THE OBSCURE LIMITATION OF MODULAR MULTILINGUAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We expose the limitation of modular multilingual language models (MLMs) in multilingual inference scenarios with unknown languages. Existing evaluations of modular MLMs exclude the involvement of language identification (LID) modules, which obscures the performance of real-case multilingual scenarios of modular MLMs. In this work, we showcase the effect of adding LID on the multilingual evaluation of modular MLMs and provide discussions for closing the performance gap of caused by the pipelined approach of LID and modular MLMs.

1 INTRODUCTION

Multilingual language models (MLMs) suffer from the capacity limitation problem known as the **curse of multilinguality**, which penalizes the efficiency of MLMs, both in terms of training and inference, for acquiring new languages. Prior works (Pfeiffer et al., 2020; Ansell et al., 2021; Pfeiffer et al., 2022) alleviate the inference inefficiency bottleneck of the curse of multilinguality by introducing modularity in MLMs through language adapters. This modularity allows MLMs to scale the number of parameters with minimal cost on the training and inference speed. One limitation of modular MLMs is that, as shown in Figure 1, the language of the input needs to be known prior to the inference step for selecting the language adapter. Nevertheless, multilingual evaluations of these modular MLMs make an assumption that an ideal language identification is given and use the language metadata provided on the evaluation data to select the correct language adapter. This produces a gap between modular MLMs in the simulated setting and in the real multilingual scenario. In this work, we address the evaluation gap and further discuss how to mitigate the limitation of modular MLMs.

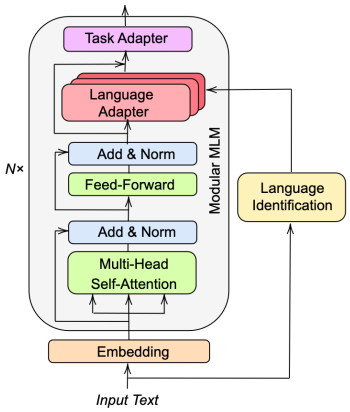

Figure 1: Modular MLMs incorporate language-specific adapters to learn new languages. This renders them language-dependent and reliant on external LID for inference.

2 RELATED WORKS

Multilingual Language Model MLMs (Conneau et al., 2020; Liu et al., 2020; Xue et al., 2021; Workshop et al., 2022) are effective for solving various language understanding and generation in various languages (Hu et al., 2020; Wilie et al., 2020; Cahyawijaya et al., 2021; Adelani et al., 2022; Kumar et al., 2022). To solve the curse of multilinguality of MLMs, the modular MLM approach is introduced. MAD-X (Pfeiffer et al., 2020) and MAD-G Ansell et al. (2021) use adapt MLMs to new languages by using language adapters. X-MOD (Pfeiffer et al., 2022) introduces modularity during pre-training which better aligns modular MLMs across languages.

Language Identification (LID) The LID task is introduced over five decades ago (Gold, 1967). Since then, various methods for LID have been introduced, such as n-gram similarity (Cavnar & Trenkle, 1994), naive bayes (Baldwin & Lui, 2010; Lui & Baldwin, 2012; Sites, 2013), and gaussian mixture (Lui et al., 2014). More recently, embedding-based methods using character (Alex Salcianu, 2020) and subwords (Joulin et al., 2017) have also been introduced. In this work, we explore the effect of utilizing these LID modules on the performance of modular MLMs.

LID Model	HRL	MRL	LRL	AVG
Fully support languages under study				
FastText	**97.22**	96.26	88.96	**93.89**
CLD3	87.84	89.30	**91.47**	89.57
CLD2	76.07	90.85	85.14	83.17
Partially support languages under study[2]				
langid.py	92.00	93.04	76.12	86.31
LangDetect	69.26	**96.45**	42.97	66.20

Table 1: Accuracy score of LIDs on MASSIVE. Most LIDs perform well on **HRL** and **MRL**, but the score falls short on **LRL**. **Bold** and underline denote first and second best, respectively.

NLU Model	HRL	MRL	LRL	AVG
Direct fine-tuning				
XLMR	86.03	84.76	83.20	84.65
mBERT	84.76	82.50	80.62	82.64
Language adapter tuning				
MAD-X (No LID)	83.30	80.96	79.46	81.27
MAD-X (FastText)	75.21	78.08	72.46	74.90
MAD-X (CLD3)	72.90	75.20	72.89	73.47

Table 2: Accuracy score of MLMs on MASSIVE. Incorporating LID decays the performance of the language-adapter model.

3 Experimental Setting

For our experiments, we utilize MASSIVE (FitzGerald et al., 2022), a multilingual intent classification dataset covering 52 typologically-diverse languages. We select 24 languages from MASSIVE and group them into 3 different resource groups based on the language size in CommonCrawl[1], i.e., high-resource languages (HRL), medium-resource languages (MRL), and low-resource languages (LRL). A detailed list of languages under study and the resource grouping is described in Appendix A. For the LID, we incorporate 5 off-the-shelf LID models, i.e., LangDetect (Nakatani, 2011), langid.py (Lui & Baldwin, 2012), FastText LID (Joulin et al., 2017), CLD2 (Sites, 2013), and CLD3 (Alex Salcianu, 2020). We evaluate these LIDs and take the best two LIDs for the multilingual evaluation with unknown languages. For the modular MLM, we utilize MAD-X Pfeiffer et al. (2020) with mBERT backbone. We compare the MAD-X with LID against two direct fine-tuned MLMs and MAD-X without LID. We use accuracy score as the evaluation metric in our experiment.

4 Result & Discussion

Based on the result of the LID experiment in Table 1, we select FastText and CLD3 for evaluating modular MLMs with unknown languages. The modular MLMs result is shown in Table 2. For the modular MLM without LID, our result aligns with prior works Pfeiffer et al. (2020); Ansell et al. (2021) yielding a slightly lower score compared to the direct fine-tuned models. Both modular MLMs with LID produce an even lower performance in all language resource groups compared to the modular MLM without LID, resulting in a gap of $\sim$7-8% accuracy score over all language groups. The detailed result of our experiment is shown in Appendix B.

We clearly observe that existing off-the-shelf LID is far from the ideal case which widens the gap to the direct fine-tuning approach and raises an open question for closing the performance gap. To address the question, it is important to understand the limitations of using modular MLMs with off-the-shelf LIDs. Several potential limitations that might occur include: 1) distribution shift of LIDs caused by domain and time differences , 2) label mismatch between LID and the language adapter, and 3) other linguistic problems that affect LIDs such as code-mixing and creole language. We leave the exploration of the solution to these potential limitations for future works.

5 Conclusion

In this work, we show the limitation of modular multilingual language models (MLMs) in inferencing with unknown languages. We evaluate the effect of using off-the-shelf LID modules on the evaluation of modular MLMs. Our result suggests that using off-the-shelf LID modules significantly decreases the performance of modular MLMs by $\sim$7-8% accuracy which widens the gap between modular MLMs and non-modular MLMs. In addition, we discuss several potential limitations that might contribute to the performance gap of using off-the-shelf LID with modular MLMs.

[1] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

[2] We zero out the performance for all the unsupported languages.

URM STATEMENT

All authors of this paper qualify as an underrepresented minority (URM) for the "Tiny Papers" track at ICLR 2023.

REFERENCES

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4488–4508, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.298.

Anton Bakalov Chris Alberti Daniel Andor David Weiss Emily Pitler Greg Coppola Jason Riesa Kuzman Ganchev Michael Ringgaard Nan Hua Ryan McDonald Slav Petrov Stefan Istrate Terry Koo Alex Salcianu, Andy Golding. Compact language detector v3 (cld3), 2020. URL https://github.com/google/cld3.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4762–4781, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.410. URL https://aclanthology.org/2021.findings-emnlp.410.

Timothy Baldwin and Marco Lui. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 229–237, Los Angeles, California, June 2010. Association for Computational Linguistics. URL https://aclanthology.org/N10-1027.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8875–8898, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.699. URL https://aclanthology.org/2021.emnlp-main.699.

William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161–175, Las Vegas, US, 1994.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages, 2022.

E Mark Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967. ISSN 0019-9958. doi: https://doi.org/10.1016/S0019-9958(67)91165-5. URL https://www.sciencedirect.com/science/article/pii/S0019995867911655.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4411–4421. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hu20b.html.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2068.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M. Khapra, and Pratyush Kumar. IndicNLG benchmark: Multilingual datasets for diverse NLG tasks in Indic languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5363–5394, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.360.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47.

Marco Lui and Timothy Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pp. 25–30, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/P12-3005.

Marco Lui, Jey Han Lau, and Timothy Baldwin. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2:27–40, 2014. doi: 10.1162/tacl_a_00163. URL https://aclanthology.org/Q14-1003.

Shuyo Nakatani. Language detection library for java, 2011. URL https://github.com/shuyo/language-detection.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7654–7673, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL https://aclanthology.org/2020.emnlp-main.617.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3479–3495, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.255. URL https://aclanthology.org/2022.naacl-main.255.

Richard Sites. Compact language detector v2 (cld2), 2013. URL https://github.com/CLD2Owners/cld2.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 843–857, 2020.

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 Haji-Hosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, 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. Bloom: A 176b-parameter open-access multilingual language model, 2022. URL https://arxiv.org/abs/2211.05100.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

A Language Under Study

We provide the list of all languages under study along with the language resource group in Table 3. Language resource is grouped by the size of language data in CommonCrawl, i.e., high-resource languages (≥**1%**, medium-resource languages ((≥**0.1%**), and low-resource languages (<**1%**).

Language	#Speaker	CC Size	Resource Group
ar-SA	360M	0.665%	MRL
bn-BD	300M	0.093%	LRL
de-DE	95M	5.662%	HRL
el-GR	13.5M	0.597%	MRL
en-US	373M	46.320%	HRL
es-ES	493M	4.435%	HRL
fi-FI	5.4M	0.398%	LRL
fr-FR	300M	4.604%	HRL
hi-IN	528M	0.155%	LRL
hu-HU	13M	0.599%	MRL
hy-AM	5.4M	0.032%	LRL
id-ID	300M	0.781%	MRL
is-IS	0.3M	0.038%	LRL
ja-JP	128M	4.532%	HRL
jv-ID	82M	0.002%	LRL
ka-GE	3.7M	0.037%	LRL
ko-KR	79.3M	0.679%	MRL
lv-LV	1.2M	0.082%	LRL
my-MM	33M	0.012%	LRL
pt-PT	250M	1.482%	HRL
ru-RU	258M	5.717%	HRL
vi-VN	70M	0.962%	MRL
zh-CN	920M	4.837%	HRL
zh-TW	4.6M	4.837%[3]	HRL

Table 3: List of languages under study in our experiments. The number of speaker information is retrieved from Wikipedia.

[3]We use the same number of zh-CN and zh-TW, since there is no Chinese (zh) language variation in CommonCrawl.

B DETAILED EXPERIMENT RESULT

We provide the complete per language result for the language identification and the modular MLMs experiments in Table 4 and Table 5.

Language	LID-Fasttext	CLD3	CLD2	langid	LangDetect
ar-SA	94.25	86.45	81.58	91.78	94.13
bn-BD	99.72	97.52	89.57	96.93	99.76
de-DE	97.70	88.59	89.73	92.83	82.54
el-GR	99.68	96.91	99.77	99.84	99.64
en-US	98.61	79.44	93.43	93.96	87.82
es-ES	96.20	78.24	73.14	86.87	86.55
fi-FI	97.70	92.91	92.90	92.08	96.09
fr-FR	98.35	87.53	85.23	94.77	94.80
hi-IN	98.44	88.21	97.83	87.94	93.54
hu-HU	98.54	92.24	93.89	95.34	96.71
hy-AM	99.90	98.37	99.92	99.17	0.00
id-ID	87.20	65.86	73.54	72.68	89.32
is-IS	89.93	92.64	90.88	92.97	0.00
ja-JP	99.41	96.63	99.04	99.11	96.23
jv-ID	24.75	68.10	0.00	22.04	0.00
ka-GE	99.56	98.49	99.95	99.65	0.00
ko-KR	99.50	98.47	99.03	99.96	99.36
lv-LV	90.73	90.06	95.25	94.33	97.32
my-MM	99.93	96.90	99.97	0.00	0.00
pt-PT	92.17	83.42	77.39	77.74	84.05
ru-RU	99.27	84.48	82.35	83.79	91.32
vi-VN	98.41	95.85	97.26	98.62	99.53
zh-CN	97.55	98.07	84.33	99.64	0.00
zh-TW	95.76	94.19	0.03	99.31	0.00
Average	**93.89**	**89.57**	**83.17**	**86.31**	**66.20**

Table 4: Per language results of language identification evaluation in MASSIVE.

Language	XLMR	mBERT	MAD-X	MAD-X w/ FastText	MAD-X w/ CLD3
ar-SA	79.32	78.35	75.72	71.92	67.79
bn-BD	83.25	80.23	78.61	76.36	74.95
de-DE	85.54	83.59	81.81	79.49	76.90
el-GR	85.07	81.74	80.93	79.56	78.51
en-US	88.16	86.45	85.78	83.89	83.15
es-ES	86.18	84.97	82.58	80.97	76.43
fi-FI	85.24	82.55	82.55	79.86	77.07
fr-FR	86.48	86.11	83.69	82.35	80.03
hi-IN	84.63	82.38	80.73	78.14	72.73
hu-HU	85.68	82.65	81.57	80.13	76.40
hy-AM	84.23	81.20	80.43	78.78	77.91
id-ID	86.52	84.67	82.01	76.03	69.30
is-IS	84.16	82.21	80.40	71.49	73.57
ja-JP	85.78	84.70	83.22	82.04	81.27
jv-ID	81.20	81.57	78.58	45.70	59.68
ka-GE	79.19	75.25	73.23	70.85	70.17
ko-KR	85.51	84.30	82.99	81.14	80.56
lv-LV	84.73	82.18	82.08	74.58	74.95
my-MM	82.18	78.01	78.48	76.36	74.98
pt-PT	86.35	85.27	83.59	80.56	77.77
ru-RU	86.65	83.96	83.52	81.74	75.45
vi-VN	86.48	83.32	82.52	79.72	78.61
zh-CN	85.41	85.24	84.23	53.09	52.69
zh-TW	83.73	82.55	81.27	52.79	52.45
Average	**84.65**	**82.64**	**81.27**	**74.90**	**73.47**

Table 5: Per language accuracy score of multilingual language models in MASSIVE.