MERaLiON-TextLLM: Cross-Lingual Understanding of Large Language Models in Chinese, Indonesian, Malay, and Singlish

MERaLiON Team

Xin Huang, Tarun Kumar Vangani, Minh Duc Pham, Xunlong Zou
Bin Wang, Zhengyuan Liu, Ai Ti Aw*

Institute for Infocomm Research (I^2R), A*STAR, Singapore

Abstract

Multilingual large language models (MLLMs) have shown impressive capabilities across a variety of languages. However, efficacy can differ greatly between different language families, especially for those with limited linguistic resources. This report presents **MERaLiON-TextLLM**, a series of open-source language models specifically tailored to improve understanding and generation in Chinese, Indonesian, Malay, and Singlish. The initial released model is built on Llama-3-8B-Base and refined through a meticulously crafted process of continued pre-training and weight merging. Our approach achieves performance improvements across benchmarks in these languages, exceeding the capabilities of the official Llama-3 models. We provide the model checkpoints as a resource to support further research and development in cross-lingual language understanding.

1 Introduction

Our first released model is named MERaLiON-LLaMA-3-8B-Instruct. MERaLiON-LLaMA-3-8B-Instruct has been extensively pre-trained on English, Chinese, and Indonesian, building upon LLaMA-3.1-8B-Base Dubey et al. [2024], with a primary emphasis on enhancing its understanding and generation capabilities in Chinese and Indonesian. Leveraging innovative corpus mixing strategies tailored to multilingual regional datasets, we carefully diversified the training materials using domain classification, hyperparameter optimization, and strategic replay techniques. These methods are specifically designed to prevent catastrophic forgetting, enabling the model to retain previously acquired knowledge while significantly improving its ability to produce high-quality, contextually accurate responses in Southeast Asian languages.

2 Pre-Training

MERaLiON-LLaMA-3-8B-Instruct training was conducted using the MaxText AI-Hypercomputer [2024] platform, utilizing both NVIDIA H100 GPUs and TPU v4-128 chips. Specifically, we utilized 64 H100 GPUs, achieving approximately 400 TFLOPS per GPU, and TPU v4-128 configurations, attaining around 168 TFLOPS per TPU chip.

---

*Please cite this report as authored by MERaLiON team.

Corresponding: {huangx2, vangani_tarun_kumar}@i2r.a-star.edu.sg

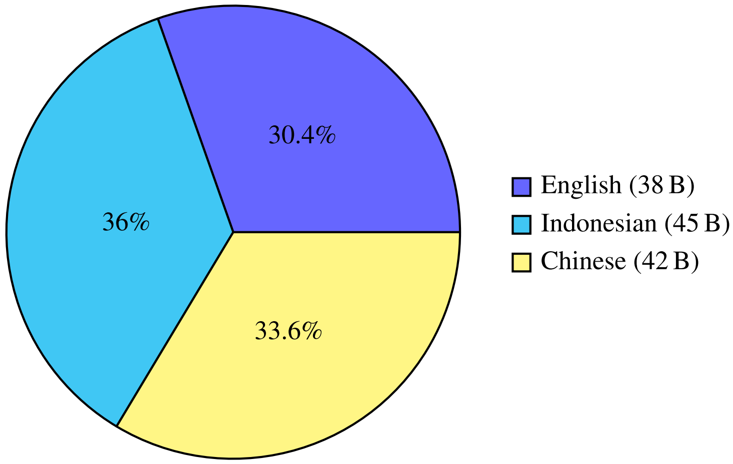

Figure 1: Distribution of Tokens

These performance metrics were achieved through optimized sharding, checkpoint strategies, and the selection of optimal batch sizes, to ensure efficient and effective model training. We conducted multiple runs to achieve optimal performance and sampling strategies.

The pre-training data was allocated across three primary languages: English, Indonesian, and Chinese. Specifically, 38 billion tokens were allocated to English, 45 billion to Indonesian, and 42 billion to Chinese. This allocation is designed to ensure robust performance across all three languages, catering to a variety of linguistic and cultural contexts.

Figure 1 illustrates the token distribution in a pie chart format for clear visualization. English accounted for 30.4% of the total tokens, Indonesian for 36.0%, and Chinese for 33.6%. The allocation highlights the model's multilingual focus, ensuring no language dominates the training set, thereby fostering balanced performance.

3 Instruction Tuning and Model Merging

We present updated experimental results for our latest variant of the Southeast Asian language model, focusing on three representative languages: English, Chinese, and Indonesian. Our work began with the construction of a multilingual instruction-response corpus containing approximately 3M pairs. Using this dataset, we performed instruction tuning on the Llama-3.1-8B base model to evaluate potential performance gains. However, our initial experiments revealed that this approach alone did not surpass the robust baseline performance of the Llama-3.1-8B-Instruct model. Specifically, the instruction-tuned model underperformed the original Llama-3.1-8B-Instruct across the three target languages in the Cross-MMLU and Cross-LogiQA datasets Wang et al. [2024].

To address these limitations, we explored model merging techniques aimed at enhancing instruction-following capabilities without requiring extensive additional tuning. By carefully merging the weight sets of Llama-3.1-8B-Base and Llama-3.1-8B-Instruct, we leveraged the strong instruction-following performance of the instructed variant while integrating domain-specific knowledge from our custom dataset. As illustrated in Table 1, the resulting "MERaLiON-Llama-3.1-8B-Instruct" model

Table 1: Performance comparison of the merged model (MERaLiON-LLaMA-3.1-8B-Instruct) against baseline and instruction-tuned models on Cross-MMLU and Cross-LogiQA benchmarks.

Model	Cross-MMLU			Cross-LogiQA		
	English	Chinese	Indonesian	English	Chinese	Indonesian
MERaLiON-LLaMA-3-8B-Instruct	**0.85**	**0.69**	**0.71**	**0.591**	0.528	**0.494**
Meta-Llama-3.1-8B-Instruct	0.82	0.63	0.66	0.585	**0.585**	0.455
Instruction-Tuned Llama-3.1-8B	0.73	0.60	0.58	0.563	0.523	0.466

consistently outperformed both the baseline Llama-3.1-8B-Instruct and our standalone instruction-tuned variant across English, Chinese, and Indonesian test sets.

These results underscore the effectiveness of merging pretrained instructed weights with domain-specific models to achieve enhanced instruction-following capabilities. The MERaLiON-Llama-3.1-8B-Instruct model demonstrates improved performance in all evaluated languages, confirming the viability of this approach. This hybrid methodology offers a promising avenue for the development of resource-efficient, multilingual, and domain-adaptive instruction-following models without necessitating extensive iterative tuning on large-scale, specialized corpora.

4 Benchmarks and Evaluation

Table 2: Cross-MMLU

Model Series	Model	English	Chinese	Indonesian	Malay	Avg
LLaMA3	**MERaLiON-LLaMA-3-8B-Instruct**	**0.847**	**0.693**	**0.713**	0.613	**0.717**
	Meta-Llama-3.1-8B-Instruct	0.82	0.633	0.66	**0.647**	0.690
	Llama3-8B-SEA-LION-v2.1-Instruct Singapore [2024]	0.753	0.667	0.693	0.64	0.688
	Meta-Llama-3-8B-Instruct	0.767	0.653	0.573	0.573	0.642
Others	**GPT4o-0513**	**0.927**	**0.887**	**0.88**	**0.907**	**0.900**
	Gemma-2-9B-IT Team et al. [2024]	0.84	0.793	0.78	0.747	0.790
	Gemma2-9B-SEA-LION-v3-Instruct Singapore [2024]	0.847	0.787	0.793	0.733	0.790
	Qwen2.5-7B-Instruct Yang et al. [2024]	0.847	0.84	0.753	0.713	0.788
	SeaLLMs-v3-7B-Chat Zhang et al. [2024]	0.833	0.727	0.74	0.687	0.747

Table 3: Cross-LogiQA

Model Series	Model	English	Chinese	Indonesian	Malay	Avg
LLaMA3	**Meta-Llama-3.1-8B-Instruct**	0.585	**0.585**	0.455	**0.523**	**0.537**
	MERaLiON-LLaMA-3-8B-Instruct	0.591	0.528	**0.494**	0.489	0.526
	Meta-Llama-3-8B-Instruct	**0.602**	0.523	0.438	0.483	0.512
	Llama3-8B-SEA-LION-v2.1-Instruct	0.528	0.517	0.403	0.443	0.473
Others	**Qwen2.5-7B-Instruct**	**0.693**	**0.71**	**0.631**	0.534	**0.642**
	Gemma-2-9B-IT	0.659	0.636	0.585	**0.602**	0.621
	Gemma2-9B-SEA-LION-v3-Instruct	0.636	0.642	0.557	0.551	0.597
	SeaLLMs-v3-7B-Chat	0.568	0.585	0.494	0.517	0.541

We conducted comprehensive evaluations using several benchmarks to assess the multilingual and instruction-following performance of MERaLiON. Key benchmarks include:

- **Cross-MMLU** Wang et al. [2024] : A subset of the MMLU dataset translated into multiple languages, including English, Chinese, Indonesian, and Malay. It aims to measure the model's ability to handle general knowledge queries across these diverse linguistic contexts.
- **Cross-LogiQA** Wang et al. [2024] : Building on the original LogiQA dataset, Cross-LogiQA introduces multilingual versions of logical reasoning tasks. It provides parallel question sets in English, Chinese, and Indonesian that are designed to maintain logical equivalence across these languages.
- **IndoMMLU** Koto et al. [2023] : A benchmark designed to assess general knowledge and language understanding of large language models for Indonesian, particularly for domains such as law, medicine, and social science Koto et al. [2023].
- **CN-Eval:** A selected subset of C-Eval Huang et al. [2023] and CMMLU Li et al. [2024] curated to specifically assess a model's knowledge about the Chinese language, culture and socio-political context. This subset provides a focused metric for assessing LLM knowledge of China.

4.1 Results

The results of our evaluation highlight several strengths of MERaLiON:

- On Cross-MMLU and Cross-LogiQA, as shown in Table 2 and Table 3, MERaLiON outperforms the baseline Meta-Llama-3.1-8B-Instruct model on Cross-MMLU, demonstrating

Table 4: IndoMMLU

Model Series	Model	Accuracy
LLaMA Series	**MERaLiON-LLaMA-3-8B-Instruct**	**0.576**
	Llama3-8B-SEA-LION-v2.1-Instruct	0.560
	Meta-Llama-3.1-8B-Instruct	0.548
	Meta-Llama-3-8B-Instruct	0.521
Others	**GPT4o-0513**	**0.760**
	Gemma2-9B-SEA-LION-v3-Instruct	0.626
	Gemma-2-9B-IT	0.621
	Qwen2.5-7B-Instruct	0.582
	SeaLLMs-v3-7B-Chat	0.541

Table 5: CN-Eval

Model Series	Model	Accuracy
LLaMA Series	**MERaLiON-LLaMA-3-8B-Instruct**	**0.514**
	Llama3-8B-SEA-LION-v2.1-Instruct	0.505
	Llama3-8B-SEA-LION-v2-Instruct	0.495
	Meta-Llama-3-8B-Instruct	0.467
	Meta-Llama-3.1-8B-Instruct	0.457
Others	**Qwen2-7B-Instruct**	**0.829**
	GPT4o-0513	0.810
	Qwen2.5-7B-Instruct	0.800
	Gemma2-9B-SEA-LION-v3-Instruct	0.590
	Gemma-2-9B-IT	0.581

superior general knowledge coverage in Chinese and Indonesian, while maintaining a strong, balanced performance in English. Similarly, MERaLiON improves logical reasoning in Indonesian, indicating that its continued pre-training and model merging strategies effectively enhance both factual knowledge and reasoning capabilities.

- On IndoMMLU, as reported in Table 4, MERaLiON significantly outperforms the baseline Llama-3.1-8B-Instruct model, achieving 0.576 accuracy versus 0.548, highlighting its improved understanding of Indonesian language and domain-specific nuances.
- For CN-Eval, MERaLiON achieves 0.514 accuracy compared to 0.457 for Llama-3.1-8B-Instruct, as shown in Table 5. This result demonstrates MERaLiON's efficacy in retaining and enhancing knowledge related to China.

These benchmarks demonstrate the superior capability of MERaLiON-TextLLM to handle multilingual tasks and deliver improved cross-lingual understanding. Our training approach ensures the model is well-suited for diverse language tasks.

5 Conclusion and Future Work

The MERaLiON-LLaMA-3-8B-Instruct model improves multilingual NLP for Indonesian and Chinese, by addressing cross-lingual challenges with effective pretraining and model merging.

Evaluation results demonstrate MERaLiON-TextLLM's strengths, such as better reasoning and question-answering on Cross-MMLU and Cross-LogiQA through balanced multilingual training, strong accuracy on IndoMMLU and CN-Eval from effective dataset preparation, enhanced instruction-following via weight merging with minimal tuning, and resource efficiency through optimized TPU/GPU strategies for scalable training.

Future directions include:

- Expanding to more underrepresented languages like Tagalog, Thai, and Vietnamese.
- Developing enhanced evaluation frameworks with human judgment and multimodal input.
- Applying the model to tasks such as translation, summarization, and content analytics.

MERaLiON demonstrates the potential for resource-efficient multilingual NLP, setting the foundation for broader linguistic coverage and consistent multilingual reasoning.

Acknowledgement

We extend our sincere gratitude to Nattadaporn Lertcheva, Xi Wang, Kui Wu, Yang Ding, Nabilah Binte Md Johan for their invaluable contributions to data, insightful discussions, and future work explorations.

This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority, Singapore.

The resources and platforms provided by Singapore NSCC Aspire2A+ and The TPU Research Cloud. We thank all contributors and collaborators who have made this effort possible.

References

AI-Hypercomputer. Maxtext: A simple, performant, and scalable jax llm, 2024. URL `https://github.com/AI-Hypercomputer/maxtext`. Accessed: 2024-12-16.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Y. Huang, Y. Bai, Z. Zhu, J. Zhang, J. Zhang, T. Su, J. Liu, C. Lv, Y. Zhang, J. Lei, Y. Fu, M. Sun, and J. He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models, 2023. URL `https://arxiv.org/abs/2305.08322`.

F. Koto, N. Aisyah, H. Li, and T. Baldwin. Large language models only pass primary school exams in indonesia: A comprehensive test on indommlu, 2023. URL `https://arxiv.org/abs/2310.04928`.

H. Li, Y. Zhang, F. Koto, Y. Yang, H. Zhao, Y. Gong, N. Duan, and T. Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese, 2024. URL `https://arxiv.org/abs/2306.09212`.

A. Singapore. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. `https://github.com/aisingapore/sealion`, 2024.

G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, J. Ferret, P. Liu, P. Tafti, A. Friesen, M. Casbon, S. Ramos, R. Kumar, C. L. Lan, S. Jerome, A. Tsitsulin, N. Vieillard, P. Stanczyk, S. Girgin, N. Momchev, M. Hoffman, S. Thakoor, J.-B. Grill, B. Neyshabur, O. Bachem, A. Walton, A. Severyn, A. Parrish, A. Ahmad, A. Hutchison, A. Abdagic, A. Carl, A. Shen, A. Brock, A. Coenen, A. Laforge, A. Paterson, B. Bastian, B. Piot, B. Wu, B. Royal, C. Chen, C. Kumar, C. Perry, C. Welty, C. A. Choquette-Choo, D. Sinopalnikov, D. Weinberger, D. Vijaykumar, D. Rogozińska, D. Herbison, E. Bandy, E. Wang, E. Noland, E. Moreira, E. Senter, E. Eltyshev, F. Visin, G. Rasskin, G. Wei, G. Cameron, G. Martins, H. Hashemi, H. Klimczak-Plucińska, H. Batra, H. Dhand, I. Nardini, J. Mein, J. Zhou, J. Svensson, J. Stanway, J. Chan, J. P. Zhou, J. Carrasqueira, J. Iljazi, J. Becker, J. Fernandez, J. van Amersfoort, J. Gordon, J. Lipschultz, J. Newlan, J. yeong Ji, K. Mohamed, K. Badola, K. Black, K. Millican, K. McDonell, K. Nguyen, K. Sodhia, K. Greene, L. L. Sjoesund, L. Usui, L. Sifre, L. Heuermann, L. Lago, L. McNealus, L. B. Soares, L. Kilpatrick, L. Dixon, L. Martins, M. Reid, M. Singh, M. Iverson, M. Görner, M. Velloso, M. Wirth, M. Davidow, M. Miller, M. Rahtz, M. Watson, M. Risdal, M. Kazemi, M. Moynihan, M. Zhang, M. Kahng, M. Park, M. Rahman, M. Khatwani, N. Dao, N. Bardoliwalla, N. Devanathan, N. Dumai, N. Chauhan, O. Wahltinez, P. Botarda, P. Barnes, P. Barham, P. Michel, P. Jin, P. Georgiev, P. Culliton, P. Kuppala, R. Comanescu, R. Merhej, R. Jana, R. A. Rokni, R. Agarwal, R. Mullins, S. Saadat, S. M. Carthy, S. Cogan, S. Perrin, S. M. R. Arnold, S. Krause, S. Dai, S. Garg, S. Sheth, S. Ronstrom, S. Chan, T. Jordan, T. Yu, T. Eccles, T. Hennigan, T. Kocisky, T. Doshi, V. Jain, V. Yadav, V. Meshram,

V. Dharmadhikari, W. Barkley, W. Wei, W. Ye, W. Han, W. Kwon, X. Xu, Z. Shen, Z. Gong, Z. Wei, V. Cotruta, P. Kirk, A. Rao, M. Giang, L. Peran, T. Warkentin, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, D. Sculley, J. Banks, A. Dragan, S. Petrov, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, S. Borgeaud, N. Fiedel, A. Joulin, K. Kenealy, R. Dadashi, and A. Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL `https://arxiv.org/abs/2408.00118`.

B. Wang, Z. Liu, X. Huang, F. Jiao, Y. Ding, A. Aw, and N. Chen. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.22. URL `https://aclanthology.org/2024.naacl-long.22`.

A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan. Qwen2 technical report, 2024. URL `https://arxiv.org/abs/2407.10671`.

W. Zhang, H. P. Chan, Y. Zhao, M. Aljunied, J. Wang, C. Liu, Y. Deng, Z. Hu, W. Xu, Y. K. Chia, X. Li, and L. Bing. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages, 2024. URL `https://arxiv.org/abs/2407.19672`.