Efficiently Adapting Pretrained Language Models to New Languages

Zoltan Csaki, Pian Pawakapan, Urmish Thakker, Qiantong Xu
SambaNova Systems, Inc.
Palo Alto, CA, USA
`zoltan.csaki@sambanovasystems.com`

Abstract

Recent large language models (LLM) exhibit sub-optimal performance on low-resource languages, as the training data of these models is usually dominated by English and other high-resource languages. Furthermore, it is challenging to train models for low-resource languages, especially from scratch, due to a lack of high quality training data. Adapting pretrained LLMs reduces the need for data in the new language while also providing cross lingual transfer capabilities. However, naively adapting to new languages leads to catastrophic forgetting and poor tokenizer efficiency. In this work, we study how to efficiently adapt any existing pretrained LLM to a new language without running into these issues. In particular, we improve the encoding efficiency of the tokenizer by adding new tokens from the target language and study the data mixing recipe to mitigate forgetting. Our experiments on adapting an English LLM to Hungarian and Thai show that our recipe can reach better performance than open source models on the target language, with minimal regressions on English.

1   Introduction & Related work

Multilingual large language models have become prevalent recently [1, 2, 3, 4, 5, 6], and have shown strong cross lingual knowledge and capability transfer [7, 8, 9, 10, 11, 12, 13]. However, these multilingual models tend to perform poorly on low-resource languages. On top of this, training models for low-resource languages from scratch is also challenging due to a lack of training data and prohibitive computational requirements. These challenges, along with the prevalence open sourced English models creates an interesting opportunity to see how they can be adapted to new languages quickly, without wasting resources by pretraining from scratch. While prior work [9, 10, 11, 14, 13] has studied this concept, there are two important questions that warrant further investigation.

How to efficiently encode the new language? Byte Pair Encoding (BPE) [15] tokenizers are commonly used in LLMs including GPT[16, 17], Llama [18, 19] and BLOOM [1, 2]. These tokenizers are able to encode text at the byte level so that they can generalize to characters that are outside of their vocabulary; this means that any BPE tokenizer can be used for all languages. However, the BPE tokenizer has poor tokenization efficiency if it was not trained on a given language. For example, the original English-centric GPT2 tokenizer with a vocabulary size of 50k needs to use 3.8 times more tokens to encode Thai compared to a smaller tokenizer with a vocabulary size of 5k that is trained on Thai. This will inevitably cost us 3.8 times more compute in both training and inference. Furthermore, it has been shown that models with sub-optimal tokenizers can also have worse evaluation results [20, 21]. In our work, we show how to improve tokenizer fertility[22] by replacing the least frequent tokens in the base model with tokens from the new language.

How to avoid catastrophic forgetting? Many works have shown that when continuing to train a LLM on data from a new domain, it undergoes catastrophic forgetting of the original domain it was trained on [23], and similar issues appear when training on a new language [23, 9, 24, 2, 25, 10, 26]. Different training paradigms including instruction-align[24], MAD-X [27], (IA)³ [28] are proposed to

Figure 1: Fertility score of the bilingual tokenizers (left: Hungarian, right: Thai) with different number of tokens replaced by new language. The red lines represent the original GPT-2 tokenizer (50k vocabulary), while the green lines represent the tokenizer trained purely on the new language. Every tokenizer has the same total vocabulary size. The number of replaced tokens are in 10^3 scale.

(a) Hungarian Tokenizer   (b) Thai Tokenizer
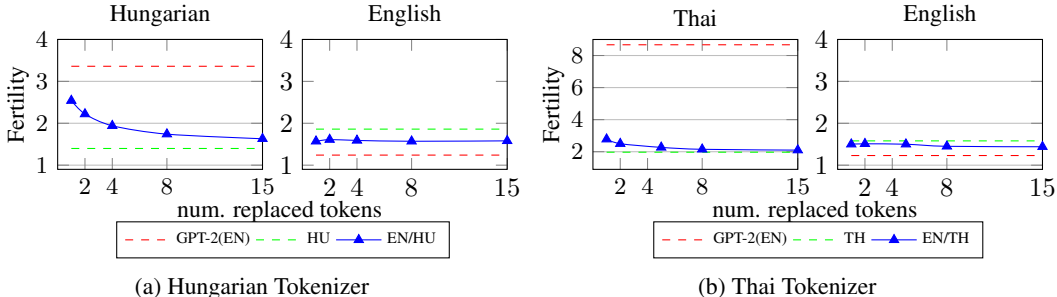
alleviate this issue, while mixing the training corpus from different languages [9, 11, 14, 29, 12, 13] is an approach shared among all the methods above. Thus, in order to avoid forgetting, we study how to use the minimum amount of mixed training data in both continuous pretraining and instruction tuning stages.

We adapt an English-centric model to Hungarian and Thai, and our evaluations show that adding new tokens and mixing training data from both languages can retain the model's English capabilities in addition to improving the models ability to learn the new language. Some contemporary works explore similar, but far less efficient methods of training LLMs on low resource languages. [30] builds an English-Arabic bilingual LLM, but they train it from scratch; while [29] builds one for English-Portuguese, but it does not optimize the tokenizer or mix the training data.

2 Implementation Details

2.1 Improving Tokenizer Efficiency

To adapt an existing tokenizer to a new language, tokens from the low resource language can be added to the existing tokenizer's vocabulary to improve its fertility. Fertility is defined as the average number of tokens per word [22], and details about how we calculated it can be found in appendix A.1. In our work, instead of extending the tokenizer's vocabulary, we replace the least frequent tokens from it with tokens from the new language. This way, we keep the model capability the same by controlling the vocabulary and embedding table size. In particular, we train a BPE tokenizer on the new language with vocabulary size k and check the number of overlapping tokens o with the original tokenizer. Then we replace the least important $k - o$ non-overlapping tokens from the original tokenizer with the new ones. We also reinitialize the corresponding embeddings in the model. For more details see appendix A.2.

As shown in figure 1, as the number of replaced tokens k increases, the fertility of the tokenizer approaches the monolingual tokenizer fertility on the new language, with minimal regressions on the English fertility. We choose to replace around 5000 tokens, which is only 10% of the overall vocabulary, because it improves the fertility by 42% on Hungarian and 73% on Thai. Note that 50% fertility drop entails two times faster training and inference. In addition, replacing more tokens beyond 5000 provides diminishing returns on the fertility, while also increases the difficulty of model adaptation due to having more randomly re-initialized token embeddings.

2.2 Training Data Mixtures

Training data for both pretraining and finetuning are prepared following the details in Section 3. Once the datasets are prepared for both languages, we shuffle them at sample level, so that every batch contains text from both languages during training. Note that in our experiments, we do not make any further transformations to either the model or the datasets, after the data is prepared on each side, so that our study is orthogonal and complementary to existing proposed methods [24, 27, 28, 9, 12] focusing on training paradigm studies.

Table 1: Each model is labeled by the language it is adapted for, followed by what style of training was done, The EN PT model is the base model for all following rows. PT stands for pretrained and IT stands for instruction tuned. Each column represents the average of all the benchmarks from a classified under a language and category, the constituent benchmarks can be found in appendix E

Tasks Metrics	English	Hungarian				Thai			
	Multi-choice Acc. (↑)	Multi-choice Acc. (↑)	QA F1 (↑)	Sum. Rouge-2 (↑)	Trans. BLEU (↑)	Multi-choice Acc. (↑)	QA F1 (↑)	Sum. Rouge-2 (↑)	Trans. BLEU (↑)
Llama2-7B	59.2%		4.3%	2.8		47.1%	48.6%	**30.9**	7.7
XGLM-7.5B	51.9%	42.8%	15.8%	0.4	1.1	46.8%	27.9%	0.0	0.2
mt0-xxl	52.7%	50/6%	30.6%	2.0	**14.0**	46.2%	**85.3%**	21.9	0.9
PULI-GPT-3SX	33.5%	39.8%	35.9%	3.1	1.3	-	-	-	-
openthaigpt-7b-chat	**60.9%**	-	-	-	-	43.7%	43.4%	26.5	4.0
EN PT	57.3%	46.2%	32.8%	0.9	1.9	46.0%	14.4%	0.0	0.2
EN PT + IT	57.3%	45.9%	34.4%	1.3	1.7	46.2%	16.1%	2.7	0.0
HU PT	55.3%	44.9%	48.5%	3.7	7.7	-	-	-	-
HU PT + IT	58.0%	**54.9%**	**64.3%**	**9.2**	6.1	-	-	-	-
TH PT	57.4%	-	-	-	-	48.4%	31.1%	11.4	3.9
TH PT + IT	56.4%	-	-	-	-	**49.9%**	48.9%	13.9	**12.5**

3 Experiments

3.1 Experimental Setup

Training is done in a two stage pipeline. The first stage is adaptive pretraining (PT) where a base pretrained English 13B GPT-2 model (B) is continuously trained on a mixture composed of the new language and English. Then, the adapted checkpoint is instruction tuned (IT) on a collection of prompt completion pairs from the new language and English. For more information see appendix B,C

We categorize all evaluation tasks into 4 categories. **Multiple Choice**, for this category we append each candidate answer to the prompt and pick the highest probability answer. **Open-ended Question Answering**, where we let the model generate an answer for each question, and report the average F1 score between the model output and the ground truth. **Summarization**, where we let the model generate a summary and report the average ROUGE-2 score between the model output and ground truth. **Translation**, where we let the model generate translated text and report the BLEU score between the model output and the ground truth. When we report the score for each category, it is the averaged score of all the evaluation tasks that we classified into that category in appendix E.

3.2 Main Results

We list all the results in table 1. The HU PT model is trained from EN PT with 50% HU, 50% EN data, and TH PT is similarly trained but with Thai data. The HU PT + IT model is trained from the HU PT checkpoint with 50% HU, 50% EN IT data. The TH PT + IT is similarly trained from the TH PT checkpoint with 50% TH, 50% EN IT data. The EN models trained on purely English data are used as baselines. We list out the dataset and training details in appendix C, D.

Table 1 shows that with our proposed training recipe, the adapted models are able to maintain the performance on English benchmarks, and improve significantly on the benchmarks of the new languages. This confirms the effectiveness of replacing the tokens in the tokenizer and mixing training data to efficiently adapt a LLM to a new language. On top of this, the adapted models perform as well or better than the state of the art baseline models we have evaluated.

3.3 Ablation Studies

3.3.1 Tokenizer

Table 2: Performance of Hungarian model with different tokenizers.

	GPT2	Bilingual
EN - Multi-choice	57.9%	**58.0%**
HU- Multi-choice	50.8%	**54.9%**
HU - QA	63.2%	**64.3%**
HU - Sum.	7.9	**9.2**
HU - Trans.	**8.8**	6.1

In this experiment, we evaluate models trained on identical data during both the continuous pretraining and IT stages. These models follow the same training recipe but use different tokenizers. Table 2 shows that **the model trained with the bilingual tokenizer performs as well or better than the model that uses the original tokenizer**, on both English and Hungarian tasks. While at the same time, the bilingual tokenizer has much better encoding efficiency as illustrated in Figure 1.

3.3.2 Pretraining data mixture

Given the same total amount of training data, we tested varying the percentage of English data (50%, 25% and 0%) in the English/Hungarian bilingual data mixture. All training is run for 30k steps. We also compare this to training a pure Hungarian model using only Hungarian data [31], a Hungarian tokenizer, from scratch for 100k steps. All the training details can be found in appendix D.1.

We summarize the comparison results in Figure 2. The first finding is that **it is effective to add in training data from the new language during pretraining**, because those models greatly outperformed the baseline English model. Second, **it is better to adapt a pretrained LLM than train a new one from scratch**, as the adapted checkpoint performs better on both languages, even though they are trained for one third as long. Third, when comparing the results from the models trained with and without English data mixed in, we can see that **mixing English data can mitigate the catastrophic forgetting on English and improve the model performance on Hungarian**. Note that there is no significant difference between mixing 50% and

Figure 2: Varying pretraining data mixtures. "EN" and "HU" models are monolingual models trained from scratch, while the other models are trained from the "EN" model with the labeled data mixture.

25% of English data during training, which implies that adaptation is not sensitive to the exact mixture ratio as long as the original language and new language are included.
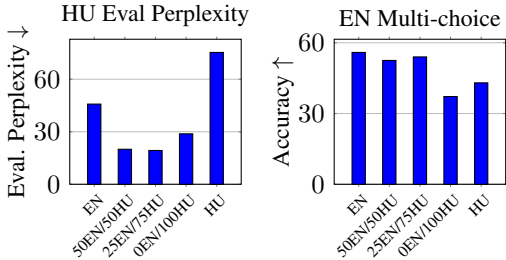
3.3.3 IT data mixture

We run all instruction tuning experiments from the Hungarian pretrained model using 6 gigabytes of instruction tuning text data (2 billion tokens) and the same training settings. Each experiment is repeated 3 times with different random dataset samples. For more details on instruction tuning datasets or training settings see appendix C.2, C.4 and D.2

There is a lack of diverse high quality instruction tuning data in most languages besides English. Thus, we study the impact of the amount of IT data from the new language on the final model performance by varying the Hungarian instruction tuning data mixing rate from 0% to 50%. Figure 3 shows that when no Hungarian instruction tuning data is

Figure 3: Model performance with different IT data mixture. ROUGE-2 score is reported for HU Sum, while accuracy and F1 scores are reported for the rest of the tasks.

included, the model undergoes catastrophic forgetting of Hungarian. However, the model performance on Hungarian improves as more Hungarian data is mixed in, with marginal returns after more than 1% of the data is Hungarian. This indicates that a small amount of IT data from the new language gives most of the model performance on the new language.
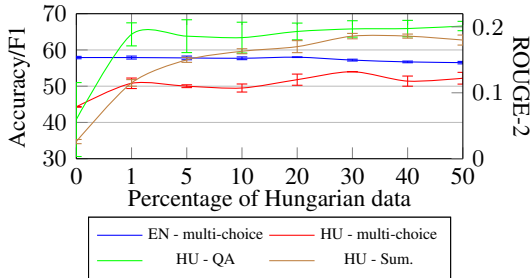
4 Conclusion

In our paper, we study the recipe to efficiently adapt an existing pretrained LLM to a new language, with better tokenizer efficiency and without catastrophic forgetting of its original knowledge. With only 10% of tokens in the tokenizer replaced by the the new ones from the target language, it can drop the fertility by 50% and 70% on Hungarian and Thai respectively, with limited regression on English. This can greatly improve the efficiency of both training and inference on the new language by 2x and 3x. In addition, with mixing training data from both languages in pretraining and IT stages, we show that this can improve the model performance on the new language while retaining the model capability on the original language.

References

[1] B. Workshop, "Bloom: A 176b-parameter open-access multilingual language model," 2023.

[2] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen, Z.-X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, and C. Raffel, "Crosslingual generalization through multitask finetuning," 2023.

[3] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," 2021.

[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2020.

[5] N. Goyal, J. Du, M. Ott, G. Anantharaman, and A. Conneau, "Larger-scale transformers for multilingual masked language modeling," 2021.

[6] O. Shliazhko, A. Fenogenova, M. Tikhonova, V. Mikhailov, A. Kozlova, and T. Shavrina, "mgpt: Few-shot learners go multilingual," 2022.

[7] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O'Horo, J. Wang, L. Zettlemoyer, Z. Kozareva, M. Diab, V. Stoyanov, and X. Li, "Few-shot learning with multilingual language models," 2022.

[8] H. Xu, B. V. Durme, and K. Murray, "Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation," 2021.

[9] Z.-X. Yong, H. Schoelkopf, N. Muennighoff, A. F. Aji, D. I. Adelani, K. Almubarak, M. S. Bari, L. Sutawika, J. Kasai, A. Baruwa, G. I. Winata, S. Biderman, E. Raff, D. Radev, and V. Nikoulina, "Bloom+1: Adding language support to bloom for zero-shot prompting," 2023.

[10] J. Phang, I. Calixto, P. M. Htut, Y. Pruksachatkun, H. Liu, C. Vania, K. Kann, and S. R. Bowman, "English intermediate-task training improves zero-shot cross-lingual transfer too," 2020.

[11] A. Ebrahimi and K. Kann, "How to adapt your pretrained multilingual model to 1600 languages," 2021.

[12] J. Ye, X. Tao, and L. Kong, "Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability," 2023.

[13] J. Armengol-Estapé, O. de Gibert Bonet, and M. Melero, "On the multilingual capabilities of very large-scale english language models," 2021.

[14] K. Ogueji, Y. Zhu, and J. Lin, "Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages," in *Proceedings of the 1st Workshop on Multilingual Representation Learning*, 2021, pp. 116–126.

[15] P. Gage, "A new algorithm for data compression," *The C Users Journal Archive*, vol. 12, pp. 23–38, 1994. [Online]. Available: https://api.semanticscholar.org/CorpusID:59804030

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[17] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[18] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[19] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[20] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, and I. Gurevych, "How good is your tokenizer? on the monolingual performance of multilingual language models," 2021.

[21] F. Stollenwerk, "Training and evaluation of a multilingual tokenizer for gpt-sw3," 2023.

[22] J. Ács. (2019, February) Exploring bert's vocabulary. [Online]. Available: https://juditacs.github.io/2019/02/19/bert-tokenization-stats.html

[23] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[24] S. Cahyawijaya, H. Lovenia, T. Yu, W. Chung, and P. Fung, "Instruct-align: Teaching novel languages with to llms through alignment-based cross-lingual instruction," 2023.

[25] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos, "Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer," 2021.

[26] T. Vu, A. Barua, B. Lester, D. Cer, M. Iyyer, and N. Constant, "Overcoming catastrophic forgetting in zero-shot cross-lingual generation," 2022.

[27] J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder, "MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7654–7673. [Online]. Available: https://aclanthology.org/2020.emnlp-main.617

[28] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1950–1965, 2022.

[29] R. Pires, H. Abonizio, T. S. Almeida, and R. Nogueira, "Sabiá: Portuguese large language models," 2023.

[30] N. Sengupta, S. K. Sahu, B. Jia, S. Katipomu, H. Li, F. Koto, O. M. Afzal, S. Kamboj, O. Pandit, R. Pal *et al.*, "Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models," *arXiv preprint arXiv:2308.16149*, 2023.

[31] D. M. Nemeskey, "Natural language processing methods for language modeling," Ph.D. dissertation, Eötvös Loránd University, 2020. [Online]. Available: https://hlt.bme.hu/media/pdf/nemeskey_thesis.pdf

[32] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman, "Universal Dependencies v1: A multilingual treebank collection," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 1659–1666. [Online]. Available: https://aclanthology.org/L16-1262

[33] V. Vincze, D. Szauter, A. Almási, G. Móra, Z. Alexin, and J. Csirik, "Hungarian dependency treebank," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), May 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/465_Paper.pdf

[34] D. Zeman, M. Popel, M. Straka, J. Hajič, J. Nivre, F. Ginter, J. Luotolahti, S. Pyysalo, S. Petrov, M. Potthast, F. Tyers, E. Badmaeva, M. Gokirmak, A. Nedoluzhko, S. Cinková, J. Hajič jr., J. Hlaváčová, V. Kettnerová, Z. Urešová, J. Kanerva, S. Ojala, A. Missilä, C. D. Manning, S. Schuster, S. Reddy, D. Taji, N. Habash, H. Leung, M.-C. de Marneffe, M. Sanguinetti, M. Simi, H. Kanayama, V. de Paiva, K. Droganova, H. Martínez Alonso, Ç. Çöltekin, U. Sulubacak, H. Uszkoreit, V. Macketanz, A. Burchardt, K. Harris, K. Marheinecke, G. Rehm, T. Kayadelen, M. Attia, A. Elkahky, Z. Yu, E. Pitler, S. Lertpradit, M. Mandl, J. Kirchner, H. F. Alcalde, J. Strnadová, E. Banerjee, R. Manurung, A. Stella, A. Shimada, S. Kwak, G. Mendonça, T. Lando, R. Nitisaroj, and J. Li, "CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies," in *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 1–19. [Online]. Available: https://aclanthology.org/K17-3001

[35] N. Silveira, T. Dozat, M.-C. de Marneffe, S. Bowman, M. Connor, J. Bauer, and C. D. Manning, "A gold standard dependency corpus for English," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014.

[36] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, "The pile: An 800gb dataset of diverse text for language modeling," 2020.

[37] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2023.

[38] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei *et al.*, "The flan collection: Designing data and methods for effective instruction tuning," *arXiv preprint arXiv:2301.13688*, 2023.

[39] H. Nguyen, "The oig dataset," Mar 2023. [Online]. Available: https://laion.ai/blog/oig-dataset/

[40] S. Iyer, X. V. Lin, R. Pasunuru, T. Mihaylov, D. Simig, P. Yu, K. Shuster, T. Wang, Q. Liu, P. S. Koura, X. Li, B. O'Horo, G. Pereyra, J. Wang, C. Dewan, A. Celikyilmaz, L. Zettlemoyer, and V. Stoyanov, "Opt-iml: Scaling language model instruction meta learning through the lens of generalization," 2023.

[41] J. Abadji, P. O. Suarez, L. Romary, and B. Sagot, "Towards a cleaner document-oriented multilingual crawled corpus," 2022.

[42] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 483–498. [Online]. Available: https://aclanthology.org/2021.naacl-main.41

[43] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave, "Ccnet: Extracting high quality monolingual datasets from web crawl data," 2019.

[44] C. Mou, C. Ha, K. Enevoldsen, and P. Liu, "Chenghaomou/text-dedup: Reference snapshot," Sep. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.8364980

[45] A. Broder, "On the resemblance and containment of documents," in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, 1997, pp. 21–29.

[46] H. Nomoto, "Interpersonal meaning annotation for asian language corpora: The case of tufs asian language parallel corpus (talpco)," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:209441604

[47] B. Buschbeck-Wolf and M. Exel, "A parallel evaluation data set of software documentation with document structure annotation," in *Workshop on Asian Translation*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221095497

[48] H. Riza, M. Purwoadi, Gunarso, T. Uliniansyah, A. A. Ti, S. M. Aljunied, L. C. Mai, V. T. Thang, N. P. Thai, V. Chea, R. Sun, S. Sam, S. Seng, K. M. Soe, K. T. Nwet, M. Utiyama, and C. Ding, "Introduction of the asian language treebank," *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pp. 1–6, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:45848332

[49] F. Ladhak, E. Durmus, C. Cardie, and K. McKeown, "Wikilingua: A new benchmark dataset for multilingual abstractive summarization," *ArXiv*, vol. abs/2010.03093, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:222177239

[50] M. Cettolo, C. Girardi, and M. Federico, "WIT3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Annual conference of the European Association for Machine Translation*. Trento, Italy: European Association for Machine Translation, May 28–30 2012, pp. 261–268. [Online]. Available: https://www.aclweb.org/anthology/2012.eamt-1.60

[51] N. team, M. R. Costa-jussà, J. Cross, O. cCelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. L. Spruit, C. Tran, P. Y. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzm'an, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "No language left behind: Scaling human-centered machine translation," *ArXiv*, vol. abs/2207.04672, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250425961

[52] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, "Xnli: Evaluating cross-lingual sentence representations," in *Conference on Empirical Methods in Natural Language Processing*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52271711

[53] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," *CoRR*, vol. abs/1910.11856, 2019.

[54] K. Viriyayudhakorn and C. Polpanumas, "iapp_wiki_qa_squad," Feb. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4539916

[55] A. Suriyawongkul, E. Chuangsuwanich, P. Chormai, and C. Polpanumas, "Pythainlp/wisesight-sentiment: First release," Sep. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3457447

[56] N. Chumpolsathien, "Using knowledge distillation from keyword extraction to improve the informativeness of neural cross-lingual summarization," Master's thesis, Beijing Institute of Technology, 2020.

[57] T. Hasan, A. Bhattacharjee, M. S. Islam, K. Mubasshir, Y.-F. Li, Y.-B. Kang, M. S. Rahman, and R. Shahriyar, "XL-sum: Large-scale multilingual abstractive summarization for 44 languages," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 4693–4703. [Online]. Available: https://aclanthology.org/2021.findings-acl.413

[58] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," https://github.com/tatsu-lab/stanford_alpaca, 2023.

[59] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, and R. Xin. (2023) Free dolly: Introducing the world's first truly open instruction-tuned llm. [Online]. Available: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm

[60] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, "How close is chatgpt to human experts? comparison corpus, evaluation, and detection," *arXiv preprint arxiv:2301.07597*.

[61] A. Kopf, Y. Kilcher, D. von Rutte, S. Anagnostidis, Z. R. Tam, K. Stevens, A. Barhoum, N. M. Duc, O. Stanley, R. Nagyfi, E. Shahul, S. Suri, D. Glushkov, A. Dantuluri, A. Maguire, C. Schuhmann, H. Nguyen, and A. Mattick, "Openassistant conversations - democratizing large language model alignment," *ArXiv*, vol. abs/2304.07327, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258179434

[62] R. Prabhakar and S. Jairath, "Sambanova sn10 rdu:accelerating software 2.0 with dataflow," in *2021 IEEE Hot Chips 33 Symposium (HCS)*, 2021, pp. 1–37.

[63] L. Gao, J. Tow, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, K. McDonell, N. Muennighoff, J. Phang, L. Reynolds, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou, "A framework for few-shot language model evaluation," Sep. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5371628

[64] D. Paperno, G. Kruszewski, A. Lazaridou, Q. N. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, and R. Fernández, "The lambada dataset: Word prediction requiring a broad discourse context," 2016.

[65] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[66] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a suit of armor conduct electricity? a new dataset for open book question answering," 2018.

[67] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," 2019.

[68] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," 2018.

[69] Y. Bisk, R. Zellers, R. L. Bras, J. Gao, and Y. Choi, "Piqa: Reasoning about physical commonsense in natural language," 2019.

[70] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, "Adversarial nli: A new benchmark for natural language understanding," 2020.

[71] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, "Winogrande: An adversarial winograd schema challenge at scale," 2019.

[72] N. Ligeti-Nagy, G. Ferenczi, E. Héja, K. Jelencsik-Mátyus, L. J. Laki, N. Vadász, Z. G. Yang, and T. Vadász, "Hulu: magyar nyelvű benchmark adatbázis kiépítése a neurális nyelvmodellek kiértékelése céljából," in *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*, 2022, pp. 431–446.

[73] O. M. Q. L. I. V. Edoardo M. Ponti, Goran Glava s and A. Korhonen, "XCOPA: A multilingual dataset for causal commonsense reasoning," *arXiv preprint*, 2020. [Online]. Available: https://ducdauge.github.io/files/xcopa.pdf

[74] D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth, "Looking beyond the surface:a challenge set for reading comprehension over multiple sentences," in *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[75] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," *arXiv preprint arXiv:1905.00537*, 2019.

[76] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova, L. Jones, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: a benchmark for question answering research," *Transactions of the Association of Computational Linguistics*, 2019.

[77] S. Zhang, X. Liu, J. Liu, J. Gao, K. Duh, and B. V. Durme, "Record: Bridging the gap between human and machine commonsense reading comprehension," *ArXiv*, vol. abs/1810.12885, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:53116244

[78] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," 2018.

A Tokenizer Details

A.1 Tokenizer Fertility Definition

Fertility is defined as the average number of tokens per word [22]. Words are defined by the Universal Dependencies framework, which gives a "consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages" [32]. To calculate the fertility we tokenize each word from the treebank individually to get the sum total number of tokens, and then divide this by the number of words in the treebank. For Hungarian we used the test set of the "Szeged Dependency Treebank" [33], for Thai we used the test set of the Thai "CoNLL 2017" treebank [34] and for English we used the test set of the "Gold Standard Universal Dependencies Corpus" [35].

A.2 Token Replacement Details

In our work, instead of extending the tokenizer's vocabulary, we replace the least frequent tokens from it with tokens from the new language. This way, we keep the model capacity the same by controlling the vocabulary and embedding table size.

1. Train a tokenizer limited to a vocabulary size k, where k is the number of tokens you want to replace in the original tokenizer
2. Find the number of overlapping tokens o between the new tokenizer of vocab size k, and the original tokenizer
3. Replace the least frequently used $k - o$ tokens from the original tokenizer with the $k - o$ tokens from the new tokenizer. Ensure that all the unchanged tokens from the original tokenizer keep the same vocabulary indices as they had before.
   (a) note that in the GPT2 tokenizer the tokens in the vocabulary and merges file are ordered from most frequent to least frequent, so we replace the last $k - o$ vocabulary indices in a GPT2 Tokenizer.
4. The GPT2 Tokenizer executes the merges rules in the merges.txt file line by line, so to improve the efficiency on the newly added tokens, Add the merges rules from the $k - o$ new tokens to the beginning of the merges.txt file.
   (a) Note that various BPE encoding algorithms are implemented without using the merges rules, so ensure that you examine your tokenizer to see how to improves the tokenizer efficiency.
5. Randomize the embeddings of the replaced tokens in the original model so the new embeddings can be learned.

We tested this tokenizer to ensure that the encoding and decoding of text works properly, and figure 1 shows that it also improves the fertility.

B Base Model

We train our base model with the same tokenizer and architecture as GPT-2 model [16]. The model has 40 layers of transformer blocks with hidden dimension 5120 and 13 billion parameters in total. The vocabulary size is 50260. The base model was pretrained on 300B English tokens from the PILE[36] and C4[37] datasets, filtered for only natural language English text.

C Datasets

C.1 English pretraining data

For the continuous pretraining phase, we often mix English data with either Hungarian or Thai. The English data we used is a 100 gigabyte sample of data from the base model pretraining corpus introduced in section B.

C.2 English instruction tuning data

To construct our English instruction tuning dataset, we sample each constituent task from FlanV2 [38] and OIG [39] equally by raw text size with a fixed dataset size budget. This creates an instruction

tuning dataset that is task diverse and reasonably sized. The benefit of sampling instruction tuning data at the task level is shown in [40], and provides a compute efficient alternative to training on all the data. The dataset is 2.6 gigabytes of raw text and about 1.9 million samples.

C.3 Hungarian pretraining tuning data

The dataset used for Hungarian pretraining is the Hungarian Webcorpus 2.0 [31]. Our dataset is 96 gigabytes and 11,152,900 documents.

C.4 Hungarian instruction tuning tuning data

There is a lack of naturally written Hungarian instruction tuning datasets, so we use google translate to translate our English Instruction tuning corpus C.2 to Hungarian. During translation the prompt and completion are translated separately and only concatenated during training.

C.5 Thai pretraining data

For the Thai pre-training corpus, we combine the Thai subsets of OSCAR [41], MC4 [42], and CCNet [43], which are all derived from Common Crawl. The entire combined corpus was processed with MinHash deduplication [44, 45] with 1-grams and a Jaccard similarity of 0.6, with sentence level n-grams (split by whitespace), and totals 15.32 million documents.

C.6 Thai instruction tuning data

For Thai instruction tuning data, we use a mixture of manually templated Thai datasets, as well as existing IT datasets translated from English.

We take various Thai NLP datasets and create a variety of prompting templates for each task to form an instruction tuning dataset. These consist of translation [46] [47] [48] [49] [50] [51], NLI [52], QA [53] [54], text categorization, sentiment analysis [55], and summarization [56] [57] tasks. This collection of datasets totals 6.26 million instruction tuning examples.

The English-translated IT datasets consist of traditional instruction tuning, multi-turn conversation, and domain-specific QA (i.e. general knowledge, finance, science, mathematics) sourced from collections like FLAN [38], OIG [39], Alpaca [58], Dolly [59], HC3 [60], and OpenAssistant [61], totalling 1.06 million examples.

D   Training Details

D.1   Pretraining Hyperparameters

The training process utilized cross-entropy loss to optimize the CLM objective. All training runs shared the same hyperparameters to ensure a fair comparison and to avoid hyperparameter searches. When comparing two pre-training ablations, the runs were trained to token parity, training on the same number of tokens regardless of available data or tokenizer efficiency

The hyperparameters used were batch size = 512, fixed learning rate = 0.000015 and weight decay = 0.1. All the tokens were packed into the training sequences, if they did not fit in a sequence then they would be placed in the next sequence so no training tokens are lost.[1] An attention mask was applied so that only tokens from the same article attend to each other.

D.2   Instruction Tuning Hyperparameters

All instruction tuning studies share the same hyper-parameters. But ablations comparing runs are not run to step parity, rather they are all trained to 1 epoch to ensure they see all the data.

The hyperparameters used are batch size = 128, fixed learning rate = 0.000015, weight decay = 0.1 and prompt loss weight = 0.0 to ensure that prompts are attended to but not trained on. All the tokens were packed into sequences in a greedy fashion, if they did not fit in a sequence then they would be discarded.[1] An attention mask was applied so that only tokens from the same article attend to each other.

---

[1]https://github.com/sambanova/generative_data_prep

D.3 Hardware Configuration

All training is run on SambaNova's Reconfigurable Data Units (RDU) [62].

E Evaluation

The EAI evaluation harness[63] is used for all benchmarking. The code[2] was adapted for new tasks to evaluate the models on non-English languages. We categorize all the tasks into 4 categories:

Multiple-choice In this category of tasks, we append each candidate option after the prompt and let the model pick answer with the highest probability. We report average accuracy on each tasks.

Table 3: Multiple Choice Evaluation Benchmarks

Language	Evaluation Tasks
English	Lambada [64], HellaSwag [65], Openbookqa [66], Boolq [67], Arc Easy and Challenge [68], PiQA [69], ANLI R1 [70] and Winogrande [71].
Hungarian	HULU evaluation suite [72], which is composed of human translated tasks HuCB, HuSST, HuWNLI, HuCOPA, HuCOLA and HuRTE.
Thai	XCOPA [73] and WiseSight Sentiment Analysis [55] corpus. Translated versions of HellaSwag [65], MultiRC [74], RTE [75].

Open-ended Question Answering In this category of tasks, we let the model to freely generate completions for each question prompt and we report the average F1 score between the model output and the ground truth answer.

Table 4: Open-ended Question Answering Evaluation Benchmarks

Language	Evaluation Tasks
Hungarian	translated versions of BoolQ [67] and Natural Questions [76]
Thai	XQuAD [53] and a translated version of ReCoRD [77]

Summarization In this category of tasks, we let the model freely generate a summary for each prompt and we report average ROUGE-2 score between the model output and ground truth.

Table 5: Summarization Evaluation Benchmarks

Language	Evaluation Tasks
Hungarian	Translated version of XSum [78].
Thai	ThaiSum [56]

Translation In this category of tasks, we let the model to freely generate translated text and we report BLEU score between the model output and the ground truth answer.

Table 6: Translation Evaluation Benchmarks

Language	Evaluation Tasks
Hungarian	wmt 2009 en-hu and wmt 2009 hu-en [3] datasets
Thai	WIT3 Ted Talks Corpus [50]

---

[2]our open source eval suite link