Findings of WMT 2024's MultiIndic22MT Shared Task for Machine Translation of 22 Indian Languages

Raj Dabre[1] **Anoop Kunchukuttan**[2]
[1]National Institute of Information and Communications Technology, Japan
[2]Microsoft, India
raj.dabre@nict.go.jp ankunchu@microsoft.com

Abstract

This paper presents the findings of the WMT 2024's MultiIndic22MT Shared Task, focusing on Machine Translation (MT) of 22 Indian Languages. In this task, we challenged participants with building MT systems which could translate between any or all of 22 Indian languages in the 8th schedule of the Indian constitution and English. For evaluation, we focused on automatic metrics, namely, chrF, chrF++ and BLEU.

1 Introduction

India is a linguistically diverse region, with 1,369 distinct natively spoken languages which were identified in the census conducted in 2011. Among these native languages, 22 have been listed in the 8th Schedule of the Constitution of India. Furthermore, about 97% of the population of India speaks one of these 22 languages as their first language in their daily lives. It is important to note that English is widely spoken and serves as the default medium of formal communication in many areas, particularly in business, education, government, and judiciary. However, the percentage of the population speaking English is approximately 10% and in the interest of smooth and clear communication, the importance in India of language translation for effective communication, social inclusion, equitable access, and national integrity cannot be over-emphasized.

Having established that Indian language MT is important, the only way to improve it is via active involvement from MT researchers and MT system developers to push the boundaries of translation quality. To this end, we offered the first of its kind shared task focusing on MT for all 22 scheduled Indian languages. Over half a decade ago, in the Workshop on Machine Translation 2018 (WAT 2018) (Nakazawa et al., 2018), the organizers introduced the IndicMT task for the first time spanning covering 7 Indic languages. Over the years they gradually added languages in WAT from 2018 to 2023 (Nakazawa et al., 2019, 2020, 2021, 2022, 2023), with WAT 2023 boasting 19 Indian languages. Over the years, with the increasing number of languages and datasets for Indian languages, these tasks have garnered growing attention, however the challenge still remains since Indian languages are still resource poor in comparison with European languages.

This year the multilingual Indian languages MT task, referred to as MultiIndic22MT, is hosted under the Ninth Conference on Machine Translation (WMT24) and for the first time ever, the task spans all 22 scheduled languages of India belonging to 4 language families and written in 12 scripts. The languages exhibit both genetic and contact relatedness (Kunchukuttan et al., 2018). Some of these languages are extremely low-resource. This diversity makes this language group ideal for studies in multilingual learning, language relatedness and low-resource MT. Our primary goal behind having this shared task was to attract both researchers and developers to identify effective practices for pushing the quality of Indian language Machine Translation, especially for the lower resourced languages. Our secondary goal was also to identify some interesting but yet unexplored practices, even if they do not lead to state-of-the-art MT performance.

2 MultiIndic22MT Shared Task

The task covered English and 22 Indic Languages, as follows:

1. Assamese
2. Bengali
3. Bodo
4. Dogri

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 669–676
November 15–16, 2024.

5. Konkani
6. Gujarati
7. Hindi
8. Kannada
9. Kashmiri (Arabic script)
10. Maithili
11. Malayalam
12. Marathi
13. Manipuri (Meitei script)
14. Nepali
15. Oriya
16. Punjabi
17. Sanskrit
18. Santali
19. Sindhi (Devanagari script)
20. Tamil
21. Telugu
22. Urdu.

We evaluated user submissions on 44 translation directions (English-Indic and Indic-English). We also evaluate the performance of 5 Indic-Indic pairs: Bengali-Hindi, Tamil-Telugu, Hindi-Malayalam and Sindhi-Punjabi. We encouraged the use of multilingualism and transfer-learning by leveraging monolingual data, back-translation and (potentially) LLMs, to develop high quality systems. Although the intention is to have users develop multilingual systems and submit translations for all directions, we also welcomed submissions for specific language pairs. The link to the shared task page is here[1].

3 Datasets and Pre-trained Models

For this shared task, we prepared a fairly extensive list of resources for the participants to train their MT systems. We also describe the evaluation sets.

3.1 Datasets

We allowed participants to use existing mined as well as back-translated parallel data along with monolingual data.

Parallel Data: As a source of parallel corpora, we recommended using the Bharat Parallel Corpus Collection (BPCC) dataset[2] (Gala et al., 2023) which spans all 22 languages in the shared task. BPCC is a comprehensive and publicly available parallel corpus that includes both existing and new data for all 22 scheduled Indic languages. It comprises two parts: BPCC-Mined and BPCC-Human, totaling approximately 230 million bitext pairs. BPCC-Mined contains about 228 million pairs, with nearly 126 million pairs newly added as a part of this work. On the other hand, BPCC-Human consists of 2.2 million gold standard English-Indic pairs, with an additional 644K bitext pairs from English Wikipedia sentences (forming the BPCC-H-Wiki subset) and 139K sentences covering everyday use cases (forming the BPCC-H-Daily subset). It is worth highlighting that BPCC provides the first available datasets for 7 languages and significantly increases the available data for all languages covered. Note that one may pivot via English to obtain Indic-Indic parallel corpora.

Parallel Back-translated Data: Additionally, BPCC also contains back-translation data generated by intermediate checkpoints of IndicTrans2 (Gala et al., 2023) models for training purposes. This data consists of English original sentences translated to 22 Indic languages for a total of 401.9M back-translated sentences and Indian language original sentences translated to English for a total of 400.9M back-translated sentences. The mined, human curated and back-translated corpora represent an extensive training dataset which we expect will be sufficient for training MT systems of reasonable quality.

Monolingual Data: We also recommended the use of monolingual data from Varta[3] (Aralikatte et al., 2023), IndicCorp v2[4] (Doddapaneni et al., 2023) and Sangraha[5] (Khan et al., 2024) corpora. Sangraha subsumes IndicCorp v2 but does not explicitly include Varta. Sangraha covers 22 languages, containing a total of 251B tokens, of which con-

[1] https://www2.statmt.org/wmt24/multiindicmt-task.html

[2] https://github.com/AI4Bharat/IndicTrans2

[3] https://huggingface.co/datasets/rahular/varta

[4] https://github.com/AI4Bharat/IndicBERT/tree/main#indiccorp-v2

[5] https://github.com/AI4Bharat/IndicLLMSuite?tab=readme-ov-file#sangraha

tains verified[6] (64B), unverified[7] (24B), and synthetic[8] (162B) tokens. On the other hand, Varta spans only 9B tokens and belongs to the NEWS domain, whereas Sangraha spans multiple domains. Our evaluation sets, which we will describe later, are multi-domain (including news) and hence we expected Sangaraha to be a better source but could not neglect Varta due to its domain specificity and high quality.

3.2 Pre-trained Models

In addition to datasets, following recently followed trends in shared tasks, we encouraged participants to leverage one or all of the following publicly available models for fine-tuning or synthetic data generation:

IndicTrans2 (Gala et al., 2023): This consists of the 3 IndicTrans2 models, one-to-many, many-to-one, and many-to-many, for English to Indic, Indic to English and Indic to Indic translation. These are the current state-of-the-art open-source MT systems, and we encouraged participants to build on top of these models to improve performance, especially for the lower resourced languages like Santali, Sindhi, Bodo, Dogri, Konkani, Kashmiri, Maithili and Manipuri.

mT5 (Xue et al., 2021): mT5 is a well known pre-trained model which supports half of the Indian languages in this shared task. However, it is only pre-trained and not fine-tuned for MT and is more suitable for focused domain specific fine-tuning investigations.

IndicBART (Dabre et al., 2022): IndicBART is a small pre-trained model for 11 Indic languages and English which, when fine-tuned, is known to outperform mBART (Liu et al., 2020) and give comparable performance as a mT5, despite both models being twice its size.

VartaT5 (Aralikatte et al., 2023): This is a T5 model specific for Indic languages and is analogous to IndicBART.

BLOOM (Workshop et al., 2023): BLOOM is a family of decoder only pre-trained models supporting 44 languages, some of them being a subset of the Indian languages we focus on in this shared task. Model sizes range from 500 million parameters to 176 billion parameters. However, BLOOM is known to be an under-trained model, and thus we expected participants to focus more on using Gemma.

Gemma (Team et al., 2024): Is another family of decoder only models with 2 and 7 billion parameters. Gemma is theoretically capable of tokenizing all 22 Indian languages of this task but its primary support is more in favor of the higher resource languages like Hindi, Marathi, Bengali, etc. We expected that participants would explore some prompting approaches on top of Gemma to determine its viability for Indian language translation.

4 Submission Criteria

We expected two types of submissions: Constrained and Unconstrained. Constrained submissions were those which used the data and models stipulated by the organizers explicitly. Unconstrained submissions were those where any other data or models were used without confirmation from the organizers. Furthermore, we encouraged primary and contrastive submissions, where participants could submit one Primary (ranked) and one Contrastive (unranked, optional).

5 Evaluation Sets and Metrics

Evaluation Sets: We provide participants with a validation set and 3 test sets. The validation set is an extension of FLORES-200 for the 22 Indian languages[9], as described in Gala et al. (2023) and consists of 997 23-way sentences. As for the test sets, 2 out of 3 are publicly available and one is a hidden test set. The publicly available sets are In22-Conv[10] and In22-Gen[11] spanning 1,503 and 1,024 23-way parallel sentences, for the conversational and general styles, respectively. The hidden test set was originally described in Chitale et al. (2024) and is an Indic language original test set where Indic sentences were translated into English by linguists. This is different from all other test sets which are English original and were translated into Indic languages. This hidden test set was released to the participants 2 weeks before the deadline and unlike In22-Conv and In22-Gen, the references

[6]The URLs of webpages from which the corpora were crawled were manually verified by linguists.

[7]The urls of webpages from which the corpora were crawled were unverifiable.

[8]These were obtained by translating English documents into Indian languages.

[9]https://indictrans2-public.objectstore.e2enetworks.net/flores-22_dev.zip

[10]https://huggingface.co/datasets/ai4bharat/IN22-Conv

[11]https://huggingface.co/datasets/ai4bharat/IN22-Gen

were kept hidden. This test set covered only 13 of the 22 Indic languages namely, Assamese, Bengali, Bodo, Gujarati, Hindi, Kashmiri, Malayalam, Nepali, Santali, Sanskrit, Sindhi, Telugu and Urdu. While we asked participants to work on translation to and from English for In22-Conv and In22-Gen, for the hidden test set, only translation from Indic to English direction was possible in order to keep the test set hidden[12].

Evaluation Metrics: We asked participants to submit their translations to us which we would then evaluate using BLEU (Papineni et al., 2002), chrF (Popović, 2015) and chrF++ (Popović, 2017) using sacreBLEU[13] (Post, 2018). We follow the appropriate tokenization of Indic languages as done by Gala et al. (2023) before computing scores.

6 Participants and Submissions

Although 32 teams had registered initially, only 4 teams ended up submitting systems and 3 submitted system description papers (1 withdrew). The teams and their submitted systems are as follows:

6.1 BV-SLP Team

The BV-SLP team (Joshi et al., 2024), short for the Banasthali Vidyapith Speech and Language Processing Lab, focused on Sindhi to English translation and only submitted translations for the hidden test set. Their approach focuses on special handling of named entities. They first extract named entities from the source Sindhi sentence and translate it first using a knowledge base of Sindhi-English named entity pairs. This intermediate output is then translated using a NMT system, which is trained to retain the translated named entities and only translate the Sindhi part. To develop the NMT system itself, they converted the existing Sindhi-English parallel corpus into a form where the Sindhi sentences had their named entities replaced with their English translations. This pre-translation approach is well known to work well for handling named entities. They used two approaches for translation itself, one (Primary) where Sindhi is directly translated into English and one (Contrastive) where Sindhi is first translated into Hindi and then into English.

[12]Asking for English to Indic translation meant that we would have to release English sentences too and this would lead to the test set references being exposed.

[13]https://github.com/mjpost/sacrebleu

6.2 NITS-CNLP Team

The NITS-CNLP team (Singh et al., 2024), short for the National Institute of Technology Silchar's Centre for Natural Language Processing, focused on English to Manipuri translation and submitted a primary and a contrastive system. Their approach was rather straightforward, where they used the English-Manipuri data from BPCC (Gala et al., 2023) and trained a transformer model. They submitted results for the In22-Conv and In22-Gen test sets. They also performed some manual evaluations.

6.3 NLIP-Lab Team

The NLIP-Lab (Brahma et al., 2024), short for the Natural Language and Information Processing Lab, was the only team that went all out and submitted translations for all translation directions and test sets. The NLIP-Lab team use an approach based on pre-training models using codemixed data which was synthetically created. Specifically, they take BPCC parallel data and replace words in English sentences with semantically similar words of the target Indic language sentences. They then pre-train a model with both the original and code-mixed data. They further refine their pre-trained model with original and code-mixed data obtained only from the high quality BPCC-seed datasets. Finally, they fine-tune their models only on the seed datasets without the code-mixed counterparts. They hypothesized that this leads to fairly strong MT systems.

7 Results and Findings

Overall, the NLIP-Lab team got 1st rank for all language pairs, directions and test sets, including In22-Conv, In22-Gen and the hidden test set for Indic to English translation.

7.1 Sindhi to English Translation

NLIP-Lab had a contender in the form of BV-SLP team for Sindhi to English translation but where the primary system of BV-SLP got BLEU, chrF and chrF++ scores of 19.4, 44.6 and 43.0, respectively. NLIP-Lab translations scored BLEU, chrF and chrF++ scores of 21.2, 47.1 and 45.5, respectively. This showed that NLIP-Lab's RASP pre-training and fine-tuning approach was definitely better than the named entity handling approach. The likely explanation was that NLIP-Lab used a

lot more parallel data and trained a larger model than their competitior.

7.2 English to Manipuri Translation

Once again, NLIP-Lab's contender for the English to Manipuri task was the NITS-CNLP lab. This was for the In22-Conv and In22-Gen test sets. NITS-CNLP got BLEU, chrF and chrF++ scores of 6.4, 28.6 and 26.6 for In22-Conv and 8.1, 32.1 and 29.4 for In22-Gen. However, NLIP-Lab got better scores of 15.2, 43.6 and 41.1 for In22-Conv and 18.2, 48.0 and 45.0 for In22-Gen. This shows that NLIP-Lab's systems are substantially better. However, this is to be expected given that NITS-CNLP did not train massively multilingual models and the latter did.

7.3 Did NLIP-Lab Beat IndicTrans2?

Unfortunately, NLIP-Lab's systems did not beat IndicTrans2. For the Indic to English directions, IndicTrans2 was almost 10 BLEU better on In22-Gen and almost 4 BLEU better on In22-Conv. For the English to Indic directions, however, the gap narrowed down to about 2 BLEU. This implies that despite IndicTrans2 being trained on significantly larger data (mostly backtranslated) and in multiple stages, its performance can still be approached by systems not leveraging massive amounts of data. This highlights then need for investigating better approaches for translating into Indic languages. As a side note, these same observations hold for Indic to Indic translation.

8 Conclusion

In this report we present the findings of the MultiIndic22MT shared task for machine translation involving 22 Indian languages. Despite the initial enthusiasm shown by participants during task registration, only 3 out of 32 teams submitted their translations and system description papers. Of these 3, only NLIP-Lab submitted translations for all directions and got first rank for all their submissions. Approaches explored varied from named entity replacement, pivot language translation (using Hindi as a pivot), code-mixed pretraining and training from scratch. Overall, code-mixed pre-training stood tall and led to the best systems. However, none of the systems could still beat IndicTrans2, indicating that there is much effort needed for pushing the state of the art for translation involving Indian languages. Given the advent of LLMs and the focus on decoder-only architectures which are well suited for document level MT, we expect that the next batch of innovations will be focused on the same. However, most LLMs dont support Indic languages that well and thus participants may have to resort to using approaches like transliteration to bridge the gap or even reduce it between the type of scripts that LLMs have seen and those that they have not (J et al., 2024; Dabre et al., 2020, 2022; Gala et al., 2023). We hope that more people will participate in another iteration of this task with interesting approaches.

References

Rahul Aralikatte, Ziling Cheng, Sumanth Doddapaneni, and Jackie Chi Kit Cheung. 2023. Varta: A large-scale headline-generation dataset for Indic languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3468–3492, Toronto, Canada. Association for Computational Linguistics.

Maharaj Brahma, Pramit Sahoo Maunendra, and Sankar Desarkar. 2024. Nlip_lab-iith multilingual mt system forwat24 mt shared task. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Pranjal Chitale, Jay Gala, and Raj Dabre. 2024. An empirical study of in-context learning in LLMs for machine translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7384–7406, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M

Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Jaavid J, Raj Dabre, Aswanth M, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. RomanSetu: Efficiently unlocking multilingual capabilities of large language models via Romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615, Bangkok, Thailand. Association for Computational Linguistics.

Nisheeth Joshi, Pragya Katyayan, Palak Arora, and Bharti Nathani. 2024. System description of bvslp for sindhi-english machine translation in multiindic22mt 2024 shared task. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Mohammed Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh Khapra. 2024. IndicLLMSuite: A blueprint for creating pre-training and fine-tuning datasets for Indian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15831–15879, Bangkok, Thailand. Association for Computational Linguistics.

Anoop Kunchukuttan, Mitesh Khapra, Gurneet Singh, and Pushpak Bhattacharyya. 2018. Leveraging orthographic similarity for multilingual neural transliteration. *Transactions of the Association for Computational Linguistics*, 6:303–316.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Yusuke Oda, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35, Hong Kong, China. Association for Computational Linguistics.

Toshiaki Nakazawa, Kazutaka Kinugawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Makoto Morishita, Ondřej Bojar, Akiko Eriguchi, Yusuke Oda, Chenhui Chu, and Sadao Kurohashi. 2023. Overview of the 10th workshop on Asian translation. In *Proceedings of the 10th Workshop on Asian Translation*, pages 1–28, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Anoop Kunchukuttan, Makoto Morishita, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2022. Overview of the 9th workshop on Asian translation. In *Proceedings of the 9th Workshop on Asian Translation*, pages 1–36, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Online. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44, Suzhou, China. Association for Computational Linguistics.

Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. 2018. Overview of the 5th workshop on Asian translation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ningthoujam Justwant Singh, Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, and Thoudam Doren Sing. 2024. Wmt24 system description for the multiindic22mt shared task on manipuri language. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan

Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. 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*, pages 483–498, Online. Association for Computational Linguistics.