Building Bridges between Cultures to Empower Better Large Language Models

Dmytro Chaplynskyi

Lang-uk initiative, Kyiv, Ukraine chaplinsky.dmitry@gmail.com

Abstract. This paper focuses on the invisible gap between cultures. The language barrier stops the flow of ideas and concepts and results in poor representation of non-English cultures in the modern large language models. Fixing these will help to bring more unique data for the LLMs, make modern tools more accessible around the world, and give better access to cultural diversity. The paper proposes a wide and flexible framework that can help to build bridges over the gaps using different techniques, including dataset collection, better machine translation, and LLMs finetuning. Such a framework might be applied to any low-resource language.

Keywords: Large Language Models · Neural Machine Translation · NLP datasets.

1 Introduction

For the last few years, my co-authors and I have focused our research on low-resource languages. I often ask my friends and colleagues if they want to learn one and what their motivation is to do so. What I love about this little study is the common answer: I'd like to learn it to access another unique culture. And this is where we start.

In recent years, many significant achievements have been made in the field of natural language processing. Not only were state-of-the-art results for various tasks beaten [10] [35], but new tasks were also outlined [20], [9]. Most importantly, NLP-powered products of the new generation were made available to the general public and changed how we read, write, and learn [16]. That created a whole new generation of products unseen before.

That was made possible with the help of the Large Language Models or LLMs. While LLMs might come in different shapes and colors, in general, they are doing the same job: compress the vast amounts of (predominantly) textual information into an efficient inner representation and allow to query this bank of concepts, facts, and ideas in a free form and pack the answer in a requested format. While it is far from Artificial General Intelligence (AGI), it is still a product of capacities that a) wasn't seen before and b) can be improved even further.

In this paper, we will focus on the ways and methods to bring such tools to other languages and to bring other languages to such tools.

2 The current state of the research

Ideas brought to the public by the release of the seminal papers [38], [10], [5] on LLMs influenced a lot of new research, including those that touch on multilingualism.

One focus was on training the model on multilingual datasets to achieve suboptimal performance in the included languages. Aya 101 model is trained on 101 languages [42], Gemma models employing a massive tokenizer which should cover more languages of the non-Latin script [34], Unbabel Tower offers Llama-based models which are pre-trained on 10 more languages [1].

Some authors explored the flaws of multilinguality (so-called Curse of Multilinguality) in the existing architectures, such as [6], [3], [27]. A "curse" here means that the more languages are used to train the model the worse the cross-lingual language understanding is. Such research is important to know the issues with the existing models and to help alleviate the curse.

Another area of the research was to try to train the models of the existing architecture on a limited number of languages (often from scratch) to improve the model's performance for the target languages, for example, EuroLLM, trained on the most popular languages of the European Union [21] or Jais, Arabic-centric model [31]. While this is probably the most promising direction, it requires an enormous amount of computational resources and is not feasible for the under-represented languages.

One interesting direction is the bag of tricks to jump-start model training on the new data by carefully altering the tokenization and embedding parts of the existing model [22], [23] before further pre-training. Such an approach can speed up training of the new models a lot, by piggybacking on the existing models.

It is worth mentioning the research devoted to the creation of multilingual and parallel corpora which are extremely useful in training NMT models and multilingual LLMs such as CCMatrix [30], [13], and CCAligned [12]. Another important area here is the methods that allow mining the parallel data [2] [14], [33] and new metrics to measure the translation quality using LLMs [28].

3 Motivation

As per the Chinchilla scaling law [15], the performance of the LLM can be improved by:

1. Bigger and better models
2. Trained for longer
3. On a bigger number of texts

While the first two elements can be solved with a bigger budget and Moore's law [24], the third has a caveat. The general number of texts available for training is limited. Not all texts are available for various reasons, like copyrights, language, media, or other limitations. But also, the number of texts is increasing almost linearly and is proportional to the size of the literate human population.

Moreover, the texts generally represent ideas, concepts, or facts. If one takes a random text from the internet, the chance that this text conveys new knowledge might be very low. On the other hand, powerful ideas and concepts create numerous texts. Think about seminal papers like The Bible or the work of Socrates. Therefore, one might speculate that ideas and concepts empower LLMs rather than raw texts. Also, the number of texts is hard-capped, and unless new generations of LLMs learn how to create new knowledge, NLP researchers are scarce in the available texts, ideas, and concepts.

Now, how can more texts be made available for LLMs? If we look at the composition of the training data for modern LLMs [37], [34], [25], English texts dominate the mix [41]. While English is not the most popular language on the planet [40], the number of datasets available in English is clearly bigger than any other language. There might be three reasons for that:

1. The NLP community is still very English-centric
2. English speakers produce more texts on average than speakers of other languages
3. English texts and datasets are more accessible in bulk

This effectively means that not only is LLM's performance worse in other languages, but ideas and concepts from other cultures are represented less in the compressed storage of knowledge provided by large language models. While the number of texts available in other languages might be lower, the number of unique ideas and concepts that can be brought to the new generation of LLMs might be substantial.

However, it is also important to pass the ideas the other way around, bringing the world's knowledge to the low-resource language. Providing a massive amount of general knowledge to speakers of non-English languages can help people learn and create in their mother tongue and fuse their local cultural aspects with the world's best ideas.

4 Methodology

There are a few strategies to shorten the gap between languages and cultures. If we dissect the general problem into smaller problems, we can see that there are two components in there:

- Make models "speak" and "understand" more languages
- Bring concepts from one language into another

As both concepts are heavily intertwined, we might focus on the particular tasks and then see how these can contribute to both components.

4.1 Paving the path to collect more datasets in a particular language

If we look at the number of datasets available through the popular platform [17], we might quickly conclude that english dominates the landscape. Therefore, it

is important to outline and share the methodology to create more datasets in other languages and create the datasets per se. Such datasets might be:

1. The massive collection of raw texts in a target language is to improve the representation of that language in the next generations of LLMs.
2. The smaller datasets, manually annotated for the various tasks, like POS and NER tagging, question answering, sentiment analysis, hate speech detection, et cetera. These might be used as is [25] or turned into the instruction format for the casual language models [19]
3. The dictionaries and the Wordnets which require a lot of curating and manual supervision.
4. Parallel corpora between languages

Not only is collecting and making the data available in a target language important, but it is also important to create and share a solid methodology that will allow the creation of similar datasets in other languages [7], [8], [32].

Some datasets might be created by translating existing datasets in other languages to a target language. This makes more data available for researchers in a particular language domain and creates parallel corpora suitable for tasks like machine translation.

4.2 Making more data accessible in a text format

While the Internet provides the researcher with vast text, some information is still inaccessible in a digitized format. This might include printed textbooks, recorded speech, music, and graphical information. Multi-modal models can help close this gap. Such models, provided with enough data and proper architecture, can access information from various media and extract ideas and concepts. Many books available in the libraries are not accessible for direct processing, and no one can explain the genius of a painter using only words.

4.3 Creating machine translation models

Another fruitful idea is to create high-quality translation models [26]. This will help increase the number of texts available in low-resource languages and the proportion of text in such languages in the general mix of the training data for general models. Not only will the speakers of the low-resource language have better access to the information and modern LLMs in their own language, but the general population can also benefit from the variety of ideas from other cultures that were added through translations.

4.4 Creating Ukrainian LLMs

Training large language models on Ukrainian texts [18] or fine-tuning existing models [4] will help generate more Ukrainian datasets, which can be used for various downstream tasks after proofreading, similar to the Self-instruct [39] and Alpaca paper [29].

4.5 Tapping into the interlingual representation of concepts in LLMS

Recent research [11], [36] has shown that modern LLMs might maintain an internal representation of language-agnostic concepts and facts. While big LLMs are still a black box for the researcher, carefully conducted experiments might allow us to understand better how the unique concepts are represented inside the LLM's weights and can be leveraged to unlock the truly multilingual models. While this area is new and yet uncharted, understanding the representation's inner mechanics might help introduce new concepts to the Large Language Models without retraining the whole model and also help to make them accessible in more languages, including low-resource ones.

The aforementioned activities might be structured as follows:

1. Preparation of the methodology to mine and filter multilingual data, especially parallel corpora to obtain the training dataset of a high quality.
2. Training of the NMT models of different architectures to set the new State Of The Art results for English-Ukrainian pair.
3. Expansion such approach to other language pairs, for example, German-Ukrainian or French-Ukrainian.
4. Preparation of the data printer, to use obtained models to translate the important texts into Ukrainian in an automatic fashion.
5. Collaboration on the creation of Ukrainian LLMs.

5 Conclusion

While it is hard to catch up with the current progress on LLMs, we, as researchers, can still make an invaluable contribution, building the bridges that can bring data, concepts, ideas, and facts from one culture into another. It doesn't require a massive amount of the compute which is still scarce in Ukraine, but thorough work on data collection and processing, creative ideas on the models training and fine-tuning, and a lot of passion to make Ukraine a part of this movement again and expand the frontier of large language models even further. In addition, the methodology created in the research process can and should be applied to other low-resource languages to preserve the knowledge and spirit of other cultures and connect them to the global bank of ideas and concepts.

References

1. Alves, D.M., Pombal, J., Guerreiro, N.M., Martins, P.H., Alves, J., Farajian, A., Peters, B., Rei, R., Fernandes, P., Agrawal, S., Colombo, P., de Souza, J.G.C., Martins, A.F.T.: Tower: An open multilingual large language model for translation-related tasks (2024)
2. Artetxe, M., Schwenk, H.: Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. Transactions of the Association for Computational Linguistics **7**, 597–610 (09 2019). https://doi.org/10.1162/tacl_a_00288, https://doi.org/10.1162/tacl_a_00288

3. Blevins, T., Limisiewicz, T., Gururangan, S., Li, M., Gonen, H., Smith, N.A., Zettlemoyer, L.: Breaking the curse of multilinguality with cross-lingual expert language models (2024), https://arxiv.org/abs/2401.10440
4. Boros, T., Chivereanu, R., Dumitrescu, S., Purcaru, O.: Fine-tuning and retrieval augmented generation for question answering using affordable large language models. In: Romanyshyn, M., Romanyshyn, N., Hlybovets, A., Ignatenko, O. (eds.) Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024. pp. 75–82. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.unlp-1.10
5. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020), https://arxiv.org/abs/2005.14165
6. Chang, T.A., Arnett, C., Tu, Z., Bergen, B.K.: When is multilinguality a curse? language modeling for 250 high- and low-resource languages (2023), https://arxiv.org/abs/2311.09205
7. Chaplynskyi, D.: Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. In: Romanyshyn, M. (ed.) Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP). pp. 1–10. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.unlp-1.1, https://aclanthology.org/2023.unlp-1.1
8. Chaplynskyi, D., Romanyshyn, M.: Introducing NER-UK 2.0: A rich corpus of named entities for Ukrainian. In: Romanyshyn, M., Romanyshyn, N., Hlybovets, A., Ignatenko, O. (eds.) Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024. pp. 23–29. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.unlp-1.4
9. Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H.P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F.P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W.H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A.N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., Zaremba, W.: Evaluating large language models trained on code (2021), https://arxiv.org/abs/2107.03374
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019), https://arxiv.org/abs/1810.04805
11. Dumas, C., Veselovsky, V., Monea, G., West, R., Wendler, C.: How do llamas process multilingual text? a latent exploration through activation patching. In: ICML 2024 Workshop on Mechanistic Interpretability (2024), https://openreview.net/forum?id=0ku2hIm4BS
12. El-Kishky, A., Chaudhary, V., Guzmán, F., Koehn, P.: CCAligned: A massive collection of cross-lingual web-document pairs. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020) (November 2020)

13. Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., Joulin, A.: Beyond english-centric multilingual machine translation (2020), https://arxiv.org/abs/2010.11125
14. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic bert sentence embedding (2022), https://arxiv.org/abs/2007.01852
15. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L.A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J.W., Vinyals, O., Sifre, L.: Training compute-optimal large language models (2022), https://arxiv.org/abs/2203.15556
16. Huber, S.E., Kiili, K., Nebel, S., Ryan, R.M., Sailer, M., Ninaus, M.: Leveraging the potential of large language models in education through playful and game-based learning. Educational Psychology Review **36**(1), 25 (2024)
17. Huggingface datasets by language, https://huggingface.co/datasets?modality=modality:textsort=trending
18. Kiulian, A., Polishko, A., Khandoga, M., Kostiuk, Y., Gabrielli, G., Łukasz Gagała, Zaraket, F., Obaida, Q.A., Garud, H., Mak, W.W.Y., Chaplynskyi, D., Amor, S.B., Peradze, G.: From english-centric to effective bilingual: Llms with custom tokenizers for underrepresented languages (2024), https://arxiv.org/abs/2410.18836
19. Kyrylov, V., Chaplynskyi, D.: GPT-2 metadata pretraining towards instruction finetuning for Ukrainian. In: Romanyshyn, M. (ed.) Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP). pp. 32–39. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.unlp-1.4, https://aclanthology.org/2023.unlp-1.4
20. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks (2021), https://arxiv.org/abs/2005.11401
21. Martins, P.H., Fernandes, P., Alves, J., Guerreiro, N.M., Rei, R., Alves, D.M., Pombal, J., Farajian, A., Faysse, M., Klimaszewski, M., Colombo, P., Haddow, B., de Souza, J.G.C., Birch, A., Martins, A.F.T.: Eurollm: Multilingual language models for europe (2024), https://arxiv.org/abs/2409.16235
22. Minixhofer, B., Paischer, F., Rekabsaz, N.: Wechsel: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. p. 3992–4006. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.naacl-main.293, http://dx.doi.org/10.18653/v1/2022.naacl-main.293
23. Minixhofer, B., Ponti, E.M., Vulić, I.: Zero-shot tokenizer transfer (2024), https://arxiv.org/abs/2405.07883
24. Mollick, E.: Establishing moore's law. Annals of the History of Computing, IEEE **28**, 62 – 75 (08 2006). https://doi.org/10.1109/MAHC.2006.45
25. Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T.L., Bari, M.S., Shen, S., Yong, Z.X., Schoelkopf, H., Tang, X., Radev, D., Aji, A.F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., Raffel, C.: Crosslingual generalization through multitask finetuning (2023), https://arxiv.org/abs/2211.01786

26. Paniv, Y., Chaplynskyi, D., Trynus, N., Kyrylov, V.: Setting up the data printer with improved english to ukrainian machine translation (2024), https://arxiv.org/abs/2404.15196
27. Pfeiffer, J., Goyal, N., Lin, X., Li, X., Cross, J., Riedel, S., Artetxe, M.: Lifting the curse of multilinguality by pre-training modular transformers. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3479–3495. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.255, https://aclanthology.org/2022.naacl-main.255
28. Rei, R., Guerreiro, N.M., Pombal, J., van Stigt, D., Treviso, M., Coheur, L., de Souza, J.G.C., Martins, A.F.T.: Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task (2023), https://arxiv.org/abs/2309.11925
29. Rohan Taori, Ishaan Gulrajani, T.Z.Y.D.X.L.C.G., Percy Liang, T.B.H.: Alpaca: A strong, replicable instruction-following model (2024), https://crfm.stanford.edu/2023/03/13/alpaca.html
30. Schwenk, H., Wenzek, G., Edunov, S., Grave, E., Joulin, A.: Ccmatrix: Mining billions of high-quality parallel sentences on the web (2020), https://arxiv.org/abs/1911.04944
31. Sengupta, N., Sahu, S.K., Jia, B., Katipomu, S., Li, H., Koto, F., Afzal, O.M., Kamboj, S., Pandit, O., Pal, R., Pradhan, L., Mujahid, Z.M., Baali, M., Aji, A.F., Liu, Z., Hock, A., Feldman, A., Lee, J., Jackson, A., Nakov, P., Baldwin, T., Xing, E.: Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models (2023)
32. Shvedova, M., Lukashevskyi, A.: Creating parallel corpora for Ukrainian: A German-Ukrainian parallel corpus (ParaRook||DE-UK). In: Romanyshyn, M., Romanyshyn, N., Hlybovets, A., Ignatenko, O. (eds.) Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024. pp. 14–22. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.unlp-1.3
33. Tan, W., Heffernan, K., Schwenk, H., Koehn, P.: Multilingual representation distillation with contrastive learning (2023), https://arxiv.org/abs/2210.05033
34. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., Tafti, P., Hussenot, L., Sessa, P.G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., Tacchetti, A., Bulanova, A., Paterson, A., Tsai, B., Shahriari, B., Lan, C.L., Choquette-Choo, C.A., Crepy, C., Cer, D., Ippolito, D., Reid, D., Buchatskaya, E., Ni, E., Noland, E., Yan, G., Tucker, G., Muraru, G.C., Rozhdestvenskiy, G., Michalewski, H., Tenney, I., Grishchenko, I., Austin, J., Keeling, J., Labanowski, J., Lespiau, J.B., Stanway, J., Brennan, J., Chen, J., Ferret, J., Chiu, J., Mao-Jones, J., Lee, K., Yu, K., Millican, K., Sjoesund, L.L., Lee, L., Dixon, L., Reid, M., Mikuła, M., Wirth, M., Sharman, M., Chinaev, N., Thain, N., Bachem, O., Chang, O., Wahltinez, O., Bailey, P., Michel, P., Yotov, P., Chaabouni, R., Comanescu, R., Jana, R., Anil, R., McIlroy, R., Liu, R., Mullins, R., Smith, S.L., Borgeaud, S., Girgin, S., Douglas, S., Pandya, S., Shakeri, S., De, S., Klimenko, T., Hennigan, T., Feinberg, V., Stokowiec, W., hui Chen, Y., Ahmed, Z., Gong, Z., Warkentin, T., Peran, L., Giang, M., Farabet, C., Vinyals, O., Dean, J., Kavukcuoglu, K., Hassabis, D., Ghahramani, Z., Eck, D., Barral, J., Pereira, F., Collins, E., Joulin,

A., Fiedel, N., Senter, E., Andreev, A., Kenealy, K.: Gemma: Open models based on gemini research and technology (2024), https://arxiv.org/abs/2403.08295
35. Team, N., Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G.M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K.R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N.F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., Wang, J.: No language left behind: Scaling human-centered machine translation (2022), https://arxiv.org/abs/2207.04672
36. Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N.L., McDougall, C., MacDiarmid, M., Freeman, C.D., Sumers, T.R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., Henighan, T.: Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread (2024), https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html
37. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023), https://arxiv.org/abs/2302.13971
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023), https://arxiv.org/abs/1706.03762
39. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H.: Self-instruct: Aligning language models with self-generated instructions (2023), https://arxiv.org/abs/2212.10560
40. List of languages by number of native speakers (2024), https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers
41. Workshop, B., :, Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., Tow, J., Rush, A.M., Biderman, S., Webson, A., Ammanamanchi, P.S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A.V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P.O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A.F., Alfassy, A., Rogers, A., Nitzav, A.K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D.I., Radev, D., Ponferrada, E.G., Levkovizh, E., Kim, E., Natan, E.B., Toni, F.D., Dupont, G., Kruszewski, G., Pistilli, G., Elsahar, H., Benyamina, H., Tran, H., Yu, I., Abdulmumin, I., Johnson, I., Gonzalez-Dios, I., de la Rosa, J., Chim, J., Dodge, J., Zhu, J., Chang, J., Frohberg, J., Tobing, J., Bhattacharjee, J., Almubarak, K., Chen, K., Lo, K., Werra, L.V., Weber, L., Phan, L., allal, L.B., Tanguy, L., Dey, M., Muñoz, M.R., Masoud, M., Grandury, M., Šaško, M., Huang, M., Coavoux, M., Singh, M., Jiang, M.T.J., Vu, M.C., Jauhar, M.A., Ghaleb, M., Subramani, N., Kassner, N., Khamis, N., Nguyen, O., Espejel, O., de Gibert, O., Villegas, P., Henderson, P., Colombo, P., Amuok, P., Lhoest, Q., Harliman, R., Bommasani, R., López, R.L., Ribeiro, R., Osei, S., Pyysalo, S., Nagel, S., Bose, S., Muhammad, S.H., Sharma, S., Longpre, S., Nikpoor, S., Silberberg, S., Pai, S., Zink, S., Torrent, T.T., Schick, T., Thrush, T., Danchev, V., Nikoulina, V., Laippala, V., Lepercq, V., Prabhu, V., Alyafeai, Z., Talat, Z., Raja, A., Heinzerling, B., Si, C., Taşar, D.E., Salesky, E., Mielke, S.J., Lee, W.Y., Sharma, A., Santilli, A., Chaffin, A., Stiegler, A., Datta, D., Szczechla,

E., Chhablani, G., Wang, H., Pandey, H., Strobelt, H., Fries, J.A., Rozen, J., Gao, L., Sutawika, L., Bari, M.S., Al-shaibani, M.S., Manica, M., Nayak, N., Teehan, R., Albanie, S., Shen, S., Ben-David, S., Bach, S.H., Kim, T., Bers, T., Fevry, T., Neeraj, T., Thakker, U., Raunak, V., Tang, X., Yong, Z.X., Sun, Z., Brody, S., Uri, Y., Tojarieh, H., Roberts, A., Chung, H.W., Tae, J., Phang, J., Press, O., Li, C., Narayanan, D., Bourfoune, H., Casper, J., Rasley, J., Ryabinin, M., Mishra, M., Zhang, M., Shoeybi, M., Peyrounette, M., Patry, N., Tazi, N., Sanseviero, O., von Platen, P., Cornette, P., Lavallée, P.F., Lacroix, R., Rajbhandari, S., Gandhi, S., Smith, S., Requena, S., Patil, S., Dettmers, T., Baruwa, A., Singh, A., Cheveleva, A., Ligozat, A.L., Subramonian, A., Névéol, A., Lovering, C., Garrette, D., Tunuguntla, D., Reiter, E., Taktasheva, E., Voloshina, E., Bogdanov, E., Winata, G.I., Schoelkopf, H., Kalo, J.C., Novikova, J., Forde, J.Z., Clive, J., Kasai, J., Kawamura, K., Hazan, L., Carpuat, M., Clinciu, M., Kim, N., Cheng, N., Serikov, O., Antverg, O., van der Wal, O., Zhang, R., Zhang, R., Gehrmann, S., Mirkin, S., Pais, S., Shavrina, T., Scialom, T., Yun, T., Limisiewicz, T., Rieser, V., Protasov, V., Mikhailov, V., Pruksachatkun, Y., Belinkov, Y., Bamberger, Z., Kasner, Z., Rueda, A., Pestana, A., Feizpour, A., Khan, A., Faranak, A., Santos, A., Hevia, A., Unldreaj, A., Aghagol, A., Abdollahi, A., Tammour, A., HajiHosseini, A., Behroozi, B., Ajibade, B., Saxena, B., Ferrandis, C.M., McDuff, D., Contractor, D., Lansky, D., David, D., Kiela, D., Nguyen, D.A., Tan, E., Baylor, E., Ozoani, E., Mirza, F., Ononiwu, F., Rezanejad, H., Jones, H., Bhattacharya, I., Solaiman, I., Sedenko, I., Nejadgholi, I., Passmore, J., Seltzer, J., Sanz, J.B., Dutra, L., Samagaio, M., Elbadri, M., Mieskes, M., Gerchick, M., Akinlolu, M., McKenna, M., Qiu, M., Ghauri, M., Burynok, M., Abrar, N., Rajani, N., Elkott, N., Fahmy, N., Samuel, O., An, R., Kromann, R., Hao, R., Alizadeh, S., Shubber, S., Wang, S., Roy, S., Viguier, S., Le, T., Oyebade, T., Le, T., Yang, Y., Nguyen, Z., Kashyap, A.R., Palasciano, A., Callahan, A., Shukla, A., Miranda-Escalada, A., Singh, A., Beilharz, B., Wang, B., Brito, C., Zhou, C., Jain, C., Xu, C., Fourrier, C., Periñán, D.L., Molano, D., Yu, D., Manjavacas, E., Barth, F., Fuhrimann, F., Altay, G., Bayrak, G., Burns, G., Vrabec, H.U., Bello, I., Dash, I., Kang, J., Giorgi, J., Golde, J., Posada, J.D., Sivaraman, K.R., Bulchandani, L., Liu, L., Shinzato, L., de Bykhovetz, M.H., Takeuchi, M., Pàmies, M., Castillo, M.A., Nezhurina, M., Sänger, M., Samwald, M., Cullan, M., Weinberg, M., Wolf, M.D., Mihaljcic, M., Liu, M., Freidank, M., Kang, M., Seelam, N., Dahlberg, N., Broad, N.M., Muellner, N., Fung, P., Haller, P., Chandrasekhar, R., Eisenberg, R., Martin, R., Canalli, R., Su, R., Su, R., Cahyawijaya, S., Garda, S., Deshmukh, S.S., Mishra, S., Kiblawi, S., Ott, S., Sang-aroonsiri, S., Kumar, S., Schweter, S., Bharati, S., Laud, T., Gigant, T., Kainuma, T., Kusa, W., Labrak, Y., Bajaj, Y.S., Venkatraman, Y., Xu, Y., Xu, Y., Xu, Y., Tan, Z., Xie, Z., Ye, Z., Bras, M., Belkada, Y., Wolf, T.: Bloom: A 176b-parameter open-access multilingual language model (2023), https://arxiv.org/abs/2211.05100

42. Üstün, A., Aryabumi, V., Yong, Z.X., Ko, W.Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.L., Kayid, A., Vargus, F., Blunsom, P., Longpre, S., Muennighoff, N., Fadaee, M., Kreutzer, J., Hooker, S.: Aya model: An instruction finetuned open-access multilingual language model (2024), https://arxiv.org/abs/2402.07827