Toward LLMs Beyond English-Centric Development

Anonymous ACL submission

Abstract

Through an analysis of sequences generated by open-weight large language models (LLMs), we demonstrate that LLMs are heavily biased toward English. While continual pre-training is commonly used to adapt LLMs to a target language, we show that it does not offer a cost advantage over training from scratch, even for improving cultural understanding in the target language. These findings suggest that dedicated per-language investment may become increasingly important for future LLM development, rather than relying primarily on the expansion of English-centric resources.

1 Introduction

Since the initial report on the remarkable performance of the large language model (LLM) (Brown et al., 2020), numerous groups have introduced LLMs including the Llama, Qwen, and Gemma series (Touvron et al., 2023a; Bai et al., 2023; Team et al., 2024). It appears that novel LLMs yielding performance improvements are being reported almost on a daily basis. However, because the majority of existing LLMs focus on English and major benchmark datasets are predominantly in English, comparatively little attention has been devoted to discussions involving other languages.

In fact, some studies have reported that they trained LLMs on datasets whose significant portions consist of English text (Touvron et al., 2023b; Bai et al., 2023). For example, the Llama 2 pre-training corpus contains approximately 90% English documents. Because the large-scale Web crawled corpus, CommonCrawl, includes only about 40% English documents[1], their pre-training data is heavily biased toward English compared with the natural distribution of Web documents. In this paper, we estimate the language distribution of training data for each LLM based on its generated sequences, and we demonstrate that an LLM is strongly biased toward English even when it is designed as a multilingual model (Figure 1).

To efficiently obtain an LLM that achieves superior performance in a specific language, previous studies have conducted continual pre-training (Fujii et al., 2024). We analyze such continual pre-training models and show that continual pre-training does not offer a cost advantage over training from scratch for developing LLMs specialized for a particular language. We show that continual pre-training can improve performance in the target language, but the improvement is commensurate with the additional computational cost incurred during continual pre-training. In particular, we demonstrate that performance on cultural understanding related to a specific language is proportional to the logarithmic scale of the training cost for that language, rather than to the total training cost (Figure 2). These results confirm that the efficiency of continual pre-training for such cultural understanding tasks is comparable to that of training from scratch. Based on these findings, we discuss future directions for LLM development.

2 Language Distribution of LLMs

Although some groups report the construction process of their LLMs in technical reports (Touvron et al., 2023b; Yang et al., 2025; Team et al., 2025), these reports do not reveal detailed statistics about their pre-training datasets. In this paper, we estimate the language distribution of pre-training datasets based on the assumption that the sequences generated by each LLM strongly correlate with its training data. In fact, previous studies have shown that we can extract training data from LLMs (Carlini et al., 2021; Nasr et al., 2025; Xu et al., 2025).

Estimation of Distribution We ask each LLM to generate 100,000 sequences, each consisting

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

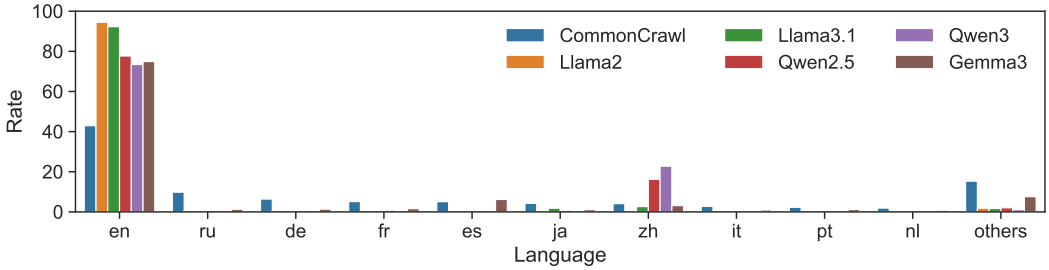


Figure 1: Estimated language distribution of pre-training data for each LLM based on randomly generated sequences. This figure focuses on top 10 languages in CommonCrawl.

Model	En	Ru	De	Fr	Es	Ja	Zh	Th	Te	Bn	Sw	Average
Qwen2.5-32B-Instruct	87.2	72.8	71.2	64.4	86.4	76.0	80.8	69.2	58.0	82.4	56.8	73.2
Qwen3-32B	95.6	86.4	88.0	82.4	91.6	77.6	89.6	88.8	84.0	83.2	74.4	85.6
Llama3.1-8B-Instruct	78.8	62.4	67.2	64.0	73.6	53.2	60.0	60.8	53.2	54.0	54.8	62.0
Gemma-3-27B-it	96.0	86.0	86.8	80.4	92.4	75.6	84.8	88.8	82.8	84.8	84.4	85.7

Table 1: 6-shot performance of each LLM on the MGSM test set. We take model names from HuggingFace Hub.

of 1,000 tokens, from the BOS token. We then identify the language of each sequence using fastText (Joulin et al., 2016; Grave et al., 2018). Finally, we regard the resulting language distribution of the generated sequences as an estimate of the language distribution of the pre-training data. To evaluate the reliability of this process, we compute the Kendall rank correlation coefficient (Kendall, 1938) between the reported language distribution and our estimated one for Llama 2 (Touvron et al., 2023b), which is one of the few studies that reveal the language distribution of their pre-training corpus. As a result, we obtained coefficients of 0.77 and 0.80 for Llama 2 7B and 13B, respectively. Because these values are significantly high, we believe that our estimation method is reasonable[2].

Comparison of Distribution Figure 1 shows the estimated language distributions of widely used open-weight LLMs and CommonCrawl. For CommonCrawl, we sampled 100,000 documents containing more than 1,000 characters and identified their languages using fastText for a fair comparison to LLMs. For the LLMs, we focus on pre-training models trained from scratch by each group. Appendix C provides the details of distributions.

Figure 1 shows that the proportion of English in the pre-training datasets is much higher than that in the CommonCrawl corpus. In other words, the LLMs in Figure 1 are strongly biased toward English. Thus, although these LLMs achieve good performance in English, they may not reach sufficient quality in other languages. In fact, such LLMs exhibit a substantial drop in performance relative to English when evaluated on language-agnostic benchmark datasets that assess language-independent abilities, such as solving math problems, across various languages. Table 1 presents the few-shot performance of each LLM on the multilingual grade school math (MGSM) dataset (Shi et al., 2023), which contains math problems translated from English into multiple languages. We set the number of shots to 6, following the setting used in the previous study (Shi et al., 2023). For example, as shown in Table 1, the performance in Russian is about 10 points lower than in English, even though Russian is the second most frequent language in CommonCrawl.

3 Effectiveness and Efficiency of Continual Pre-training

To adapt an LLM to a target language, previous studies have applied continual pre-training with the target language (Yang et al., 2023; Pires et al., 2023; Zhao et al., 2024; Cui et al., 2024; Fujii et al., 2024). These studies reported that continual pre-training can produce a strong LLM in the target language at a substantially lower computational cost than training from scratch. In this section, we re-evaluate the performance of continual pre-training on both language-agnostic benchmark datasets and benchmark datasets that require cultural understanding of the target language to answer the questions.

[2]We also investigate LLM-jp and OLMo that do not reveal their language distribution but make their pre-training corpus publicly available. Appendix A provides further details.

Model	Language-agnostic				Japanese cultural understanding		
	MGSM		MMLU ProX				
	En	Ja	En	Ja	NIILCQA	Part of JMMLU	JAQKET
Continual pre-trained models and their base model							
Llama-3.1-8B-Instruct	78.8	53.2	43.5	28.7	30.7	71.6	27.5
Llama-3.1-Swallow-8B-Instruct-v0.5	76.8	60.0	43.0	36.2	**58.3**	85.7	65.5
Qwen2.5-32B	92.0	82.4	61.1	51.0	35.4	84.3	48.4
Qwen2.5-bakeneko-32B-Instruct	**94.4**	**84.0**	**71.9**	**64.9**	32.3	83.5	51.2
Qwen2.5-32B-Instruct	87.2	76.0	68.7	61.2	26.8	84.3	49.9
ABEJA-Qwen2.5-32b-Japanese-v1.0	90.8	32.8	66.3	58.1	42.5	**90.9**	**70.5**
ELYZA-Shortcut-1.0-Qwen-32B	61.6	69.2	61.7	62.9	36.2	87.3	61.7
LLMs with pre-training data containing a large amount of Japanese text							
Sarashina2.2-3B-Instruct-v0.1	76.0	60.4	36.1	30.2	52.0	81.9	68.2
LLM-jp-3.1-13B-Instruct4	57.2	66.0	29.8	28.5	51.2	80.8	69.4

Table 2: Performance on language-agnostic benchmarks and Japanese cultural understanding benchmarks. We use 6-shot for MGSM, 5-shot for MMLU ProX, 4-shot for NIILCQA and JMMLU, and zero-shot for JAQKET.

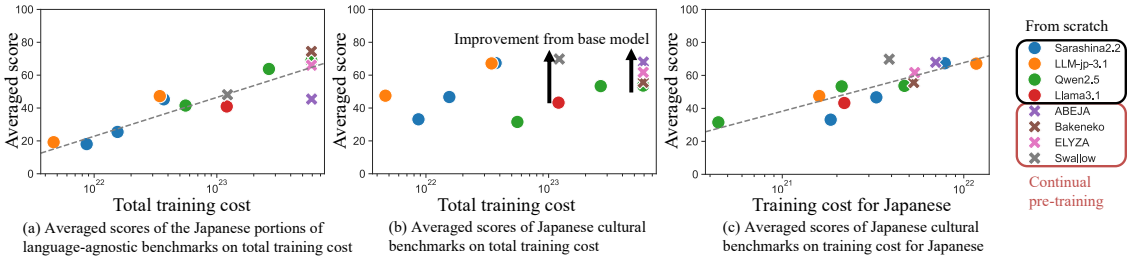
Model We focus on continual pre-training models adapted to Japanese because there are various models. Specifically, we evaluate Llama-3.1-Swallow-8B-Instruct-v0.5 trained from Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-bakeneko-32B-Instruct trained from Qwen2.5-32B (Yang et al., 2025), and ABEJA-Qwen2.5-32B-Japanese-v1.0 and ELYZA-Shortcut-1.0-Qwen-32B trained from Qwen2.5-32B-Instruct (Yang et al., 2025). These models are instruction-tuned after continual pre-training.

Dataset For language-agnostic benchmark tasks, we focus on math problem and massive multilingual language understanding (MMLU) datasets. We use the Japanese portions of MGSM (Shi et al., 2023) and MMLU ProX (Xuan et al., 2025), which are translated from English grade school math problems (Cobbe et al., 2021) and MMLU Pro (Wang et al., 2024), respectively. We also report the English portions of these datasets as a reference. In addition, we use the NIILC question answering dataset (NIILCQA) (Sekine, 2003), the subset of JMMLU related to Japanese culture (Yin et al., 2024), and JAQKET[3] as benchmark datasets to evaluate cultural understanding of Japan. We use the LM Evalution Harness[4] for language-agnostic benchmarks and FlexEval[5] for Japanese cultural understanding benchmarks.

Result 1: Effectiveness Table 2 shows the results of each LLM on language-agnostic benchmarks and Japanese cultural understanding benchmarks. This table also reports the performance of LLMs whose pre-training data contains a large amount of Japanese text, for reference. For the language-agnostic benchmarks, Llama-3.1-Swallow-8B-Instruct-v0.5 improves performance over its base model in Japanese portions. Qwen2.5-bakeneko-32B-Instruct also improves the performance. Thus, continual pre-training in the target language can improve performance on language-agnostic benchmarks. However, as shown in the results of ABEJA-Qwen2.5-32B-Japanese-v1.0 and ELYZA-Shortcut-1.0-Qwen-32B, continual pre-training may also degrade performance in the target language. Therefore, it is important to design an appropriate training strategy carefully when adapting a strong model. Moreover, for Japanese cultural understanding benchmarks, all models except Qwen2.5-bakeneko-32B-Instruct improve upon the performance of their base models.

Result 2: Efficiency Table 2 demonstrates that continual pre-training mostly improves performance in the target language. However, how efficient is continual pre-training? In other words, does continual pre-training provide a larger performance gain per unit of computational cost compared to pre-training from scratch? To investigate this question, we illustrate the relationship between performance and computational cost. Specifically, we compute the average score of the Japanese portions of the language-agnostic benchmarks and Japanese cultural benchmarks. We then examine the relationship among these averaged scores, the total computational cost, and the computational cost spent on training with Japanese texts. We use the product of the parameter size and the number of consumed tokens as the measure of computational cost. To compute the amount of Japanese training data, we use the estimated ratio obtained in Section 2.

[3] github.com/kumapo/JAQKET-dataset
[4] github.com/EleutherAI/lm-evaluation-harness
[5] github.com/sbintuitions/flexeval/



(a) Averaged scores of the Japanese portions of language-agnostic benchmarks on total training cost

(b) Averaged scores of Japanese cultural benchmarks on total training cost

(c) Averaged scores of Japanese cultural benchmarks on training cost for Japanese

Figure 2: Averaged benchmark scores for each training cost. The vertical arrows in (b) indicate the performance improvement from the base model through continual pre-training.

Figure 2 shows the relationship between performance and training cost. To construct this figure, we additionally evaluated models with varying parameter sizes for Qwen2.5, Sarashina2.2, and LLM-jp-3.1. We also plot a regression line based on the points. As shown in Figure 2 (a), the scores on the Japanese portions of language-agnostic benchmarks improve linearly with respect to the logarithmic scale of the total training cost. This result implies that continual pre-training with the target language does not have a substantial impact on the performance of language-agnostic benchmarks for the target language.

On the other hand, Figure 2 (b) shows that continual pre-training models achieve substantial improvements on the Japanese cultural understanding benchmarks when evaluated in terms of the total training cost. Based on this figure, continual pre-training appears to be a remarkably cost-efficient approach when considering only the total training cost. In contrast, when we focus on the training cost specifically spent on Japanese, Figure 2 (c) indicates that improvements are proportional to the logarithmic scale of the training cost for Japanese. Thus, the gains obtained through continual pre-training arise only from the additional training cost allocated to Japanese. In other words, the cost-performance trend of continual pre-training follows approximately the same scaling as training from scratch for the target language.

4 Discussion

Based on the findings from Sections 2 and 3, performance on language-agnostic benchmarks is likely to continue improving under recent English-centric LLM development, as such performance primarily depends on the total training cost. Furthermore, previous studies have shown that LLMs can improve performance on language-agnostic benchmarks by prompting the model to think in the predominant language, i.e., English, rather than by explicitly learning each language (Shi et al., 2023). As a result, there is limited incentive to incorporate large amounts of non-English data into LLM training when the objective is solely to improve performance on language-agnostic benchmarks. Consequently, LLMs tend to exhibit weaker cultural understanding for languages that are not major in their training corpora.

Moreover, it is worth questioning whether language-agnostic benchmarks are truly language-independent. In common benchmark construction pipelines, reasoning task datasets are first created in English and then translated into multiple languages. Thus, these benchmarks are implicitly influenced by English (or Western-centric) concepts. In fact, Singh et al. (2025) reported that the MMLU dataset contains questions biased toward Western cultural perspectives, such as *US History* and *US Law*.

As LLMs are becoming foundational tools for a wide range of applications, it is undesirable for them to embed strong cultural biases. To better understand and mitigate such biases, we emphasize the importance of developing diverse culture-specific benchmarks in addition to language-agnostic benchmarks. At a minimum, our findings encourage consideration beyond improvements on language-agnostic benchmarks alone.

5 Conclusion

In this study, we show that the pre-training datasets of open-weight LLMs contain a disproportionately large amount of English text compared with the language distribution of the large-scale Web crawled corpus, CommonCrawl. In addition, we indicate that continual pre-training is approximately as cost-efficient as training from scratch, even for improving cultural understanding in a target language. We hope that our findings help advance LLM development beyond English-centric paradigms.

Limitations

Language Distribution Estimation To estimate the language distribution of pre-training corpora, we analyze sequences generated by each LLM. To assess the reliability of this approach, we compare the estimated language distribution with the reported distribution for Llama 2 (Touvron et al., 2023b), which is one of the few studies that disclose the language distribution of their training corpus. We observe a significantly high correlation between the two distributions. In addition, as described in Appendix A, we also investigate the correlation for LLM-jp and OLMo models and observe similarly strong correlations. We also confirm that our estimation is stable across different random seeds. However, we acknowledge that it would be desirable to validate this correlation for other models if such information were available.

Furthermore, we also investigate fine-tuned versions of Llama 2 7B and 13B. Their Kendall rank correlation coefficients are 0.62 and 0.60, respectively, which are lower than those of the corresponding pre-training models. Therefore, we focus on pre-training models when estimating language distributions.

Target Language In this study, we focus on Japanese as the target language for analyzing continual pre-training, because a wide range of continual pre-training models and cultural understanding benchmark datasets are available for Japanese. To the best of our knowledge, no other language currently offers a comparable combination of both resources for a systematic analysis.

Since the procedure of continual pre-training is not language-specific, we believe that the findings obtained from Japanese continual pre-training models can be applied to models for other languages. This is analogous to how neural scaling laws, which were primarily established through experiments on English corpora (Kaplan et al., 2020; Hoffmann et al., 2022), are widely accepted as applicable across languages. However, we acknowledge that investigating continual pre-training models for additional languages would further strengthen the generality of our findings.

Configuration of Continual Pre-training In this study, our analysis is limited to models with up to 32B parameters. Because pre-training LLMs at larger scale requires substantial computational resources, even when using continual pre-training, publicly available LLMs at substantially larger scales are limited. Although we anticipate similar trends for larger models, verifying whether Figure 2 (c) generalizes to larger parameter scales remains an important direction for future work.

We note that the R^2 values of the regression lines in Figure 2 (a) and (c) are 0.60 and 0.57, respectively. Since our analysis relies on diverse models independently developed rather than a controlled training strategy and uses several datasets for evaluation, such variation is expected (Ivgi et al., 2022).

In addition, some recent LLMs adopt Mixture-of-Experts (MoE) architecture (Shazeer et al., 2017) but we focus on dense architecture in this study. While it is an important direction to investigate their properties such as neural scaling laws (Liew et al., 2025), MoE models have a large number of total parameters, and thus, they require substantial computational resources for training, similar to the discussion on model size above. As a result, publicly available MoE models that have undergone continual pre-training are scarce, and we leave this investigation for future work.

We also note that we exclude analysis of post-training such as supervised fine-tuning (SFT), as post-training is complementary to pre-training and solely relying on post-training is insufficient to adapt LLMs to a target language. In fact, prior work has shown that attempting to introduce new linguistic or cultural knowledge through SFT tends to increase hallucinations rather than improve performance (Gekhman et al., 2024). Therefore, post-training is outside the scope of our analysis.

Cultural Understanding Benchmarks Recent studies have introduced benchmarks for each language (Isbarov et al., 2025; Koto et al., 2024; Li et al., 2024; Koto et al., 2023; Togmanov et al., 2025; Yin et al., 2024), highlighting the importance of evaluating LLMs beyond English-centric perspectives. Our study complements these efforts by providing quantitative evidence of the language distribution in training data, offering an empirical basis for systematically discussing the efficacy of LLMs across languages, rather than investigating internal representations (Zhong et al., 2025). Our analysis is limited to text-based modality, and extending this investigation to multimodal models, such as constructing benchmarks for VLMs (Vayani et al., 2024; Mogrovejo et al., 2024), is beyond the scope of this study.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1877–1901.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling (COLM)*.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7765–7784.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15789–15809.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, and 3 others. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 30016–30030.

Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, Kavsar Huseynova, Dmitry Gaynullin, Anar Rzayev, Osman Tursun, Aizirek Turdubaeva, Ilshat Saetov, Rinat Kharisov, Saule Belginova, Ariana Kenbayeva, Amina Alisheva, Abdullatif Köksal, Samir Rustamov, and Duygu Ataman. 2025. TUMLU: A unified and native language understanding benchmark for Turkic languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 22816–22838.

Maor Ivgi, Yair Carmon, and Jonathan Berant. 2022. Scaling laws under the microscope: Predicting transformer performance from small scale experiments. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7354–7371.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Maurice Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12359–12374.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive

multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285.

Seng Pei Liew, Takuya Kato, and Sho Takase. 2025. Scaling laws for upcycling mixture-of-experts language models. In *Forty-second International Conference on Machine Learning (ICML)*.

LLM-jp. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms.

David Orlando Romero Mogrovejo, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Villa Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Jouitteau, David LE MEUR, Joan Nwatu, Ganzorig Batnasan, and 57 others. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. 2025. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240.

Satoshi Sekine. 2003. Development of a question answering system focused on an encyclopedia (in japanese). In *9th Annual Meeting of the Association for Natural Language Processing*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 18761–18799.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15725–15788.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report.

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, and 89 others. 2024. Gemma: Open models based on gemini research and technology.

Mukhammed Togmanov, Nurdaulet Mukhituly, Diana Turmakhan, Jonibek Mansurov, Maiya Goloburda, Akhmed Sakip, Zhuohan Xie, Yuxia Wang, Bekassyl Syzdykov, Nurkhan Laiyk, Alham Fikri Aji, Ekaterina Kochmar, Preslav Nakov, and Fajri Koto. 2025. KazMMLU: Evaluating language models on Kazakh, Russian, and regional knowledge of Kazakhstan. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14403–14416.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard

Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 95266–95290.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2025. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1513–1532.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect LLMs? a cross-lingual study on the influence of prompt politeness on LLM performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 9–35.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer.

Chengzhi Zhong, Qianying Liu, Fei Cheng, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2025. What language do non-English-centric large language models think in? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26333–26346.

A Additional Validation of Language Distribution Estimation

To examine the reasonableness of our language distribution estimation method based on sequences generated by each LLM, we investigate the Kendall rank correlation coefficient for LLM-jp and OLMo models (LLM-jp, 2024; Groeneveld et al., 2024), in addition to Llama 2. Unlike Llama 2, these models do not disclose the language distribution of their pre-training corpus but they do provide the corpus itself. Therefore, we sample data from the released corpus and compute the correlation coefficient between the language distribution of the sampled corpus and that of the sequences generated by those models.

We focus on LLM-jp-3-1.8B, LLM-jp-3-13B and OLMo-1B. For LLM-jp models, LLM-jp Corpus v3[6] was used for pre-training. Because this corpus is divided into multiple domains, we sample documents from each domain in proportion to the reported number of tokens in that domain. We then concatenate the sampled documents and further sample 100,000 documents from the concatenated set to estimate the language distribution. For OLMo, Dolma was used for pre-training (Soldaini et al., 2024), and thus, we sample 100,000 documents from it. The resulting Kendall rank correlation coefficients are 0.61 for LLM-jp-3-1.8B, 0.63 for LLM-jp-3-13B, and 0.71 for OLMo-1B. The slightly lower coefficients for LLM-jp models may be attributed to the approximate nature of our corpus sampling, as the pre-training corpus is divided into multiple domains and we estimate the overall distribution through proportional sampling from each domain. Since these values indicate strong correlations, we consider our estimation method to be reasonable.

In addition, we examine the robustness of our estimation method with respect to random seeds dur-

[6] https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3

Hyper-parameter	Value
Temperature	1.0
Top_p	1.0
Top_k	-1
max_tokens	1,000
stop	None
ignore_eos	True
repetition_penalty	1.0

Table 3: Hyper-parameters used in sequence generation.

ing inference. We generate sequences from Llama 2 13B using two different random seeds. The resulting Kendall rank correlation coefficients are 0.70 and 0.71, which are similarly high. This result indicates that our estimation is stable across different random seeds.

B Hyper-parameters during Sequence Generation

To generate sequences randomly in Section 2, we use vLLM (Kwon et al., 2023) with its default hyper-parameters. Table 3 lists the corresponding values.

C Details of Language Distribution

Table 4 shows the estimated language distribution from sampled CommonCrawl corpus and sequences generated by each LLM except for Llama 2. For Llama 2, we used the language distribution reported in the original paper (Touvron et al., 2023b). For this estimation, we used Llama-3.1-8B[7], Qwen2.5-3B[8], Qwen3-4B-Base[9], Gemma-3-4B-pt[10], LLM-jp-3-1.8B[11], and Sarashina2.2-3B[12]. We note that LLM-jp-3-1.8B is the pre-training model of LLM-jp-3.1-1.8B-Instruct4.

D Amount of Japanese Training Data

Table 5 shows the estimated number of Japanese tokens trained by base model and the number of Japanese tokens during continual pre-training for each model.

E MGSM Performance with Thinking in Each Language

Table 6 shows the few-shot performance of each LLM on the MGSM test set when we prompt the models to think in English, Chinese, or Japanese. Following Shi et al. (2023), we provide each few-shot example as a pair consisting of a question in the original language and an answer in the target language. For example, in the case of Russian with thinking in English, the few-shot example consists of a Russian question and its corresponding English answer.

Table 6 indicates that Llama3.1-8B-Instruct and Gemma-3-27B-it achieve the best performance when thinking in English, consistent with the findings reported by Shi et al. (2023). In contrast, the performance of Qwen2.5-32B-Instruct and Qwen3-32B when thinking in Chinese is comparable to or better than their performance when thinking in English. In addition, LLM-jp-3.1-13B-Instruct4 achieves the best performance when thinking in Japanese. These results indicate that an LLM can achieve superior performance when thinking in a language that is frequent in its pre-training corpus. In other words, strong performance when prompting an LLM to think in English suggests that English occupies a large proportion of its pre-training corpus.

[7] huggingface.co/meta-llama/Llama-3.1-8B
[8] huggingface.co/Qwen/Qwen2.5-3B
[9] huggingface.co/Qwen/Qwen3-4B-Base
[10] huggingface.co/google/gemma-3-4b-pt
[11] huggingface.co/llm-jp/llm-jp-3-1.8b
[12] huggingface.co/sbintuitions/sarashina2.2-3b

Model	En	Ru	De	Fr	Es	Ja	Zh	It	Pt	Nl	others
CommonCrawl	43.0	9.8	6.4	5.1	5.1	4.2	4.1	2.8	2.3	1.9	15.3
Llama 2	89.7	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	9.1
Llama 3.1	92.3	0.4	0.3	0.2	0.3	1.8	2.6	0.1	0.1	0.1	1.7
Qwen 2.5	77.7	0.8	0.6	0.8	0.5	0.8	16.2	0.3	0.3	0.1	1.9
Qwen 3	73.4	0.6	0.3	0.3	0.6	0.4	22.8	0.1	0.5	0.0	1.0
Gemma 3	74.9	1.3	1.4	1.6	6.2	1.1	3.1	0.9	1.2	0.8	7.5
LLM-jp-3	63.4	0.0	0.0	0.0	0.0	36.4	0.1	0.0	0.0	0.0	0.0
Sarashina2.2	73.7	0.5	0.3	0.4	0.4	21.3	2.0	0.1	0.2	0.0	1.1

Table 4: Details of language distribution for CommonCrawl and each LLM. We represent the percentage of each language.

Model	Estimated Japanese tokens in pre-training	Japanese tokens in continual pre-training
Llama3.1-Swallow-8B-Instruct-v0.5[13]	274B	210B
Qwen2.5-Bakeneko-32B-Instruct-v2[14]	144B	18B
ABEJA-Qwen2.5-32B-Japanese-v1.0[15]	144B	70B
ELYZA-Shortcut-1.0-Qwen-32B[16]	144B	20B
LLMs with pre-training data containing a large amount of Japanese text		
Sarashina2.2-3B-Instruct-v0.1[17]	2339B	N/A
LLM-jp-3.1-13B-Instruct4[18]	911B	N/A

Table 5: The estimated number of Japanese tokens in base model training and the number of Japanese tokens during continual pre-training.

Model	En	Ru	De	Fr	Es	Ja	Zh	Th	Te	Bn	Sw	Average
Thinking in English												
Qwen2.5-32B-Instruct	87.2	79.6	77.2	74.8	83.6	71.2	76.8	77.6	61.6	80.8	52.0	74.8
Qwen3-32B	95.6	90.0	88.4	88.4	90.4	87.2	87.6	89.6	82.8	85.6	79.2	87.7
Llama3.1-8B-Instruct	80.0	68.4	72.0	71.2	75.6	58.8	66.4	67.6	54.0	58.4	56.0	66.2
Gemma-3-27B-it	96.0	91.2	88.0	85.6	92.0	81.2	89.2	89.6	84.8	88.0	83.6	88.1
Llama3.1-Swallow-8B-Instruct-v0.5	76.8	64.8	63.6	68.4	68.8	66.4	62.8	50.4	36.0	45.2	41.6	58.6
Qwen2.5-Bakeneko-32B-Instruct-v2	94.4	90.8	90.4	85.2	87.2	85.6	88.4	89.6	64.0	83.6	60.8	83.6
ABEJA-Qwen2.5-32B-Japanese-v1.0	90.8	89.6	84.4	83.2	86.4	83.2	84.0	86.4	59.2	81.2	51.6	80.0
ELYZA-Shortcut-1.0-Qwen-32B	61.6	63.6	50.4	60.8	86.0	80.0	63.6	66.0	55.2	68.0	44.0	63.6
Sarashina2.2-3B-Instruct-v0.1	76.0	56.0	62.8	62.0	66.8	68.4	60.4	22.4	7.2	17.6	8.4	46.2
LLM-jp-3.1-13B-Instruct4	57.2	55.6	52.8	54.0	57.6	64.0	50.8	28.4	19.2	30.0	11.6	43.7
Thinking in Chinese												
Qwen2.5-32B-Instruct	85.6	86.4	83.6	82.0	86.0	76.4	80.8	86.8	54.0	76.0	62.0	78.1
Qwen3-32B	94.4	88.4	88.4	84.8	91.6	80.8	89.6	88.4	85.2	85.6	73.2	86.4
Llama3.1-8B-Instruct	74.8	66.4	65.2	53.6	70.4	54.4	60.0	17.6	38.0	48.0	55.2	54.9
Gemma-3-27B-it	92.0	87.2	85.2	84.4	90.0	78.4	84.8	86.4	80.4	82.0	82.4	84.8
Llama3.1-Swallow-8B-Instruct-v0.5	62.4	50.8	54.8	52.0	62.4	53.6	55.6	40.8	26.4	26.8	34.0	47.2
Qwen2.5-Bakeneko-32B-Instruct-v2	94.0	90.8	89.6	84.4	90.8	83.2	87.2	88.4	62.4	86.0	62.0	83.5
ABEJA-Qwen2.5-32B-Japanese-v1.0	88.0	85.2	86.0	86.4	86.8	80.4	84.0	57.2	62.4	44.0	55.2	74.1
ELYZA-Shortcut-1.0-Qwen-32B	73.6	70.8	60.8	69.6	84.8	79.6	82.0	68.0	56.4	63.2	59.2	69.8
Sarashina2.2-3B-Instruct-v0.1	65.6	47.6	46.4	51.6	52.8	60.8	54.0	24.8	5.6	13.6	8.0	39.2
LLM-jp-3.1-13B-Instruct4	61.2	52.0	55.2	53.2	49.2	61.6	16.4	26.0	20.4	22.8	6.8	38.6
Thinking in Japanese												
Qwen2.5-32B-Instruct	78.4	74.0	75.2	62.4	79.2	76.0	76.4	58.4	57.6	64.8	54.8	68.8
Qwen3-32B	96.4	85.2	90.0	86.8	92.0	77.6	87.2	74.8	83.2	84.0	72.4	84.5
Llama3.1-8B-Instruct	57.6	62.0	58.0	43.2	58.4	53.2	50.8	41.2	42.0	44.0	48.4	50.8
Gemma-3-27B-it	90.8	85.6	85.2	81.2	86.0	75.6	82.8	83.6	79.6	81.6	81.2	83.0
Llama3.1-Swallow-8B-Instruct-v0.5	68.8	62.4	63.2	55.2	61.2	60.0	51.2	43.6	33.2	46.4	38.0	53.0
Qwen2.5-Bakeneko-32B-Instruct-v2	91.2	86.4	84.4	83.2	86.4	84.0	82.4	87.6	60.8	82.8	60.4	80.9
ABEJA-Qwen2.5-32B-Japanese-v1.0	39.2	60.4	67.2	45.6	71.2	32.8	50.4	37.2	28.4	34.4	44.4	46.5
ELYZA-Shortcut-1.0-Qwen-32B	74.4	59.2	69.2	63.2	75.2	69.2	34.4	55.6	54.4	64.4	45.2	60.4
Sarashina2.2-3B-Instruct-v0.1	66.0	55.2	50.0	57.2	54.8	60.4	58.8	21.2	8.8	15.2	10.8	41.7
LLM-jp-3.1-13B-Instruct4	77.6	56.4	57.6	64.4	62.8	66.0	56.0	35.2	22.8	35.2	7.6	49.2

Table 6: 6-shot performance with thinking in each language on the MGSM test set.