CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	UCCIX: Irish-eXcellence Large Language Model
Authors	Tran, Khanh-Tung;O'Sullivan, Barry;Nguyen, Hoang D.
Publication date	2024
Original Citation	Tran, K.-T., O'Sullivan, B. and Nguyen, H.D. (2024) 'UCCIX: Irish-excellence Large Language Model'. arXiv. https://doi.org/10.48550/ARXIV.2405.13010
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://arxiv.org/abs/2405.13010 - 10.48550/ARXIV.2405.13010.
Rights	© 2024, the Authors. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-11-07 12:22:17
Item downloaded from	https://hdl.handle.net/10468/17310



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

arXiv:2405.13010v1 [cs.CL] 13 May 2024

UCCIX: Irish-eXcellence Large Language Model

Khanh-Tung Tran[a], **Barry O'Sullivan**[a;*] **and Hoang D. Nguyen**[a]

[a]University College Cork, Cork, Ireland

Abstract. The development of Large Language Models (LLMs) has predominantly focused on high-resource languages, leaving extremely low-resource languages like Irish with limited representation. This work presents UCCIX, a pioneering effort on the development of an open-source Irish-based LLM. We propose a novel framework for continued pre-training of LLMs specifically adapted for extremely low-resource languages, requiring only a fraction of the textual data typically needed for training LLMs according to scaling laws. Our model, based on Llama 2-13B [23], outperforms much larger models on Irish language tasks with up to 12% performance improvement, showcasing the effectiveness and efficiency of our approach. We also contribute comprehensive Irish benchmarking datasets, including IrishQA, a question-answering dataset, and Irish version of MT-bench [28]. These datasets enable rigorous evaluation and facilitate future research in Irish LLM systems. Our work aims to preserve and promote the Irish language, knowledge, and culture of Ireland in the digital era while providing a framework for adapting LLMs to other indigenous languages.

1 Introduction and Contribution

Large Language Models (LLMs) have demonstrated exceptional performance in a wide range of natural language processing (NLP) tasks, capable of understanding and generating human-like text. Prominent models, whether closed-source like ChatGPT, Google Gemini, or open-source variants such as Llama 2 [23] and BLOOM [25], predominantly cater to popular languages like English, however, are limited in supporting for other indigenous languages. Moreover, the training of such LLMs often demands vast amounts of text data, numbering in the hundreds of billions or even trillions of tokens, adhering to recent scaling laws [11, 19]. This poses a significant challenge for most languages, particularly those with extremely limited resources. For instance, consider the case of Irish, even though it is the first official language of the Republic of Ireland, it is extremely underpresented, in both daily usages and on the Internet space. Classified as Definitely Endangered by UNESCO [24], Irish faces an uphill battle for preservation and digitization.

In this work, we attempt to develop open-source Irish-based LLMs with up to tens of billions of parameters for a language with only hundreds of million tokens of text [5], only a fraction of the data typically required for training. This challenge presents an extreme scenario that has been overlooked in relevant works. For instance, FinGPT [15] utilized 300B Finnish tokens in order to be able to continued pre-train the base BLOOM model, and [26] employed up to 20B tokens for five non-English languages. Similarly, Llemma [3] leveraged up to 200B tokens in mathematical topics for adapting CodeLlama. Approaches suggesting synthetic data generation, such as [10, 16], are resource-intensive, and the quality of off-the-shelf LLMs in generating Irish text remains uncertain. BLOOM+1 [27] explored language adaptation at smaller model scales, resulting in modest performance gains. gaBERT [5], a mono-Irish language model, operates at a much smaller scale with less than 300 million parameters. To address these limitations, we present *UCCIX*, the first Irish-based LLMs, trained on our curated set of Irish training datasets, covering broad domains and topics. We carry out experiments with the language adaptation strategy of continued pre-training, including tailored data mixture and scheduler during continued pre-training. An example of the interactions with our LLMs is provided in Figure 1. We hope UCCIX can be used as a tool to preserve the language, knowledge, and culture of Irish in the digital era. Moreover, while Irish is our case study in this research, we believe our framework can be adapted as a blueprint for language adaptation of large models to other low-resource languages.

Developing an LLM for an extremely low-resource language presents challenges not only in terms of training resources but also in the evaluation process. To our best knowledge, there is a scarcity of benchmarking datasets specifically tailored for evaluating LLMs in the context of low-resource languages [6]. This limitation poses a significant challenge in effectively assessing the performance and capabilities of generative language models designed for languages with limited textual data, such as Irish. Existing benchmark datasets that



Figure 1. Example of responses generated by our model, UCCIX, and other baselines[1]. We demonstrate the English translation version on the right.

* Corresponding Author. Email: b.osullivan@cs.ucc.ie

[1] Demo webpage: https://aine.chat/

support Irish are often focused on traditional NLP tasks [14, 5, 2] and may not fully capture the capabilities of modern LLMs, such as language understanding and reasoning. This poses a need for tailoring a new set of benchmark data for Irish-based LLMs. To address this gap, we create new datasets that enable comprehensive benchmarking across various scenarios and settings, such as open-book question answering, and multi-turn conversational interactions.

Our contributions:

- An innovative framework for continued pre-training of large language models for extremely low-resource languages.
- UCCIX, an Irish-based LLM, based on Llama 2-13B. UCCIX outperforms much larger models on Irish tasks with up to 12% performance improvement, demonstrating the effectiveness and efficiency of our proposed framework.
- Irish benchmarking datasets, including IrishQA, our curated question-answering dataset on topics surrounding Ireland and its cultural nuances, available in both English and Irish; and MT-bench, translated to Irish and verified by native Irish speakers.

2 UCCIX: Irish-eXcellence Large Language Model

Data Collection and Pre-processing. For pre-training, we meticulously curate all available open-source Irish data. Our primary sources include CulturaX [20] and Glot500 [12], both of which provide valuable content from multilingual websites, including a subset dedicated to Irish. Additionally, we incorporate data from the Irish segment of the ga-en bitext pair of ParaCrawl v7 [4], text sourced from Irish Wikipedia, and data from Corpora Irish [9]. To ensure data integrity, we conduct thorough cleaning and filtering, following the heuristics introduced in [15]. This is then followed by the removal of duplicates across datasets using n-gram ($n = 5$ in this work) matching, as there are potential overlaps between data sources. This preprocessing process ensures a high quality corpora of data, particularly important in resource-constrained environments [22]. The size and proportion of each dataset in the final data mixture are summarized in Table 1. We believe our mixture is relatively larger than all previous works for Irish NLP [17, 5].

Table 1. Pre-train data statistics.

Dataset	Chars (before pre-processing)	Chars (after pre-processing)	Ratio
CulturaX-ga [20]	1.3B	1.2B	65.0%
Glot500-ga [12]	1.3B	483.7M	27.1%
Gaparacrawl [4]	395.2M	106.3M	6.0%
Gawiki[2]	41.9M	23.4M	1.3%
Corpora Irish [9]	37.1M	11.1M	0.6%
Total	3.1B	1.8B	100%

Continued Pre-training. Given the constraints of extremely low-resource languages, where the amount of mono-Irish data is insufficient for the pre-training of LLMs from scratch, we adopt a strategy of language adaptation for continued pre-training. Starting from a pre-trained LLM, we iteratively enhance its proficiency in understanding additional languages by exposing it to target language data. We follow the hypothesis from recent works [27, 26] that we can transfer world knowledge and English proficiency to Irish through this process. Moreover, we employ a tailored data scheduler to feed parallel English-Irish data first, facilitating the establishment of linguistic connections between the two languages. Parallel pairs of English-Irish sentences from ELRC [7] are leveraged. The parallel data is approximately 1% of the size of the mono-Irish corpus introduced in Table 1. Then, the larger set of mono-Irish is fed to the model to further refine the proficiency of the language and coverage of Irish cultural nuances.

Model Architecture. The selection of an appropriate base model is vital for maximizing support for the Irish language. We employ perplexity as the main metrics to compare between base models. We select a subset of the pre-training corpora with $\approx 2M$ characters, named PT, to compute the metric on. We identify Llama 2-13B as the optimal base model. With a lower perplexity score (8.94) compared to alternative models such as Mistral-7B [13](11.68) or BLOOM-7B [25](63.75), Llama 2-13B serves as a strong foundation. Additionally, we only consider models of adequate size (less than 20B parameters), as we fear the problem of underfitting larger models.

Tokenizer's Vocabulary Expansion. We extend the original vocabulary of the chosen LLM, as it contains mostly tokens for languages such as English. From the original size of 32k tokens, the vocabulary is expanded to include 10k Irish tokens. This not only allows for efficiency improvement in generating Irish text but also maintains the same generation speed for English. Moreover, the expanded vocabulary provides a more meaningful representation of Irish text and enables bilingual capabilities in the final LLM.

Supervised Instruction Fine-tuning. After pre-training phase, to align the model more closely with human preferences, we apply Supervised Instruction Fine-tuning [18, 21]. This process enhances the model's ability to follow human instructions effectively. We denote the pre-trained model as UCCIX, and the instruction fine-tuned version as UCCIX-Instruct.

Evaluation Data. In evaluating pre-trained LLMs, our focus lies on assessing their language understanding capabilities. We utilize various existing NLP benchmarking datasets tailored for Irish, including the Cloze Test [5], the Irish subset of SIB-200 [2], and gaHealth [14] However, we observe a lack of natural language generation benchmarking tasks, such as question-answering, and Irish-specific cultural content in these datasets. To address this gap, we introduce *IrishQA*, a question-answering dataset on topics surrounding Ireland and its cultural nuances. The dataset has a total of 60 question-answer pairs, each pair is accompanied by a paragraph, containing the contextual information necessary to answer the question that the LLMs will be prompted with alongside the question. The dataset is available in both English and Irish, created by native Irish speakers. Furthermore, for the evaluation of instruction-tuned LLMs, we follow a recent approach [28] that highlights the potential of using really large language models, such as GPT-4, as an automated judge. We translate and adapt the MT-bench dataset to Irish using Google Cloud Translation API [1], carefully verifying and correcting the translation through native Irish speakers.

3 Benchmarking Results

Irish Performance. The evaluation results in Table 2 reveals the robust performances of UCCIX, surparssing larger models across datasets, with the largest gap of 12% accuracy on Cloze Test, compared to the second best model. These margins remain wider when compared against baselines of similar sizes (Llama 2-13B and BLOOM-7B1). Notably, baseline models seem to have an understanding of the Irish language, as evidenced by their ability to translate from Irish to English (gaHealth-ga2en task), but they failed to generate coherent Irish text, resulting in superior performance on the gaHealth-ga2en task compared to gaHealth-en2ga. Nevertheless, our model enhances the performance across both transla-

[2] We use the dumps from https://dumps.wikimedia.org/gawiki/20240201/

Table 2. Evaluation results of pre-trained models on curated set of Irish benchmarking datasets. UCCIX$_{npd}$ is our approach but without training the base model on parallel data first, and UCCIX$_{nte}$ is the ablation study without tokenizer's vocabulary expansion.

Model	Acc. on Cloze Test (0-shot)	Acc. on SIB-200 (Irish subset) (10-shot)	Exact-match on IrishQA (ga) (5-shot)	BLEU-4 on gaHealth - en2ga (5-shot)	BLEU-4 on gaHealth - ga2en (5-shot)	Tokenizer speed on PT (chars/token)	Fertility rate on PT (tokens/word)
gpt-3.5-turbo	N/A	N/A	0.2222	0.1864	0.4257	2.45	N/A
Llama 2-70B	0.63	0.7059	0.2963	0.0852	0.3861	2.26	1.95
Llama 2-13B	0.54	0.5343	0.3148	0.0325	0.2560	2.26	1.95
BLOOM-7B1	0.45	0.1471	0.0000	0.0061	0.0184	2.62	1.80
UCCIX$_{npd}$ (Ours)	0.61	0.7206	0.3704	0.2950	0.4563	3.47	1.57
UCCIX$_{nte}$ (Ours)	**0.80**	0.7402	0.3333	0.3297	0.4631	2.26	1.95
UCCIX (Ours)	0.75	**0.7794**	**0.3889**	**0.3334**	**0.4636**	3.47	1.57

Table 3. Ablation Experiment - Evaluation results of pre-trained models on English benchmarking datasets.

Model	Exact-match on Natural Question (5-shot)	Exact-match on IrishQA (en) (5-shot)	Acc. on Wino-grande (5-shot)	Acc. norm on Hel-laSwag (10-shot)
gpt-3.5-turbo	**0.4660**	0.3333	N/A	N/A
Llama-2-70B	0.3806	0.4074	**0.8374**	**0.8701**
Llama 2-13B	0.3069	**0.4444**	0.7609	0.8223
BLOOM-7B	0.0806	0.1667	0.6519	0.6202
UCCIX$_{npd}$ (Ours)	0.1848	0.3704	0.7017	0.7695
UCCIX$_{nte}$ (Ours)	0.1584	0.3704	0.6969	0.7690
UCCIX (Ours)	0.1668	0.3704	0.7135	0.7758

Table 4. Ablation Experiment - Evaluation of instruction-tuned models on English and Irish versions of MT-bench.

Model	MT-Bench (en)	MT-Bench (ga)
gpt-3.5-turbo	**8.63**	**5.47**
Llama 2-70B-chat	6.86	2.22
Llama 2-13B-chat	6.65	1.94
UCCIX-Instruct (Ours)	5.82	5.01

tion directions, with substantial improvements of 0.1471 BLEU-4 on gaHealth-ga2en and 0.0379 BLEU-4 on gaHealth-en2ga. Furthermore, The results on IrishQA and the examples from Figure 1 show that the baseline models are proned toward inherent biases, as their responses while in Irish, usually unrelated to the posed questions. We note that some results for the closed-source gpt-3.5-turbo model are not available, as they do not provide loglikelihood information in their APIs, which is required for some benchmarking tasks [8].

Catastrophic Forgetting. While our models exhibit robust performance in Irish-related tasks, we notice instances of catastrophic forgetting when benchmarked on English-related datasets, as demonstrated in Table 3. This might be due to the iterative training on Irish corpora, leading the model to forget the knowledge as previously trained on English dataset. This highlights a potential area for further improvement in model adaptation and fine-tuning techniques.

Instruction Following Capability. We follow the approach outlined in [28] and introduce our Irish-translated version of MT-bench, which is a challenging multi-turn question set spanning 8 categories. We evaluate the instruction-following capability of UCCIX-Instruct on both the original English and our proposed Irish versions. As illustrated in Table 4, UCCIX-Instruct demonstrates sophisticated performance on the Irish version, nearly rivaling that of gpt-3.5-turbo (5.47 compared to 5.01), despite being more than 10 times smaller in size. However, we also notice a performance degradation on English version, compared to the Llama 2-13B-chat, which is of the same size as ours. This aligns with the observation of catastrophic forgetting.

Generation Efficiency. To assess the efficiency gained through the expanded vocabulary with native Irish tokens, we compute two metrics: tokenizer speed and fertility rate. Tokenizer speed, representing the proportion of split tokens to the number of characters (chars/token), and fertility rate, indicating the average number of tokens required to represent a word or document (tokens/word), are computed within the vocabulary as the proportion of the number of tokens to the tokens that start with "space" character. As reported in Table 2, our expanded tokenizer exhibits a generation speed 1.54 times faster (3.47) than the base Llama 2's tokenizer (2.26), establishing it as the fastest among models. Moreover, as seen in comparisons between UCCIX and UCCIX$_{nte}$ in Table 2 and 3, the expansion of vocabulary also leads to reliable performance enhancement.

Ablation Study Without Parallel Data. We carry out a training experiment with the same base LLama 2-13B LLM, but without the scheduling of parallel data at the start of continued pre-training, denoted as UCCIX$_{npd}$. The results in Table 2 and 3 highlight the usefulness of our proposed approach of utilizing parallel corpora to bridge the gaps of linguistic differences between the two languages.

4 Conclusion

In this paper, we develop a novel Irish-based LLM, UCCIX, to challenge the task of training LLMs for low-resource languages, particularly Irish. Through meticulous data curation and cleaning and innovative training strategies and techniques, we have demonstrated the feasibility of creating robust LLMs despite constrained resources. Our model showcases impressive proficiency in understanding and generating Irish text, with notable achievements in various benchmarking tasks. We contribute comprehensive benchmarking datasets, including IrishQA and Irish version of MT-bench, which enable thorough evaluation and facilitate future research. We believe that our work has important implications for preserving and promoting the Irish language, and can serve as a blueprint for adapting LLMs to other low-resource languages, helping to bridge the gap between high- and low-resource languages in NLP.

Acknowledgements

We would like to acknowledge CloudCIX Limited for the generous collaborative support of computing resources on their NVIDIA HGX/H100 GPU cluster. This research work has emanated from research conducted with financial support from Science Foundation Ireland under Grant 12/RC/2289-P2 and 18/CRT/6223.

References

[1] Cloud Translation — cloud.google.com. https://cloud.google.com/translate?hl=en. [Accessed 09-05-2024].

[2] David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee, 'SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects', in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds., Yvette Graham and Matthew Purver, pp. 226–245, St. Julian's, Malta, (March 2024). Association for Computational Linguistics.

[3] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen Marcus McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck, 'Llemma: An open language model for mathematics', in *The Twelfth International Conference on Learning Representations*, (2024).

[4] Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza, 'ParaCrawl: Web-scale acquisition of parallel corpora', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, eds., Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, pp. 4555–4567, Online, (July 2020). Association for Computational Linguistics.

[5] James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J. Ó Meachair, and Jennifer Foster, 'gaBERT — an Irish language model', in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, eds., Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, pp. 4774–4788, Marseille, France, (June 2022). European Language Resources Association.

[6] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie, 'A survey on evaluation of large language models', *ACM Trans. Intell. Syst. Technol.*, **15**(3), (mar 2024).

[7] European Language Resource Coordination, *AI for Multilingual Europe – Why Language Data Matters*, ELRC Consortium, 2022.

[8] Leo Gao et al. A framework for few-shot language model evaluation, 12 2023.

[9] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff, 'Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages', in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, eds., Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, pp. 759–765, Istanbul, Turkey, (May 2012). European Language Resources Association (ELRA).

[10] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2023.

[11] Jordan Hoffmann et al., 'An empirical analysis of compute-optimal large language model training', in *Advances in Neural Information Processing Systems*, eds., Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, (2022).

[12] Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze, 'Glot500: Scaling multilingual corpora and language models to 500 languages', in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds., Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, pp. 1082–1117, Toronto, Canada, (July 2023). Association for Computational Linguistics.

[13] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed, 'Mistral 7b', *CoRR*, **abs/2310.06825**, (2023).

[14] Seamus Lankford, Haithem Afli, and Andy Way, 'Machine translation in the covid domain: an English-Irish case study for LoResMT 2021', in *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, eds., John Ortega, Atul Kr. Ojha, Katharina Kann, and Chao-Hong Liu, pp. 144–150, Virtual, (August 2021). Association for Machine Translation in the Americas.

[15] Risto Luukkonen et al., 'FinGPT: Large generative models for a small language', in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, eds., Houda Bouamor, Juan Pino, and Kalika Bali, pp. 2710–2726, Singapore, (December 2023). Association for Computational Linguistics.

[16] Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling, 2024.

[17] Simon Mille, Elaine Uí Dhonnchadha, Lauren Cassidy, Brian Davis, Stamatia Dasiopoulou, and Anya Belz, 'Generating Irish text with a flexible plug-and-play architecture', in *Proceedings of the 2nd Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*, eds., Mihai Surdeanu, Ellen Riloff, Laura Chiticariu, Dayne Frietag, Gus Hahn-Powell, Clayton T. Morrison, Enrique Noriega-Atala, Rebecca Sharp, and Marco Valenzuela-Escarcega, pp. 25–42, Singapore, (December 2023). Association for Computational Linguistics.

[18] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi, 'Cross-task generalization via natural language crowdsourcing instructions', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds., Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, pp. 3470–3487, Dublin, Ireland, (May 2022). Association for Computational Linguistics.

[19] Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel, 'Scaling data-constrained language models', in *Thirty-seventh Conference on Neural Information Processing Systems*, (2023).

[20] Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages, 2023.

[21] Long Ouyang et al., 'Training language models to follow instructions with human feedback', in *Advances in Neural Information Processing Systems*, eds., Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, (2022).

[22] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, 'Exploring the limits of transfer learning with a unified text-to-text transformer', *J. Mach. Learn. Res.*, **21**, 140:1–140:67, (2019).

[23] Hugo Touvron et al., 'Llama 2: Open foundation and fine-tuned chat models', (2023).

[24] *Atlas of the world's languages in danger*, ed., UNESCO, Memory of Peoples Series, UNESCO, 3 edn., February 2010.

[25] BigScience Workshop et al. Bloom: A 176b-parameter open-access multilingual language model, 2022.

[26] Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla, 'A paradigm shift in machine translation: Boosting translation performance of large language models', in *The Twelfth International Conference on Learning Representations*, (2024).

[27] Zheng Xin Yong et al., 'BLOOM+1: Adding language support to BLOOM for zero-shot prompting', in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds., Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, pp. 11682–11703, Toronto, Canada, (July 2023). Association for Computational Linguistics.

[28] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica, 'Judging LLM-as-a-judge with MT-bench and chatbot arena', in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, (2023).