SeaLLMs 3: Open Foundation and Chat Multilingual Large Language Models for Southeast Asian Languages

Wenxuan Zhang*, **Hou Pong Chan***, **Yiran Zhao***, **Mahani Aljunied***, **Jianyu Wang***, **Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, Lidong Bing**†

DAMO Academy, Alibaba Group
Project page: `https://seallms.github.io/`

Abstract

Large Language Models (LLMs) have shown remarkable abilities across various tasks, yet their development has predominantly centered on high-resource languages like English and Chinese, leaving low-resource languages underserved. To address this disparity, we present SeaLLMs 3, the latest iteration of the SeaLLMs model family, tailored for Southeast Asian languages. This region, characterized by its rich linguistic diversity, has lacked adequate language technology support. SeaLLMs 3 aims to bridge this gap by covering a comprehensive range of languages spoken in this region, including English, Chinese, Indonesian, Vietnamese, Thai, Tagalog, Malay, Burmese, Khmer, Lao, Tamil, and Javanese. Leveraging efficient language enhancement techniques and a specially constructed instruction tuning dataset, SeaLLMs 3 significantly reduces training costs while maintaining high performance and versatility. Our model excels in tasks such as world knowledge, mathematical reasoning, translation, and instruction following, achieving state-of-the-art performance among similarly sized models. Additionally, we prioritized safety and reliability by addressing both general and culture-specific considerations and incorporated mechanisms to reduce hallucinations. This work underscores the importance of inclusive AI, showing that advanced LLM capabilities can benefit underserved linguistic and cultural communities.

1 Introduction

Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023) and Gemini (Anil et al., 2023) have demonstrated remarkable capabilities across a wide array of tasks, ranging from natural language understanding and generation to more specialized domain applications (Zhao et al., 2023). These models have proven valuable, offering substantial benefits to the global community, especially through the proliferation of open-source LLMs such as Llama (Touvron et al., 2023a,b), Mistral (Jiang et al., 2023), Qwen (Bai et al., 2023), and Gemma (Mesnard et al., 2024). However, the majority of these efforts have been concentrated on high-resource languages such as English and Chinese, or well-developed regions like Europe (Zhang et al., 2023; Ahuja et al., 2023). Consequently, the development of LLMs tailored for low-resource languages or underdeveloped regions has been significantly overlooked, resulting in a lack of inclusivity and equitable distribution of AI advancements across diverse linguistic and cultural communities (Qin et al., 2024; Huang et al., 2024).

To bridge this gap, we introduced the SeaLLMs model (Nguyen et al., 2023c), specifically designed LLMs for Southeast Asian languages. Southeast Asia (SEA) is a region with a rich diversity of languages spoken by millions of people, yet it suffers from a significant lack of language technology support (Aji et al., 2022). The initiative of SeaLLMs thus aims to make the benefits of LLMs accessible to speakers of these languages, addressing their unique linguistic and cultural nuances. Following this endeavor, several other models have been dedicated to this region as well, such as SEA-LION (AI Singapore, 2023) and Sailor (Dou et al., 2024). However, these models often face significant limitations: they are typically released only as foundational or chat models, offer limited options in terms of model size, and cover a limited number of SEA languages. Moreover, the relatively scarce availability of language corpora further constrains the amount of training data available, hindering the development and performance of these models.

In this work, we introduce **SeaLLMs 3**, the latest iteration of the SeaLLMs model family. This version is designed to cover a more diverse array of Southeast Asian languages, including En-

* Equal contributions.
† Corresponding author: l.bing@alibaba-inc.com

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), pages 96–105

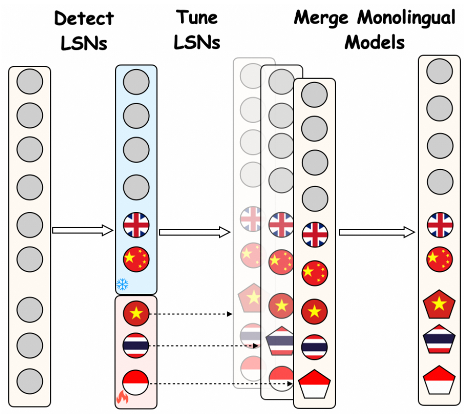

Figure 1: Illustration of Language-Specific Neuron Training. The snow and fire icons indicate frozen and non-frozen weights respectively.

glish, Chinese, Indonesian, Vietnamese, Thai, Tagalog, Malay, Burmese, Khmer, Lao, Tamil, and Javanese. Different from the conventional continue-pretraining paradigm (Zhao et al., 2024a; Nguyen et al., 2023c; Dou et al., 2024), we conduct efficient language enhancement by training language-specific neurons only based on a foundation model (Zhao et al., 2024b), significantly reducing the overall training cost. Moreover, such targeted training also ensures that the performance of high-resource languages can remain unaffected during the enhancement. Furthermore, SeaLLMs 3 is trained using a specially constructed instruction tuning dataset that encompasses a wide variety of task types and carefully balanced language distributions. This approach ensures that the model can handle the linguistic diversity of the Southeast Asian region while maintaining high performance and versatility across different applications. As a result, it achieves state-of-the-art performance among models with similar sizes, excelling across a diverse array of tasks such as world knowledge (Zhang et al., 2023), mathematical reasoning (Shi et al., 2023), translation (Costa-jussà et al., 2022), and instruction following.

In the meantime, we pay special attention to the model's reliability and trustworthiness during its development, which are often under-considered in multilingual settings (Deng et al., 2023). In particular, we address both general and culture-specific safety considerations to ensure the models provide contextually appropriate responses. The model is also specifically trained to be aware of its knowledge boundary and refuse what it does not know. This focus on safety and reliability has resulted in SeaLLMs 3 exhibiting reduced hallucination and delivering safe, coherent responses, especially in queries closely related to Southeast Asian culture.

We open-source both the foundational and chat models of SeaLLMs 3[1]. The foundational model could serve as a base for conducting instruction tuning tailored to specific application requirements. Meanwhile, the chat model has already undergone instruction tuning and is ready for direct use of handling a wide range of tasks.

2 Pre-training

2.1 Pre-training Data

Building on the efforts from previous versions of SeaLLMs (Nguyen et al., 2023c), we have incorporated corpora from a wider range of data sources to enhance diversity. Specifically, we have integrated fundamental knowledge from Wikipedia (Foundation) and textbooks (Ben Allal et al., 2024), journalistic materials such as CC-News (Crawl), web-based corpora from CulturaX (Nguyen et al., 2023a), and MADLAD-400 (Kudugunta et al., 2023). We have also improved the data processing pipeline including the language model filtering and duplicate removal to improve the data quality.

Additionally, we explore the utilization of model-synthetic data for training, which received much attention recently (Ben Allal et al., 2024). Starting with manual annotation of domain-specific knowledge points in SEA languages, we then employed stronger models to generate targeted tutorial-style content, thereby enhancing SeaLLMs 3 with enriched regional knowledge in a more explicit form.

2.2 Language-Specific Neuron Training

We built our model based on the Qwen2 model family (Yang et al., 2024) and further conducted language enhancement to augment its capability in SEA languages. This approach allows the model to quickly inherit foundational knowledge from Qwen, rather than learning it from scratch.

The most straightforward method for language enhancement is typically through continued pertaining (Zhao et al., 2024a), which we also used for previous versions of SeaLLMs. However, as discussed, the relatively scarce availability of language corpora limits the amount of training data, hindering the development and performance of these models. Furthermore, it is often observed that such direct continued pretraining can compromise the model's

[1]https://huggingface.co/collections/SeaLLMs/seallms-v3-668f3a52e1e6fbaad5752cdb

original capacity in high-resource languages like English and Chinese (Dou et al., 2024).

In this iteration, we adopt Language-Specific Neuron (LSN) training for efficient language enhancement, as shown in Figure 1. Recent studies have found that certain language-specific neurons in language models are responsible for processing specific languages. For instance, Zhao et al. (2024b) discovered that language-specific neurons comprise only about 0.1% of all parameters. Thus, the capabilities of a language can be enhanced by training its corresponding LSNs while preserving the multilingual abilities of other languages. To efficiently train SeaLLMs 3, we employ the parallel language-specific neuron detection method proposed by Zhao et al. (2024b). As shown in the left most part of Figure 1, this method allows us to identify the LSNs of SEA languages using language-specific training data, selected as a down-sampled subset of the corpora from the training data. We then specifically train these detected LSNs to develop multiple monolingual LLMs in SEA languages, which are subsequently merged to create a unified multilingual LLM for SEA languages. Additionally, to maintain proficiency in English and Chinese from the original foundation model, we detect their respective LSNs and exclude them from the entire pre-training process.

This method offers several advantages. First, it requires relatively less training data since the training is more targeted, which significantly reduces training costs. Second, because the training is targeted, we can ensure that the performance of high-resource languages from the original foundation model remains unaffected. LSNs operate independently and do not influence one another, avoiding the sacrifices seen with previous methods.

3 Supervised Fine-Tuning (SFT)

3.1 Supervised Fine-tuning Data

Most existing open-source supervised fine-tuning (SFT) datasets are predominantly in English (Wei et al., 2022; Taori et al., 2023), which presents a challenge for developing effective models for Southeast Asian languages. To address this, we employed various techniques to construct our SFT pool. For example, we selectively translate some high-quality English data to SEA languages with quality filtering, conduct self-instruction to automatically generate SFT data of certain types, and use various prompting strategies (Madaan et al.,

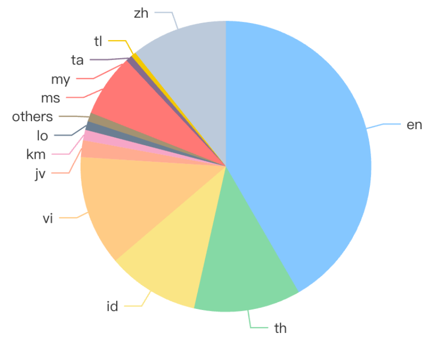

Figure 2: Language distribution of the SFT data

2023; Nguyen et al., 2023b). Following our previous practice, native speakers have been actively engaged throughout the entire SFT data construction process. They manually collect and write seed questions and topic lists, ensuring linguistic and cultural accuracy from the outset. Additionally, native speakers verify, filter, and edit the synthetic SFT data to maintain high quality.

Our preliminary experiments indicated that relying heavily on dominant English data adversely affects performance. To mitigate this, we strive to maintain a relatively good balance of language representation in our training data this time. Figure 2 shows the language distribution of our SFT data. While English remains a significant portion of the dataset, substantial representation is given to other SEA languages such as Indonesian, Vietnamese, Thai, and others, ensuring a comprehensive and diverse linguistic foundation for the model's training.

Since the first release of SeaLLMs, the task types of SFT data have been greatly expanded. The dataset now includes a diverse range of task types such as coding, math, education-related content, reasoning, general dialogue, table-related tasks, QA, and many more. This expansion ensures that the model is well-rounded and capable of handling a variety of queries and tasks. Additionally, SFT with multiple turns has been significantly increased to enhance the model's ability to engage in natural, multi-turn dialogues, improving its conversational fluency and coherence.

Model safety, trustworthiness, and reliability are also important factors for constructing the SFT pool. To address this, we specifically constructed refusal-type data, enabling the model to decline questions beyond its knowledge boundaries, such as those involving non-existing entities. Furthermore, we carefully curated safety-related data, in-

Model	en	zh	id	th	vi	avg	avg_sea
Gemma-7B	73.2	51.9	47.5	46.0	59.4	55.6	51.0
Sailor-7B-Chat	66.0	65.2	47.5	46.2	51.3	55.2	48.3
SeaLLM-7B-v2.5	75.8	58.1	49.9	50.2	62.2	59.2	54.1
Sailor-14B	74.8	84.0	53.6	52.8	62.1	65.5	56.2
Sailor-14B-Chat	74.9	84.3	55.3	56.6	63.7	67.0	58.5
Qwen2-7B	**81.5**	87.4	53.0	47.9	62.8	66.5	54.6
Qwen2-7B-Instruct	80.9	**88.0**	55.8	55.5	62.4	68.5	57.9
SeaLLMs-v3-7B	80.9	86.3	54.5	53.0	62.8	67.5	56.8
SeaLLMs-v3-7B-Chat	80.9	87.4	**55.8**	**56.9**	**64.9**	**69.2**	**59.2**

Table 1: Results of multilingual world knowledge with the M3Exam benchmark

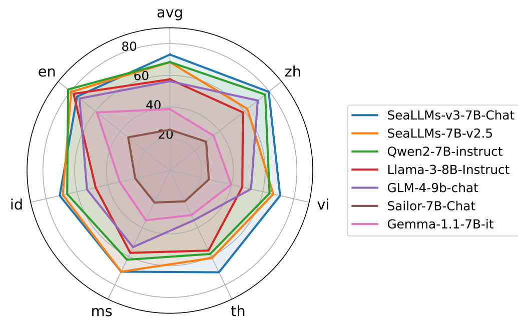

Figure 3: Results of multilingual math with the MGSM benchmark under the zero-shot setting

cluding both general safety data (which are culturally independent, such as general moral principles) and country-specific safety data (which are culturally sensitive). This approach ensures that the model can be safely deployed with cultural considerations in mind, providing accurate and appropriate responses across different cultural contexts.

3.2 Training Details

Two stages of training are employed to optimize the model's performance. In the first stage, a large volume of SFT data is used to equip the model with instruction-following capabilities and to familiarize it with different task types. In the second stage, a smaller but high-quality SFT dataset is utilized to fine-tune the model, ensuring it performs exceptionally well on important tasks. Detailed hyperparamter settings are in Appendix A.

4 Evaluations

We conduct extensive evaluations against models with similar sizes, including Sailor-7B / Sailor-7B-Chat (Dou et al., 2024), Gemma-7b / Gemma-7b-it (Mesnard et al., 2024), Qwen2-7B / Qwen2-7B-Chat (Yang et al., 2024), Llama-3-8B / Llama-3-8B-Instruct (Touvron et al., 2023b), Aya-23-8B (Aryabumi et al., 2024), and the previous versions (mainly v2.5) of SeaLLMs (Nguyen et al., 2023c).

The evaluations can be generally categorized into the following two dimensions: (1) **Model Capability**: We assess the model's performance on human exam questions, its ability to follow multi-turn instructions, its proficiency in mathematics, and its translation accuracy of different language pairs; (2) **Model Trustworthiness**: We evaluate the model's safety and tendency to hallucinate, particularly in the context of Southeast Asia.

4.1 Model Capability Evaluation

4.1.1 Multilingual World Knowledge

Dataset We utilize the M3Exam dataset (Zhang et al., 2023), comprising real human exam questions collected from different countries and spanning different subjects and educational stages. This dataset effectively tests the model's multilingual world knowledge in a manner more akin to real-world settings. We take the questions in English (en), Chinese (zh), Indonesian (id), Vietnamese (vi), and Thai (th). We also employ the translated MMLU (Hendrycks et al., 2021) questions for evaluation, which primarily tests the cross-lingual alignment of the model as the required knowledge is still mainly Western-focused. For evaluation, we use accuracy score as the metric.

Results As shown in Table 1 for results on M3Exam dataset, our models, SeaLLMs-v3-7B and SeaLLMs-v3-7B-Chat, demonstrate competitive performance, with SeaLLMs-v3-7B-Chat achieving the highest average score (0.692) and the highest average score for SEA languages (0.592). Compared to the previous version, SeaLLMs-7B-v2.5, our latest models show significant improvement in overall performance specifically in handling Southeast Asian languages. Furthermore, while the Qwen2-7B-Instruct model performs exceptionally well in English and Chinese, our models exhibit superior performance across a broader range of Southeast Asian languages, highlighting their enhanced multilingual capabilities. The evaluation results on the translated MMLU dataset are in Appendix C.2, which demonstrate similar trends.

4.1.2 Multilingual Math

Dataset We assess the multilingual mathematics capabilities using the MGSM dataset (Shi et al., 2023). Originally, MGSM comprised testing sam-

Model	en	id	jv	km	lo	ms	my	ta	th	tl	vi	zh	avg
Sailor-7B-Chat	49.40	49.78	28.33	2.68	6.85	47.75	5.35	18.23	38.92	29.00	41.76	20.87	28.24
SeaLLM-7B-v2.5	**55.09**	**53.71**	18.13	18.09	15.53	**51.33**	19.71	26.10	40.55	**45.58**	**44.56**	24.18	34.38
Llama-3-8B-Instruct	51.54	49.03	22.46	15.34	5.42	46.72	21.24	**32.09**	35.75	40.80	39.31	14.87	31.22
Qwen2-7B-Instruct	50.36	47.55	29.36	19.26	11.06	42.43	19.33	20.04	36.07	37.91	39.63	22.87	31.32
SeaLLMs-v3-7B-Chat	54.68	52.52	**29.86**	**27.30**	**26.34**	45.04	**21.54**	31.93	**41.52**	38.51	43.78	**26.10**	**36.52**

Table 2: Results of translation with Flores-200 dataset.

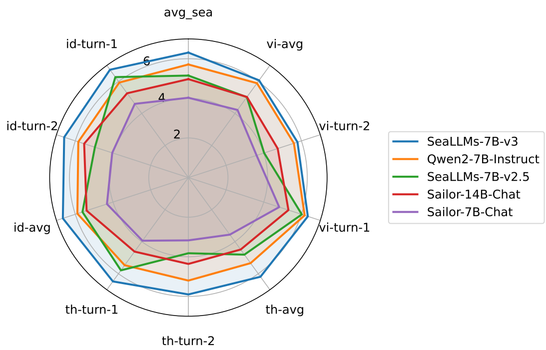

Figure 4: Results of multilingual instruction-following with SeaBench benchmark.

ples solely in English, Chinese, and Thai. To extend this dataset to other SEA languages, specifically Indonesian, Malay, and Vietnamese, we utilize Google Translate to translate the original English questions to questions in those languages.

Results Figure 3 presents the evaluation results on the MGSM benchmark under the zero-shot setting (for testing chat versions). We observe that SeaLLMs-v3-7B-Chat achieves the highest average score (73.1), showing strong performance across all languages. This highlights its superior adaptability and robustness in multilingual math tasks compared to its counterparts like Qwen2-7B-instruct (68.4). Results under the few-shot setting are presented in Appendix C.1.

4.1.3 Multilingual Instruction-following

Dataset As there is no publicly available dataset for testing the model's multi-turn instruction-following capability in SEA languages, we construct our own benchmark, namely SeaBench (Liu et al., 2025), for such evaluation. SeaBench consists of multi-turn human instructions spanning various task types for Indonesian, Vietnamese, and Thai. Further details of SeaBench are in Appendix B.1. Given the two-turn questions, the model under testing generates two-turn responses. These responses are then graded by a stronger LLM (GPT-4o was used in our experiments) with human-crafted reference answers. The scores are then assigned to each turn of the response.

Results As shown in Table 4, SeaLLMs-v3-7B-Chat outperforms all other models in multilingual instruction-following across Indonesian (id), Thai (th), and Vietnamese (vi). It achieves the highest average scores in both individual turns and overall averages for each language. Specifically, SeaLLMs-v3-7B-Chat surpasses the previous version, SeaLLMs-7B-v2.5, by a significant margin (6.31 vs 5.15) and outperforms the strongest baseline model Qwen2-7B-Instruct (6.31 vs 5.70). These results highlight SeaLLMs-v3-7B-Chat's superior ability to generate more coherent and contextually appropriate multi-turn responses.

4.1.4 Translation

Dataset We evaluate the machine translation performances with the test set of Flores-200 (Costa-jussà et al., 2022). We choose all 12 languages for a comprehensive evaluation, including Burmese (my), Chinese (zh), English (en), Indonesian (id), Javanese (jv), Khmer (km), Lao (lo), Malay (ms), Tagalog (tl), Tamil (ta), Thai (th), and Vietnamese (vi). We translate between each pair of languages and report the average 0-shot chrF scores after averaging the results for target languages.

Results As shown in Table 2, SeaLLMs-v3-7B-Chat outperforms other models in machine translation, achieving an average chrF score of 36.52. It excels particularly in Javanese, Khmer, Lao, Burmese, Thai, and Chinese, consistently achieving the highest scores in these languages. Compared to its predecessor, SeaLLMs-7B-v2.5, which has an average score of 34.38, SeaLLMs-v3-7B-Chat shows clear improvement. Additionally, SeaLLMs-v3-7B-Chat surpasses strong baselines like Llama-3-8B-Instruct and Qwen2-7B-Instruct, with average scores of 31.22 and 31.32, respectively. Notably, the model's performance in translating low-resource languages, such as Khmer (27.30) and Lao (26.34), highlights its robustness and effectiveness in handling diverse and challenging translation tasks. This consistent performance

Model	en	zh	vi	th	id	avg
Gemma-1.1-7B-it	68.6	29.9	12.1	30.2	35.5	31.9
Sailor-7B-Chat	37.1	15.8	11.3	13.0	31.7	21.1
SeaLLM-7B-v2.5	44.6	3.9	11.3	49.7	31.2	21.4
Llama-3-8B-Instruct	**80.4**	0.0	0.0	0.0	11.2	16.9
GLM-4-9B-Chat	68.3	52.5	20.2	5.8	9.5	24.7
Qwen2-7B-Instruct	66.7	42.8	62.8	60.5	63.5	57.6
SeaLLMs-v3-7B-Chat	**80.4**	**81.8**	**83.8**	**84.6**	**89.7**	**81.7**

Table 3: Performance in refusing questions about non-existing entities on SeaRefuse-H.

across multiple languages underscores the model's versatility and capability in low-resource language translation settings.

4.2 Model Trustworthiness Evaluation

4.2.1 Hallucination

Dataset A trustworthy LLM should only answer the questions that it knows and abstain from answering questions that it does not know. As there is no existing benchmark in SEA languages for testing the capability of an LLM to refuse questions that exceed its knowledge boundary, we conduct evaluation on our in-house evaluation benchmark, **SeaRefuse**, for such evaluation. SeaRefuse consists of answerable and unanswerable factoid questions across multiple SEA languages. Unanswerable questions are constructed with non-existent entities, designed to surpass the knowledge boundaries of LLMs. Our benchmark includes two test sets: **SeaRefuse-G** and **SeaRefuse-H**. The unanswerable questions in SeaRefuse-G are generated by GPT-4o while the unanswerable questions in SeaRefuse-H are written by humans. In evaluation, we report the F1-score of each model on correctly refusing questions about non-existing entities. We adopt a keyword-matching approach to detect refusal responses. Further details of the dataset and evaluation settings are in Appendix B.3.

Results The experiment results on SeaRefuse-H are shown in Table 3. We observe that SeaLLMs-v3-7B-Chat outperforms all other baseline models by a large margin in zh, vi, th, and id languages. In English, the performance of SeaLLMs-v3-7B-Chat is competitive with Llama-3-8B-Instruct. These results demonstrate the capability of SeaLLMs-v3 to refuse questions that it does not know. The evaluation results on SeaRefuse-G are shown in Appendix C.3, which reveal similar observations.

Model	en	jv	th	vi	zh	avg
Sailor-7B-Chat	78.7	54.9	62.2	67.6	76.2	67.9
Llama-3-8B-Instruct	88.3	26.4	71.1	69.8	77.1	66.5
Sailor-14B-Chat	86.9	30.5	53.7	60.9	72.7	60.9
GLM-4-9B-Chat	77.1	21.3	30.2	60.6	74.9	52.8
Qwen2-7B-Instruct	88.6	43.8	63.8	73.0	87.3	71.3
SeaLLMs-v3-7B-Chat	**88.9**	**60.0**	**73.3**	**83.8**	**92.7**	**79.7**

Table 4: Safety performance of different models.

4.2.2 Safety

To evaluate the models' safety capabilities, we use the questions of SEA languages from the MultiJail dataset (Deng et al., 2023), which includes English (en), Javanese (jv), Thai (th), Vietnamese (vi), and Chinese (zh). Each question in the dataset is potentially malicious, and the model should refuse to answer them. To determine whether the model's response is safe, we first translate the response into English and then prompt GPT-4o to check if the translated response is harmful. The results are reported as the safe rate of the responses.

Table 4 presents the safety capabilities of various models evaluated with the MultiJail dataset. Notably, SeaLLMs-v3-7B-Chat outperforms all other models with an average safe rate of 79.7%, demonstrating robust performance across all languages, particularly excelling in Vietnamese (83.8%) and Chinese (92.7%). In comparison, Qwen2-7B-Instruct follows with a distant second average of 71.3%, with its highest safe rate in Chinese (87.3%). Other models like Sailor-7B-Chat and Llama-3-8B-Instruct also show competitive performance but lag behind in consistency across languages. Notably, the exceptional performance of SeaLLMs-v3 in the three Southeast Asian languages (jv, th, and vi) underscores SeaLLM's effective design, which caters to the linguistic nuances of this region.

5 Conclusion

In conclusion, SeaLLMs 3 represents a significant advancement in the development of large language models for Southeast Asian languages, addressing the region's unique linguistic and cultural challenges. By adopting an efficient language enhancement approach and constructing a comprehensive instruction tuning dataset, SeaLLMs 3 achieves state-of-the-art performance while maintaining cost-effectiveness. Our commitment to reliability and safety, particularly in contextually appropriate responses and knowledge boundaries, further strengthens the model's applicability and trustworthiness. The open-sourcing of both foundational and chat models ensures that SeaLLMs 3 is accessible for a wide range of applications, fostering further innovation and inclusivity in AI development for Southeast Asia.

Limitations

Acknowledgments

We would like to express our special thanks to our professional and native linguists, Tantong Champaiboon, Nguyen Ngoc Yen Nhi and Tara Devina Putri, who helped build, evaluate, and fact-check our sampled pretraining and SFT dataset as well as evaluating our models across different aspects, especially safety.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Uttama Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 4232–4267.

AI Singapore. 2023. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. `https://github.com/aisingapore/sealion`.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 7226–7249. Association for Computational Linguistics.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan N. Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *CoRR*, abs/2405.15032.

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, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. Cosmopedia.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Common Crawl. Common crawl news.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models. *CoRR*, abs/2310.06474.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia. *CoRR*, abs/2404.03608.

Wikimedia Foundation. Wikimedia downloads.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt.

2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net.

Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *CoRR*, abs/2405.10936.

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

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu, and Lidong Bing. 2025. Seaexam and seabench: Benchmarking llms with local multilingual questions in southeast asia. *Preprint*, arXiv:2502.06298.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, 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-Cristian 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, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023a. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2023b. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. *CoRR*, abs/2306.11372.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023c. Seallms - large language models for southeast asia. *CoRR*, abs/2312.00738.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *CoRR*, abs/2404.04925.

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 2023*. OpenReview.net.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022*. OpenReview.net.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024a. Llama beyond english: An empirical study on language capability transfer. *CoRR*, abs/2401.01055.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024b. How do large language models handle multilingualism? *CoRR*, abs/2402.18815.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

A Details of hyperparameters

During the SFT process, different samples are packed together for efficiency, with a maximum length set at 8,192 tokens. The learning rate is set at 1.0e-5, with a warmup ratio of 0.1. Additionally, gradients are clipped at a maximum of 1.0 to prevent exploding gradients.

B Details of Evaluation Setting

B.1 Details of Multilingual Instruction-following Evaluation

Following MT-Bench (Zheng et al., 2023), SeaBench considers 8 task types including writing, roleplay, extraction, reasoning, math, coding, knowledge I (STEM), and knowledge II (humanities/social science). Additionally, considering the characteristics of the multilingual setting, SeaBench includes two more task types: safety and life. The safety task tests whether the model will respond to unsafe queries in a local context, while the life task includes questions likely asked in real-life settings, which might be informally written or even ambiguous.

All questions in SeaBench are manually written by native speakers of each language. During construction, we instructed the annotators to ensure the questions were as localized as possible, e.g., using local entities, concepts, and knowledge in the questions. Furthermore, reference answers have been constructed to ensure fair judgment.

B.2 Details of Multilingual Math Evaluation

It is important to note that in our translations, we adhere to the numerical notation conventions of each respective country. For instance, in both Indonesian and Vietnamese, dots are used as thousands separators, and commas as decimal separators, which is the reverse of the English numeral system. We follow the same convention when evaluating the model's generations.

B.3 Details of Hallucination Evaluation

	Unanswerable question	**Answerable question**
Refused	True-Positive	False-Positive
Answered	False-Negative	True-Negative

Table 5: The confusion matrix for the evaluation of the refusal ability of LLMs.

In the SeaRefuse-G test set, the unanswerable questions are generated by prompting GPT-4o. In the SeaRefuse-H test set, our linguists annotate the unanswerable questions by refining machine-generated unanswerable questions. The answerable questions in both SeaRefuse-G and SeaRefuse-H are collected from open-sourced factoid QA datasets.

We apply the confusion matrix illustrated in Table 5 to compute the F1-score of a system on refusing unanswerable questions about fake entities. We adopt a keyword-matching approach to determine whether a model refuses to answer a factoid question. Specifically, we work with professional and native linguists to devise a set of refusal keywords for English, Chinese, Vietnamese, Indonesian, and Thai, respectively. If the generated response contains any of these refusal keywords, we assume that the response is a refusal response.

C Further Experiment Analysis

C.1 Evaluation results on the MGSM benchmark under the few-shot setting

Figure 5 illustrates the experiment results on the MGSM benchmark under the few-shot setting (for testing base versions of different LLMs). SeaLLMs-v3-7B demonstrates the highest average score (63.1), outperforming other models such as Qwen2-7B (62.3) and GLM-4-9b (60.2), particularly excelling in Indonesian and Thai.

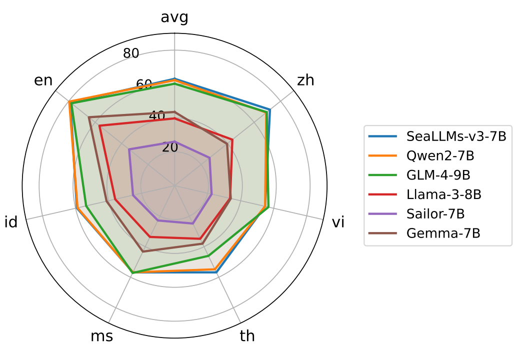

Figure 5: Results of multilingual math with the MGSM benchmark under the few-shot setting.

C.2 Evaluation results on the translated MMLU benchmark

Table 6 shows the results of different models on the translated MMLU dataset, we can see that our SeaLLMs-v3-7B and SeaLLMs-v3-7B-Chat models also outperform other models, particularly in Southeast Asian languages. Compared to the previous SeaLLMs-7B-v2.5 version, our latest models show substantial improvements, particularly in handling Southeast Asian languages (with avg_sea improved from 52.4 to 58.2).

Model	en	zh	id	th	vi	avg	avg_sea
Gemma-7B	63.4	50.9	54.5	49.0	49.4	53.5	51.0
Sailor-7B-Chat	55.8	47.2	48.4	41.4	46.2	47.8	45.4
SeaLLM-7B-v2.5	65.2	54.4	56.5	47.9	52.8	55.3	52.4
Sailor-14B	61.8	56.4	57.0	48.2	53.5	55.4	52.9
Sailor-14B-Chat	62.7	56.1	56.7	49.6	54.1	55.8	53.5
Qwen2-7B	71.0	64.2	60.2	52.0	56.6	60.8	56.3
Qwen2-7B-Instruct	70.8	63.5	59.9	52.4	56.8	60.7	56.4
SeaLLMs-v3-7B	70.6	**65.4**	61.7	53.6	**58.7**	62.0	58.0
SeaLLMs-v3-7B-Chat	**71.3**	64.7	**62.5**	**54.4**	57.8	**62.2**	**58.2**

Table 6: Results of multilingual world knowledge with the translated MMLU benchmark

C.3 Evaluation Results on the SeaRefuse-G dataset

The experiment results of different models on the SeaRefuse-G dataset are shown in Table 7.

Model	en	zh	vi	th	id	avg
Gemma-1.1-7B-it	53.6	28.2	26.2	21.3	30.4	31.9
Sailor-7B-Chat	33.8	18.8	5.2	9.7	16.4	16.8
SeaLLM-7B-v2.5	13.1	1.5	3.2	19.6	0.8	7.7
Llama-3-8B-Instruct	**72.2**	0.0	1.2	0.8	3.9	15.6
GLM-4-9B-Chat	45.0	41.0	21.5	5.4	2.4	23.1
Qwen2-7B-Instruct	63.7	35.8	52.9	46.4	55.9	50.9
SeaLLMs-v3-7B-Chat	71.1	**77.2**	**78.2**	**61.6**	**67.6**	**71.1**

Table 7: Performance in refusing questions about non-existing entities on SeaRefuse-G.