially shorter training, establishing a high performance ceiling. This suggests monolingual pre-training is effective for training in-language competence, even for small models, aligning with prior work (Chang et al., 2026). We note that scores near 100% are expected for grammatically competent models and do not indicate overfitting (Jumelet et al., 2026). MultiBLiMP is thus a highly discriminative benchmark for our purposes, and we focus ablation analyses here accordingly.

Merging causes performance collapse Linear merging of 10 monolingual pre-trained models leads to near-random performance across all benchmarks and languages (see Appendix A for full language results), we expect due to catastrophic interference of conflicting parameters during merging. Figure 2 shows this occurs for any number of merges: merging successive combinations of 2 to 10 models shows MultiBLiMP performance collapses from the first merge, with no clear gains on each added language. Further ablations indicate this result is agnostic to the training stage, as Figure 1 shows merging equivalent checkpoints throughout training consistently results in near-random performance compared to mixed pre-training.

Interference mitigation does not help Applying TIES and DARE-TIES merging, which prune unimportant parameters to reduce interference, yields negligible improvements across tasks, with performance still collapsing against monolingual and mixed models. Qualitative analysis of FLORES-200 outputs, seen in Appendix B, suggests *all* merged models fail to generate any meaningful text in any language. This confirms that independently pre-trained models are too distinct in both magnitude and weight-space alignment compared to fine-tuned versions of a pre-trained model, compromising these methods' effectiveness.

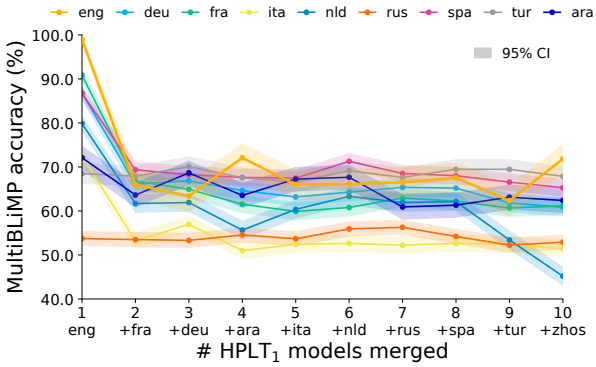

Figure 2: MultiBLiMP accuracy per-language for linear merges of 2-10 monolingual HPLT$_1$ models (randomly ordered; only combination matters). At each step one model is evaluated on all languages. We note HPLT$_{eng}$ is incidentally performant in various languages. Performance collapses to near-random from the first merge.

Mixed pre-training gives modest results Our Mixed$_{10}$ model achieves consistently good perfor-

mance across languages, underperforming HPLT_1 experts but performing similarly to EuroLLM. This suggests mixed pre-training is a good compromise for multilingual settings, while monolingual models are more performant when the task language is pre-defined. This result may arise from both limited token-exposure compared to baselines, and smaller model capacity per-language compared to HPLT_1 models. However, Figure 1 indicates that Mixed_{10} still reaches a high MultiBLiMP accuracy across languages after only 10000 training steps.

4 Analysis

Our results pose the question of *why* merging harms performance, and to what extent we can predict merge failure or success. We explore this question by calculating performance drops from merging, and testing their correlations with model similarity measures. We linearly merge all bilingual combinations of the 10 HPLT_1 models, giving 45 merged models. For each language pair, we calculate Δ as the mean *drop* in MultiBLiMP accuracy on the two merged languages between the HPLT_1 experts and the merged bilingual model. We calculate *parametric* similarity between merged models via layer-wise cosine similarity, mean stable rank difference, and mean L2 norm difference, but find no significant correlations with Δ (see Appendix C).

We then test *representational* similarity measures, which are invariant to parameter symmetries that can confound weight-space metrics (Klabunde et al., 2025). We calculate mean layer-wise linear centred kernel alignment (CKA) (Kornblith et al., 2019) of model representations, by passing the English FLORES-200 devtest through all 10 HPLT_1 models[3], then averaging layer-wise CKA. Figure 3 shows higher CKA significantly correlates with smaller Δ ($r = 0.447$, $p < 0.005$), suggesting cross-model representational similarity reduces interference and is a useful predictor of merge failure.

5 Discussion

The inefficacy of merging independently pre-trained monolingual models across all tasks, languages, and merge methods contrasts with the success of merging language-specific fine-tuned models for multilinguality (Aakanksha et al., 2024; Bandarkar and Peng, 2025). Our results align with

[3]We assume all monolingual models have been exposed to non-zero amounts of English text given difficulties in document-level language identification (Fedorova et al., 2026).

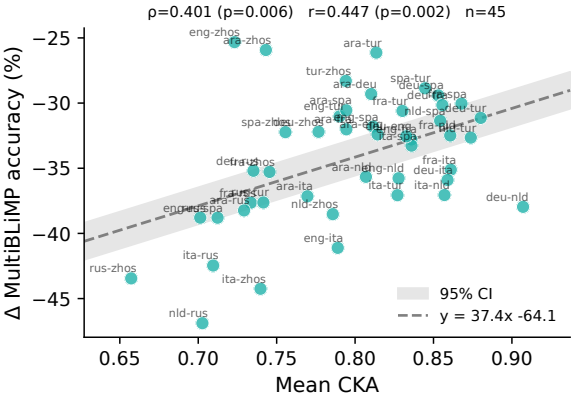

Figure 3: Mean layer-wise CKA between monolingual HPLT_1 models significantly correlates (in terms of Spearman's ρ and Pearson's r) with smaller merge performance drop Δ from monolingual to bilingual merged models. This suggests increasing representational similarity improves merge success.

prior work suggesting that representational similarity is an essential ingredient for successful merging of heterogenous models (Shaheen et al., 2026). This could be achieved through through shared pre-training (Li et al., 2022), weight-space alignment methods (Ainsworth et al., 2023), or adapting unsupervised embedding alignment techniques (Lample et al., 2018), the latter two providing open directions for future work.

While representational similarity is one ingredient for merging, recent work suggests merging heterogenous models also requires interference mitigation strategies (Chen et al., 2026). Our results indicate that the converse holds: interference mitigation alone is not sufficient to overcome the representational divergence of monolingually pre-trained models. While future ablations into the effects of initialisation and partial pre-training may prove enlightening, we speculatively conclude that merging pre-trained models requires both representational similarity and interference mitigation methods.

6 Conclusion

We find that merging small, independently pre-trained monolingual models is ineffective, despite applying interference-mitigation strategies. Mixed multilingual data pre-training is a simpler, albeit less flexible, approach to achieve modest but consistent performance across languages. Our results suggest merging requires representational similarity between models, but we find that independently pre-training monolingual models leads to divergent,

heterogeneous representations. Therefore, our core recommendation is that representational alignment, most straightforwardly achievable through some initial mixed pre-training, is required for effective language-specific model merging.

Limitations

We acknowledge the following limitations of this work. First, we only use models around 2B parameters to maintain strict control over experimental conditions; and we train our mixed model on 100B tokens, which is a smaller scale than state-of-the-art models of similar sizes. Our focus is not on leading performance, and while 100B tokens is enough to see clear trends, further training would strengthen results. Next, we test three standard merging strategies, leaving exploration of stronger alignment-based approaches to future work. We train only one mixed model as a controlled comparison against the HPLT monolingual models, since any additional pre-training required substantial resources. Similarly, due to resource limitations, it was not feasible to scale results to larger models or bigger datasets. We note recent work explores scaling laws for monolingual and multilingual pre-training (Longpre et al., 2026), and future work could extend our research to explore the scaling properties of merging pre-trained models to understand whether our findings hold for larger models.

Acknowledgements

SA was funded in part by the UvA's Language Sciences for Social Good project, the City of Amsterdam, and the Netherlands Organization for Scientific Research (NWO) under project numbers VI.C.192.080 and 2023.017. SA and AU are grateful to Booking.com, where they first collaborated during SA's internship. This partnership led to the present collaboration, which is unrelated to Booking.com and was carried out independently of the company. This project has received funding from the Horizon Europe research and innovation programme of the European Union under Grant No. 101070350 and Grant No. 101195233 (Digital Europe programme of the European Union). The authors thank CSC (Finland) for computational resources and support. We further thank both colleagues from the UvA LTL and Helsinki NLP groups for providing helpful feedback prior to submission, and the anonymous reviewers for their constructive efforts to improve this research.

References

Aakanksha, Arash Ahmadian, Seraphina Goldfarb-Tarrant, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. 2024. Mix Data or Merge Models? Optimizing for Diverse Multi-Task Learning. *arXiv preprint*. ArXiv:2410.10801 [cs].

Divyanshu Aggarwal, Sankarshan Damle, Navin Goyal, Satya Lokam, and Sunayana Sitaram. 2024. Towards exploring continual fine-tuning for enhancing language ability in large language model. In *NeurIPS 2024 Workshop on Scalable Continual Learning for Lifelong Foundation Models*.

Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. 2023. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*.

Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Laurie Burchell, Pinzhen Chen, Mariia Fedorova, Ona de Gibert, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Andrey Kutuzov, Veronika Laippala, Bhavitvya Malik, Farrokh Mehryary, Vladislav Mikhailov, Amanda Myntti, Dayyán O'Brien, and 8 others. 2025. HPLT's second data release. In *Proceedings of Machine Translation Summit XX: Volume 2*, pages 101–102, Geneva, Switzerland. European Association for Machine Translation.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 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)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Lucas Bandarkar and Nanyun Peng. 2025. The Unreasonable Effectiveness of Model Merging for Cross-Lingual Transfer in LLMs. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 131–148, Suzhuo, China. Association for Computational Linguistics.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. An Expanded Massive Multilingual Dataset for High-Performance Language Technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. When Is Multilinguality

a Curse? Language Modeling for 250 High- and Low-Resource Languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2026. Goldfish: Monolingual Language Models for 350 Languages. *arXiv preprint*. ArXiv:2408.10441 [cs].

Shilian Chen, Jie Zhou, Qin Chen, Wen Wu, Xin Li, Qi Feng, and Liang He. 2026. Can Heterogeneous Language Models Be Fused? *arXiv preprint*. ArXiv:2604.01674 [cs].

Alexandra Chronopoulou, Jonas Pfeiffer, Joshua Maynez, Xinyi Wang, Sebastian Ruder, and Priyanka Agrawal. 2024. Language and Task Arithmetic with Parameter-Efficient Layers for Zero-Shot Summarization. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 114–126, Miami, Florida, USA. Association for Computational Linguistics.

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

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, pages 1–6.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A New Massive Multilingual Dataset for High-Performance Language Technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Mariia Fedorova, Nikolay Arefyev, Maja Buljan, Jindřich Helcl, Stephan Oepen, Egil Rønningstad, and Yves Scherrer. 2026. OpenLID-v3: Improving the Precision of Closely Related Language Identification – An Experience Report. *arXiv preprint*. ArXiv:2602.13139 [cs] version: 2.

Negar Foroutan, Paul Teiletche, Ayush Kumar Tarun, and Antoine Bosselut. 2025. Revisiting Multilingual Data Mixtures in Language Model Pretraining. *arXiv preprint*. ArXiv:2510.25947 [cs].

Baban Gain, Asif Ekbal, and Trilok Nath Singh. 2026. One Model to Translate Them All? A Journey to Mount Doom for Multilingual Model Merging. *arXiv preprint*. ArXiv:2604.02881 [cs].

Kevin Glocker, Kätriin Kukk, Romina Oji, Marcel Bollmann, Marco Kuhlmann, and Jenny Kunz. 2025. Grow Up and Merge: Scaling Strategies for Efficient Language Adaptation. *arXiv preprint*. ArXiv:2512.10772 [cs].

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A Toolkit for Merging Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. 2025. Similarity of Neural Network Models: A Survey of Functional and Representational Measures. *ACM Comput. Surv.*, 57(9):242:1–242:52.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. 2022. Branch-train-merge: Embarrassingly

parallel training of expert language models. In *First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022*.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common Sense Beyond English: Evaluating and Improving Multilingual Language Models for Commonsense Reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

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

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

Natalia Moskvina, Raquel Montero, Masaya Yoshida, Ferdy Hubers, Paolo Morosi, Walid Irhaymi, Jin Yan, Tamara Serrano, Elena Pagliarini, Fritz Günther, and Evelina Leivada. 2026. Multilingual Large Language Models do not comprehend all natural languages to equal degrees. *arXiv preprint*. ArXiv:2602.20065 [cs].

OpenEuroLLM. 2025. Release of 38 Monolingual 2.15B LLMs Trained on HPLT v2.

Marinela Parović, Ivan Vulić, and Anna Korhonen. 2024. Investigating the Potential of Task Arithmetic for Cross-Lingual Transfer. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 124–137, St. Julian's, Malta. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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, Anton Tsitsulin, and 178 others. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint*. ArXiv:2408.00118 [cs].

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

Nour Shaheen, Sarath Chandar, Boris Knyazev, and Ekaterina Lobacheva. 2026. Is depth heterogeneity a barrier to model merging? In *Third Workshop on Test-Time Updates (Main Track)*.

Chen Shani, Yuval Reif, Nathan Roll, Dan Jurafsky, and Ekaterina Shutova. 2026. The Roots of Performance Disparity in Multilingual Language Models: Intrinsic Modeling Difficulty or Design Choices? *arXiv preprint*. ArXiv:2601.07220 [cs] version: 2.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *arXiv preprint*. ArXiv:1909.08053 [cs].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*. ArXiv:2307.09288 [cs].

Rob van der Goot, Esther Ploeger, Verena Blaschke, and Tanja Samardzic. 2025. DistaLs: a comprehensive collection of language distance measures. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 307–318, Suzhou, China. Association for Computational Linguistics.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, pages 23965–23998. PMLR.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. 2023. TIES-Merging: Resolving Interference When Merging Models. *Advances in Neural Information Processing Systems*, 36:7093–7115.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model Merging in LLMs, MLLMs, and Beyond: Methods, Theories, Applications and Opportunities. *arXiv preprint*. ArXiv:2408.07666 [cs].

Jinluan Yang, Dingnan Jin, Anke Tang, Li Shen, Didi Zhu, Zhengyu Chen, Ziyu Zhao, Daixin Wang, Qing Cui, Zhiqiang Zhang, Jun Zhou, Fei Wu, and Kun Kuang. 2025. Mix Data or Merge Models? Balancing the Helpfulness, Honesty, and Harmlessness of Large Language Model via Model Merging. *arXiv preprint*. ArXiv:2502.06876 [cs].

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.

Siqi Zeng, Yifei He, Weiqiu You, Yifan Hao, Yao-Hung Hubert Tsai, Makoto Yamada, and Han Zhao. 2025. Efficient Model Editing with Task Vector Bases: A Theoretical Framework and Scalable Approach. *arXiv preprint*. ArXiv:2502.01015 [cs].

A Full Task Results

We provide full results per task across all available languages for each task in Tables 2–6. We note that findings from Table 1 are reflected here: Monolingual expert models perform best on MultiBLiMP, and otherwise multilingual baselines are strong. Mixed pre-training gives consistent but not leading performance across languages, while any merging leads to performance collapse to near-random across all languages.

B Qualitative Analysis of Outputs

We perform a brief qualitative analysis of model generations for the FLORES translation task, with examples shown in Table 7. We see the following error modes: monolingual models show input copying and occasionally fail to translate into the target language. Some models including Tiny-Aya fail to stop generating after translating. The Mixed_{10} model generally provides a functional translation but sometimes still copies inputs. Finally, all merging settings lead to complete output collapse: either generating strings of numerals, or nonsense tokens. This suggests any merging leads models to completely lose their generative language capabilities.

C Model Similarity Correlation

In Table 8, we report results of Pearson's r and Spearman's ρ correlation tests between Δ performance after merging and measures of model and language similarity: layer-wise cosine similarity, mean layer-wise rank difference, absolute mean L2 norm difference across layers, mean layer-wise CKA (Kornblith et al., 2019), and Lang2vec typological distance (with kNN imputation of missing features from similar languages) of the 2 merged languages (Littell et al., 2017; van der Goot et al., 2025).

Measure	ρ	p_S	r	p_P
Mean CKA	0.40^{**}	0.006	0.45^{**}	0.002
Lang2vec (kNN)	0.35^{*}	0.017	0.32^{*}	0.033
Mean Rank Δ	0.15	0.340	0.11	0.458
Cosine Similarity	0.11	0.468	0.11	0.460
Mean L2 Norm Δ	0.03	0.830	0.04	0.798

Table 8: Spearman and Pearson correlations of various model- and language-similarity measures against Δ MultiBLiMP accuracy for 45 bilingual merges of monolingual models. $^{*}p < 0.05$, $^{**}p < 0.01$.

We observe no significant correlation between parametric similarity measures such as layer-wise cosine similarity. However, we observe a surprising correlation between *increasing* Lang2vec typological distance and *decreasing* Δ, i.e. smaller performance drops. This indicates that combining more typologically similar languages results in worse performing merges. We leave exploration of this to future work.

Model	eng	deu	fra	ita	nld	rus	spa	tur	ara	zho
Multilingual baselines										
Tiny-Aya-Base	**99.0**	99.0	99.3	97.2	96.3	97.8	98.0	92.8	95.4	–
EuroLLM-1.7B	97.8	98.0	98.8	96.4	96.9	97.2	97.5	87.8	93.0	–
Gemma-2-2B	98.7	97.4	97.9	93.9	94.5	95.1	96.3	89.0	90.6	–
Monolingual experts										
$HPLT_1$ (per-language)	**99.0**	**99.1**	**99.4**	**98.3**	**99.1**	**99.2**	**98.4**	**98.7**	**96.9**	–
Mixed pre-training										
$Mixed_{10}$	97.5	99.0	99.1	97.2	98.6	97.8	97.9	95.7	95.2	–
$Merged_{10}$ experts										
Linear	71.8	60.8	61.2	51.4	45.2	52.9	65.3	67.8	62.4	–
TIES	67.1	62.7	57.4	52.5	50.6	53.7	65.0	66.1	64.8	–
DARE-TIES	67.9	60.4	60.4	55.0	50.4	54.3	66.5	68.3	63.9	–

Table 2: MultiBLIMP accuracy (↑) for each model and language. Best per-language entry in bold. – = not available.

Model	eng	deu	fra	ita	nld	rus	spa	tur	ara	zho
Multilingual baselines										
EuroLLM-1.7B	36.8	38.9	38.6	35.2	37.2	38.6	37.4	34.1	34.6	37.1
Gemma-2-2B	59.0	52.3	52.6	46.3	48.2	49.4	52.9	42.1	46.2	47.2
Tiny-Aya-Base	**61.0**	**58.3**	**58.6**	**52.0**	**53.6**	**54.3**	**57.4**	**47.0**	**57.2**	**52.9**
Monolingual experts										
$HPLT_1$ (per-language)	44.2	38.9	40.1	35.2	39.6	40.4	40.8	34.9	35.4	30.0
Mixed pre-training										
$Mixed_{10}$	38.3	38.8	40.8	36.3	37.9	36.6	38.4	34.3	36.1	27.2
$Merged_{10}$ experts										
Linear	24.2	21.9	23.7	27.1	24.8	22.7	24.0	23.0	24.6	22.6
TIES	22.2	23.8	23.4	23.3	25.7	23.9	25.7	24.0	24.7	22.2
DARE-TIES	24.1	24.7	23.8	23.2	24.6	23.2	25.4	24.1	24.8	22.6

Table 3: Belebele accuracy (↑) per model and language. Best per-language entry in bold.

Model	eng	deu	fra	ita	nld	rus	spa	tur	ara	zho
Multilingual baselines										
EuroLLM-1.7B	60.1	45.9	51.3	49.1	45.8	44.9	49.1	41.1	38.9	42.0
Gemma-2-2B	**74.5**	50.7	58.6	53.1	50.6	51.9	58.8	45.9	40.7	50.1
Tiny-Aya-Base	73.5	**56.0**	**61.6**	**61.8**	**52.3**	**55.0**	**61.2**	**51.7**	**50.0**	**52.5**
Monolingual experts										
$HPLT_1$ (per-language)	65.0	44.1	47.8	45.2	45.6	45.0	51.0	45.5	36.0	34.7
Mixed pre-training										
$Mixed_{10}$	48.5	38.4	42.3	41.7	39.8	39.2	44.4	40.0	36.0	29.4
$Merged_{10}$ experts										
Linear	25.7	24.8	23.7	27.6	24.4	24.9	25.4	25.2	24.4	22.8
TIES	26.3	26.2	21.5	25.8	24.0	26.4	26.1	26.3	24.9	23.2
DARE-TIES	27.1	26.6	25.0	26.2	24.4	25.0	26.5	25.7	24.6	23.2

Table 4: HellaSwag accuracy normalised (↑) per model and language. Best per-language entry in bold.

Model	eng	deu	fra	ita	nld	rus	spa	tur	ara	zho
Multilingual baselines										
EuroLLM-1.7B	42.6	34.4	29.1	32.9	33.7	25.4	29.0	–	27.6	35.3
Gemma-2-2B	**68.2**	48.6	43.9	42.9	38.0	33.2	45.2	–	36.8	45.9
Tiny-Aya-Base	65.1	**49.9**	**47.7**	**51.1**	**45.5**	**38.8**	**48.8**	–	**43.3**	**47.6**
Monolingual experts										
$HPLT_1$ (per-language)	54.4	39.1	35.5	36.4	37.9	31.7	38.1	–	31.2	34.8
Mixed pre-training										
$Mixed_{10}$	45.7	35.8	33.7	37.0	34.7	30.6	35.2	–	30.9	24.3
$Merged_{10}$ experts										
Linear	21.5	20.1	18.3	22.5	20.4	22.1	21.3	–	19.1	19.6
TIES	21.9	20.1	19.8	22.8	21.1	20.2	20.4	–	19.3	20.3
DARE-TIES	21.3	20.4	19.2	21.7	20.7	21.8	22.0	–	19.2	18.3

Table 5: X-CSQA accuracy normalised (↑) per model and language. Best per-language entry in bold. – = not available.

Model	eng	deu	fra	ita	nld	rus	spa	tur	ara	zho
Multilingual baselines										
EuroLLM-1.7B	46.3	41.2	**59.0**	**43.3**	35.4	34.0	33.9	**48.0**	27.6	8.0
Gemma-2-2B	43.9	**52.0**	53.7	33.5	**42.7**	47.6	**47.5**	38.8	29.0	**13.5**
Tiny-Aya-Base	**51.4**	44.2	56.1	26.4	37.0	**50.6**	33.3	41.2	**40.2**	7.1
Monolingual experts										
$HPLT_1$ (per-language)	34.9	27.7	55.8	42.6	32.8	26.1	43.7	35.1	23.8	5.5
Mixed pre-training										
$Mixed_{10}$	42.2	38.5	40.1	34.5	26.7	29.7	35.6	35.0	30.3	3.5
$Merged_{10}$ experts										
Linear	2.1	1.8	3.0	1.9	2.0	2.1	2.3	1.9	1.9	2.4
TIES	15.8	15.9	16.4	17.7	16.7	4.8	16.7	14.1	3.9	4.9
DARE-TIES	1.5	2.2	1.9	2.3	1.4	1.9	0.9	1.6	1.6	2.3

Table 6: FLORES-200 ChrF++ (eng–xxx, ↑) per model and language. eng results are for fra–eng. Best per-language entry in bold.

Model	Translation
SOURCE	Del Potro had the early advantage in the second set, but this too required a tie break after reaching 6-6.
TARGET	Malgré le net avantage de Del Potro pendant le deuxième set, il a fallu passer par un tie-break une fois que le score a atteint 6-6.
TINY-AYA	Del Potro a eu un avantage précoce dans le deuxième set, mais cela aussi a nécessité un tie break après avoir atteint 6-6.
HPLT$_{eng}$	Del Potro had the early advantage in the second set, but this too required a tie break after reaching 6-6.
HPLT$_{fra}$	L'Argentin Del Potro avait l'avantage au début de la deuxième manche, mais cette fois-ci, il a fallu un tie break après avoir atteint 6-6.
MIXED$_{10}$	Del Potro avait l'avantage dans le premier set, mais ce fut un tie-break après avoir atteint 6-6.
MERGED$_{10}$	"000000000000000000000000000000000000000000000000000000000000000000000000000000...
TIES$_{10}$	y velalixobalixobalix ...,obal:'al:'al:'al.........gerixyinternetal.........geryinternetal.........gerix ...,obal.........gerstonal.........yideaideaideaideaideaideaideaideaideaideaideaideaideaide...
DARE-TIES$_{10}$	0199999999999999999999999999999999999999999999999999999999999999999999999999999...
SOURCE	The 35mm format is actually, somewhat confusingly, 36mm in width by 24mm in height.
TARGET	Le format 35 mm est en fait, et c'est assez déconcertant, de 36 mm de largeur par 24 mm de hauteur.
TINY-AYA	Le format 35 mm est en fait, de manière quelque peu déroutante, de 36 mm de large par 24 mm de haut.
HPLT$_{eng}$	Le format 35mm n'est pas vraiment confusingement confusingement 36mm en hauteur par 24mm en largeur.
HPLT$_{fra}$	Le format 35mm est en fait, quelque peu confus, 36mm en largeur par 24mm en hauteur.
MIXED$_{10}$	The 35mm format is actually, somewhat confusingly, 36mm in width by 24mm in height.
MERGED$_{10}$	000000000000000000000000000000000000000000000000000000000000000000000000000000...
TIES$_{10}$	yyyyyinternetalral...)alral...)al.........geryinternetal.........gerixyinternetal.........geryinternetal......... gerixyinternetal.........gerstoneritisnetal.........ireoveoveoveoveoveoveoveoveoveoveov...
DARE-TIES$_{10}$	000000000000000000000000000000000000000000000000000000000000000000000000000000...
SOURCE	It is related to but usually not involving alpine style ski touring or mountaineering, the latter ones done in steep terrain and requiring much stiffer skis and boots.
TARGET	Cela ressemble au ski de randonnée de type alpin, ainsi qu'à l'alpinisme, sans toutefois incorporer en général ces activités, qui se pratiquent sur des terrains accidentés et qui exigent l'utilisation...
TINY-AYA	C'est lié à mais généralement pas impliquant le ski-touring ou l'alpinisme de style alpin, ces derniers étant pratiqués dans des terrains raides et nécessitant des skis et des bottes beaucoup plus rigides. English: It is related to but usually not involving alpine style ski touring or mountaineering, the latter ones done
HPLT$_{eng}$	It is related to but usually not involving alpine style ski touring or mountaineering, the latter ones done in steep terrain and requiring much stiffer skis and boots.
HPLT$_{fra}$	It is related to but usually not involving alpine style ski touring or mountaineering, the latter ones done in steep terrain and requiring much stiffer skis and boots. Le ski de randonnée est une activité de plein air qui consiste à parcourir des distances de plus en plus longues à l'aide de skis spécialement conçus à cet effet.
MIXED$_{10}$	It is related to but usually not involving alpine style ski touring or mountaineering, the latter ones done in steep terrain and requiring much stiffer skis and boots.
MERGED$_{10}$	"000000000000000000000000000000000000000000000000000000000000000000000000000000...
TIES$_{10}$	dalaodyinternetiiobalix ...,obalix ...,obal.........yinternetal ....yideaideaideaideaideaideaideaidea ideaideaideaidea- ...al.........yodononet ' ...
DARE-TIES$_{10}$	8888889999999999999999999999999999999999999999999999999999999999999999999999999...

Table 7: Example generated translations from various models (FLORES devtest, English→French). All merging leads to catastrophic collapse of generation capabilities, while monolingual models exhibit input copying and/or a failure to translate.