SeaLLMs - Large Language Models for Southeast Asia

Xuan-Phi Nguyen,[*] **Wenxuan Zhang**[*], **Xin Li**[*], **Mahani Aljunied**[*], **Zhiqiang Hu,**
Chenhui Shen[‡], **Yew Ken Chia**[‡], **Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng,**
Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, Lidong Bing[†]

DAMO Academy, Alibaba Group
Video: https://youtu.be/s0mBrHYD_H4
Website: https://damo-nlp-sg.github.io/SeaLLMs

Abstract

Despite the remarkable achievements of large language models (LLMs) in various tasks, there remains a linguistic bias that favors high-resource languages, such as English, often at the expense of low-resource and regional languages. To address this imbalance, we introduce SeaLLMs, an innovative series of language models that specifically focuses on Southeast Asian (SEA) languages. SeaLLMs are built upon popular English-centric models through continued pre-training with an extended vocabulary, specialized instruction and alignment tuning to better capture the intricacies of regional languages. This allows them to respect and reflect local cultural norms, customs, stylistic preferences, and legal considerations. Our comprehensive evaluation demonstrates that SeaLLM models exhibit superior performance across a wide spectrum of linguistic tasks and assistant-style instruction-following capabilities relative to comparable open-source models. Moreover, they outperform ChatGPT-3.5 in non-Latin languages, such as Thai, Khmer, Lao, and Burmese, by large margins while remaining lightweight and cost-effective to operate.

1 Introduction

The advent of large language models (LLMs) has radically transformed the field of natural language processing, demonstrating remarkable abilities in text generation, comprehension, and decision-making tasks (Brown et al., 2020; OpenAI, 2023a,b; Touvron et al., 2023a,b; Thoppilan et al., 2022; Jiang et al., 2023; Wei et al., 2023; Bai et al., 2023). While the proficiencies of these models are extraordinary, the majority of existing LLMs embody a linguistic hierarchy overwhelmingly dominated by English (Ahuja et al., 2023; Lai et al., 2023; Zhang et al., 2023). This dominance

undermines the multilingual capability of such models, with particularly prejudicial outcomes for lower-resource and regional languages, where data scarcity and tokenization challenges lead to disproportionately poor model performance. This linguistic disparity not only impedes access to state-of-the-art AI technologies for non-English-speaking populations but also risks cultural homogenization and the loss of linguistic diversity. While hyper-polyglot models exist (Scao et al., 2022; Muennighoff et al., 2022; Wei et al., 2023), they may pay a high cost for high-resource language performance while lacking in multilingual instruction-following abilities.

Recognizing the urgent need to democratize AI and empower linguistically diverse regions, we introduce SeaLLMs[1], a suite of specialized language models optimized for Southeast Asian languages[2]. These languages, while rich and diverse, often lack the extensive dataset support available for more widely spoken languages, resulting in a stark performance gap in existing LLM applications.

As a long-term continuous effort, as of this writing, SeaLLMs come in three versions (v1, v2, v2.5). SeaLLM-13B-v1, which was pre-trained from Llama-2-13B, eclipses the performance of most available open-source LLMs in a comprehensive array of tasks including world knowledge assessments, language comprehension, and generative capabilities in SEA languages. For English and alike, SeaLLMs do not only preserve, but also demonstrate enhanced performance in tasks that were part of the original Llama training set. When evaluated on multilingual instruction-following tasks with GPT-4 as a judge (Zheng et al., 2023), SeaLLM-13B-v1 outperforms ChatGPT-3.5 by large margins in less-represented languages such

---

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

[1] https://github.com/DAMO-NLP-SG/SeaLLMs
[2] English (Eng), Chinese (Zho), Indonesian (Ind), Vietnamese (Vie), Thai (Tha), Khmer (Khm), Lao, Malay (Msa), Burmese (Mya) and Tagalog (Tgl)
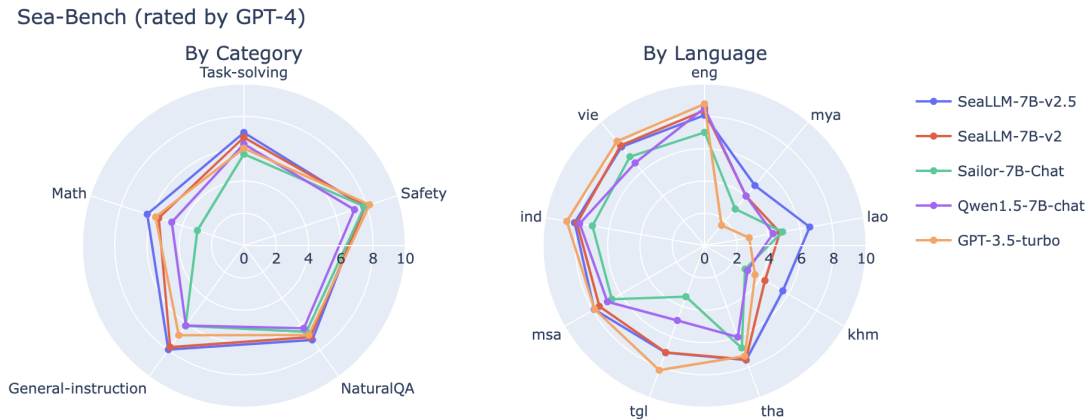
Figure 1: Sea-bench (Section 4.2) scores as evaluated by GPT-4 (Zheng et al., 2023) for different models. Each radar chart compares scores as averaged across 5 categories (left) and 9 languages (right). Detailed breakdown by each category and language is given in Figure 4 in the Appendix.

as Khmer, Lao or Burmese. Meanwhile, SeaLLM-7B-v2, which was pre-trained from Mistral-7B (Jiang et al., 2023), demonstrates better performances in math and commonsense reasoning than comparable baselines, surpassing ChatGPT-3.5 in reasoning for common SEA languages, while being much smaller in sizes. Later, SeaLLM-7B-v2.5, which was further pre-trained from Gemma-7B (Team et al., 2024), shows significant improvements in SEA languages over SeaLLM-7B-v2.

Figure 2 illustrates the four-stage training process of SeaLLMs. In the first stage, detailed in Section 2.3, we conduct continuous pre-training from the foundational models (Touvron et al., 2023b; Jiang et al., 2023) with an extended vocabulary tailored for SEA languages. Next, we fine-tune the model in a novel hybrid paradigm with a mixture of multilingual pre-training data and English-dominant instruction fine-tuning data (Section 3.2). The following stage subsequently fine-tunes the model on a balanced and custom-built multilingual SFT dataset. Finally, we conduct self-preferencing alignment optimization using the SeaLLM model itself, without relying on human annotators or more powerful LLMs (OpenAI, 2023b).

2 Pre-training

2.1 Pre-training Data

The pre-training data comprises a heterogeneous assortment of documents sourced from several publicly accessible repositories (Suárez et al., 2019; Raffel et al., 2019; Computer, 2023; Foundation). Specifically, during the creation of the pre-training data, we include web-based corpora such as Common Crawl (Wenzek et al., 2020), journalistic content such as CC-News, text corpora with expertly-curated knowledge such as Wikipedia (Foundation), and some scholarly publications. After collecting the data, we employ a language identifier (Bojanowski et al., 2017) to retain the documents for the major languages in Southeast Asia, namely Thai, Vietnamese, Indonesian, Chinese, Khmer, Lao, Malay, Burmese, and Tagalog, and discard the remaining ones. Subsequent stages of data refinement involve the multiple modules dedicated to data cleansing and content filtration. We blend such data with the highest quality English data from RedPajama subset (Computer, 2023) in more balanced ratios, as we found that such English data are useful to preserve the original learnt knowledge.

2.2 Vocabulary Expansion

Table 1 describes how expensive it is to process an under-represented non-Latin language. For example, encoding a single sentence in Thai requires 4.3 times more tokens than its English equivalent. The reason for this is that most English language models employ a BPE tokenizer (Sennrich et al., 2016) that inefficiently segments texts from non-Latin scripts into disproportionately lengthy byte sequences, which inadequately represent the underlying semantic content, resulting in diminished model performance (Nguyen et al., 2023). To that end, we propose a novel vocabulary expansion technique, as formally described in Algorithm 1 in the Appendix. This technique involves recursively merging whole-word and sub-word token pieces of a new language from a highly multilingual target tokenizer (i.e., the NLLB tokenizer (Costa-jussà et al.,
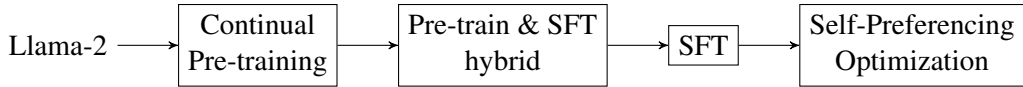
Figure 2: Complete Training Process of SeaLLMs. It begins with continual pre-training Llama-2 with more data of regional languages. Then the models undergo specialized fine-tuning process with multilingual SFT data, before finally being tuned with self-preferencing alignment.

Language	ChatGPT's	Llama's	SeaLLM's
Vie	4.41	3.46	1.48
Zho	2.80	2.36	1.40
Tha	9.09	5.10	1.87
Ind	2.00	2.09	1.36
Khm	15.56	12.14	2.67
Lao	13.29	13.50	2.07
Msa	2.07	2.16	1.50
Mya	17.11	9.85	1.93
Tgl	2.28	2.22	1.91
Eng	1.00 (baseline)	1.19	1.19

Table 1: Averaged compression ratios between the tokenized length of texts of each language produced by different tokenizers versus the baseline tokenized length of same-meaning English equivalents produced by Chat-GPT tokenizer (*i.e.,* it costs 15.6x more tokens to encode Khmer than English with ChatGPT tokenizer). SeaLLM's ratios are applicable only for v1 and v2.

2022)), to the existing LLM tokenizer. This new set of retrieved tokens are then pruned to remove rarely appearing and low-quality tokens before being added to the final SeaLLM tokenizer.

Table 1 demonstrates the efficiency of the new vocabulary. The compression ratio for Thai text has markedly improved from 4.29 to 1.57, signifying a 2.7-fold increase in the length of Thai text that can be encoded within the same context constraints. At the same time, the compression of English text has experienced a negligible reduction of 0.3%, thus maintaining its tokenization effectiveness.

We applied our vocabulary expansion for SeaLLM v1 and v2 with Llama-2 and Mistral-7B as backbones due to their limit 32K-token vocabulary. However, we did not extend the tokenizer for SeaLLM-7B-v2.5, which inherits a sufficiently large 250K-token vocabulary from Gemma-7B.

2.3 Pre-training Process

We organize our pre-training dataset based on the language of the content and the quality of the data, as mentioned in Section 2.1. We setup a separate stream of data for each language, and dynamically control and balance the sampling ratio of each lan-

guage. We pack multilingual documents into a single sequence up to the maximum context length. During the last steps of pre-training, we re-feed the model with more high quality data, which it has previously seen, to readjust the model's learning focus back towards the high-quality data, improving the model's performance.

3 Supervised Fine-tuning (SFT)

3.1 Supervised Fine-tuning Data

Our supervised finetuning (SFT) data consists of many categories, including text understanding and processing, math and logical reasoning, user-centric instruction-following, and natural dialog data. As most public and open-source SFT data are English-only (Longpre et al., 2023; Lian et al., 2023; Mukherjee et al., 2023; Lee et al., 2023), various techniques were implemented to enhance the multilingual aspect of the model. These include sourcing natural data from local websites in natural settings, selectively translating from English data, employing self-instruction, and using advanced prompting techniques (Wang et al., 2022; Madaan et al., 2023; Nguyen et al., 2023). As those synthetically generated data may remain incorrect or low-quality, native speakers[3] were then engaged to further verify, filter, and edit such synthetic responses to finalize the SFT dataset. We find that engaging the annotators to verify and modify model-generated responses is more efficient than having them write responses from scratch. Safety-related data also played a crucial role in fine-tuning SeaLLMs. We manually collected and prepared country-relevant safety data, which covered a broad range of culturally and legally sensitive topics in each of these countries. This was necessary as such topics are often overlooked or may even conflict with open-source English-centric safety data (Deng et al., 2023).

For SeaLLM-7B-v2 and SeaLLM-7B-v2.5, we incorporate significantly more SFT data relating to math and commonsense reasoning. Such data is

---

[3]Hired by our organization, they are not co-authors.

synthetically generated with SeaLLM-13B-v1, as well as strong English models (Jiang et al., 2024; Bai et al., 2023) using a combination of few-shot paraphrasing and translation techniques (Yu et al., 2023).

3.2 Supervised Fine-tuning

Pre-train and SFT Hybrid. As our SFT data is still significantly English due to contributions of open-source data, directly conducting SFT on it may overshadow the smaller SEA language datasets. Therefore, we propose incorporating an additional step prior to complete fine-tuning, namely Pre-train & SFT Hybrid. In this step, the model is further trained on a combination of the pre-training corpus and a large portion of English SFT data, leaving the remaining and more balanced amount of English SFT data to the next stage. During this hybrid stage, the model processes both general pre-training content and instruction-following examples. We mask the source side of the instruction or supervised data to prevent the model from overfitting to the training examples and to reduce the risk of it simply memorizing the input data instead of learning the more generalized ability to follow instructions.

Supervised Fine-tuning. We conduct supervised fine-tuning by compiling instructions from a variety of sources explained in Section 3.1, combining them at random into a single, consolidated sequence to maximize efficiency. To enhance the multi-turn conversation capability, in the later stage of fine-tuning, we further artificially create multi-turn conversations by randomly joining several single-turn instructions together.

3.3 Self-Preferencing Optimization

Alignment from human feedback preference has been key to the success of many AI-assistant language models (Stiennon et al., 2020; Touvron et al., 2023b; Rafailov et al., 2023; Ouyang et al., 2022). To save the cost of human preference annotation work, some have sought to use powerful LLMs like GPT-4 (OpenAI, 2023b) to play the part of a preference data generator (Tunstall et al., 2023). However, that may not even be feasible for low-resource non-Latin languages because of the unfavorable tokenization of ChatGPT as explained in Section 2.2. In other words, even short prompts would exceed their context-length and the API-call costs would explode by up to 17 times.

Therefore, we use our own SeaLLM SFT models to generate preference data by asking it to indicate its preference between two of its own responses, given a question based on certain human-written criteria. To eliminate position bias, we swap the order of the responses and remove samples with inconsistent preference. The data is later used to employ direct preference optimization (Rafailov et al., 2023) to significantly improve the model abilities as an assistant. As such, unlike other works (Mukherjee et al., 2023; Tunstall et al., 2023), our models are free from relying on powerful close-sourced models like GPT-4 to improve the performance in low-resource languages. Our self-preferencing method also shares certain flavors with another self-rewarding mechanism (Yuan et al., 2024).[4]

4 Evaluation

4.1 Model Variants

We trained multiple variants of SeaLLMs, as specified in the following.

- **SeaLLM-7B-v1**: Trained from Llama-2-7B, it supports the 10 official languages used in Southeast Asia.

- **SeaLLM-13B-v1**: Trained from Llama-2-13B, it outperforms ChatGPT-3.5 in most non-Latin SEA languages (Khm, Lao, Mya and Tha) by large margins.

- **SeaLLM-7B-v2**: Trained from Mistral-7B, it outperforms SeaLLM-13B-v1 by far in higher-resource SEA languages (Vie, Ind, Tha), and surpasses ChatGPT-3.5 in math reasoning in SEA languages.

- **SeaLLM-7B-v2.5**: Trained from Gemma-7B, it outperforms SeaLLM-7B-v2 and SeaLLM-13B-v1 remarkably and surpasses ChatGPT-3.5 in various aspects in SEA languages, especially non-Latin languages.

4.2 Sea-bench Peer Comparison

While there are popular benchmarks to evaluate LLMs as a helpful assistant, such as MT-bench (Zheng et al., 2023), they are only English-based and not suitable for evaluating performances in low-resource languages. Due to such a lack of multilingual benchmarks for assistant-style models,

---

[4]Our work was publicly available before Yuan et al. (2024).

| Model | M3Exam | | | | | MMLU |
	Eng	Zho	Vie	Ind	Tha	Eng
ChatGPT-3.5	75.46	60.20	58.64	**49.27**	37.41	**70.00**
SeaLion-7b	23.80	25.87	27.11	24.28	20.29	26.87
Llama-2-13b	61.17	43.29	39.97	35.50	23.74	53.50
Polylm-13b	32.23	29.26	29.01	25.36	18.08	22.94
SeaLLM-7B-v1	54.89	39.30	38.74	32.95	25.09	47.16
SeaLLM-13B-v1	62.69	44.50	46.45	39.28	36.39	52.68
SeaLLM-7B-v2	70.91	55.43	51.15	42.25	35.52	61.89
SeaLLM-7B-v2.5	**76.87**	**62.54**	**63.11**	48.64	**46.86**	64.05

Table 2: Multilingual world knowledge accuracy evaluation across multiple languages and various models of different sizes.

we engaged native linguists to build a multilingual test set with instructions that cover SEA languages, called **Sea-bench**. The linguists sourced such data by translating open-source English test sets, collecting real user questions from local forums and websites, collecting real math and reasoning questions from reputable sources, as well as writing test instructions themselves. Our Sea-Bench consists of diverse categories of instructions to evaluate the models, as follows:

- Task-solving: This type of data comprises various text understanding and processing tasks, such as summarization, translation, etc.

- Math-reasoning: This includes math problems and logical reasoning tasks.

- General-instruction data: This consists of general user-centric instructions, which evaluate the model's ability in general knowledge and writing.

- NaturalQA: This consists of queries posted by real users, often in popular local forums, with a variety of subjects and topics of local interest. The aim is to test the model's capacity to understand and respond coherently to colloquial language, natural expressions and idiomatic language, and locally contextualized references.

- Safety: This includes both general safety and local context-related safety instructions. While most general safety questions are translated from open sources, other local country-specific safety instructions are written by linguists of each language.

As inspired by MT-bench (Zheng et al., 2023), we evaluate and compare SeaLLMs with well-known and state-of-the-art models using GPT-4 as a judge in a score-based grading metrics and a peer comparison (or pairwise comparison) manner.

Figure 1 compares our SeaLLM (v2, v2.5) chat models with Qwen1.5-7B-chat (Bai et al., 2023) and the widely reputed ChatGPT-3.5[5] (OpenAI, 2023a). In the "By Category" chart, SeaLLM-7B-v2.5 performs on par with or surpasses ChatGPT-3.5 across various linguistic and writing tasks. This is largely thanks to the large gap in low-resource non-Latin languages, such as Burmese (Mya), Lao, Khmer and Thai, as seen in the "By language" chart on the right in Figure 1.

Model	Languages	MT-bench
GPT-4-turbo	Multi	9.32
Mixtral-8x7B (46B)	Multi	8.3
Starling-LM-7B-alpha	Mono (Eng)	8.0
OpenChat-3.5-7B	Mono (Eng)	7.81
SeaLLM-7B-v2	**Multi**	**7.54**
SeaLLM-7B-v2.5	**Multi**	**7.43**
Llama-2-70B-chat	Mono	6.86
Mistral-7B-instruct	Mono	6.84
SeaLLM-13B-v1	**Multi**	**6.32**

Table 3: MT-Bench scores (Zheng et al., 2023) for closed, open, multilingual and monolingual (as indicated by their authors on Huggingface.) models.

4.3 MT-bench

We also compare our models with certain baselines on the English MT-Bench (Zheng et al., 2023) in Table 3. As shown, SeaLLM-7B-v2 model demonstrates outstanding ability in English, given its size.

---

[5]gpt-3.5-turbo June 2023 version.

Model	Eng		Zho		Vie		Ind		Tha	
	GSM8K	MATH	GSM8K	MATH	GSM8K	MATH	GSM8K	MATH	GSM8K	MATH
ChatGPT-3.5	**80.8**	34.1	48.2	21.5	55.0	26.5	64.3	26.4	35.8	18.1
Qwen1.5-7B-chat	56.8	15.3	40.0	2.7	37.7	9.0	36.9	7.7	21.9	4.7
SeaLLM-7B-v2	78.2	27.5	**53.7**	17.6	69.9	23.8	**71.5**	24.4	59.6	22.4
SeaLLM-7B-v2.5	78.5	**34.9**	51.3	**22.1**	72.3	**30.2**	71.5	**30.1**	62.0	**28.4**

Table 4: GSM8K and MATH scores (Cobbe et al., 2021; Hendrycks et al., 2021b) and their translated-versions in Chinese, Vietnamese, Indonesian and Thai, under zero-shot chain-of-thought prompting for different models.
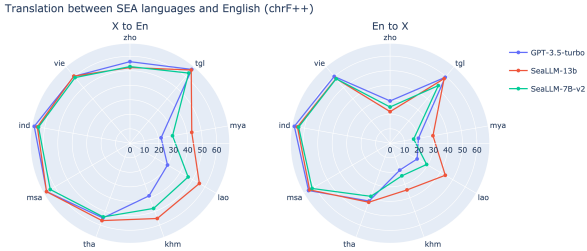

Figure 3: Translation chrF++ scores of various models for both SEA languages to English and English to SEA languages directions.

It is also a rare multilingual model in the 7B realm, especially since it focuses on non-mainstream languages.

4.4 World Knowledge

In this section, we evaluate our models and reputable chat baselines (Touvron et al., 2023b; Wei et al., 2023; OpenAI, 2023a) in terms of world knowledge. For knowledge across languages, we use the M3Exam benchmark (Zhang et al., 2023), which consists of real questions from human exam papers with various degrees of difficulty, ranging from primary school to high school examinations. We evaluate M3Exam with 3-shot native-instruction prompts across English, Chinese, Vietnamese, Indonesian and Thai. We also evaluate our models with the well-known English-centric MMLU benchmark (Hendrycks et al., 2021a).

Table 2 details the evaluations of world knowledge across multiple languages and models of different sizes. SeaLLM-7B-v2.5 exhibits the best performance given its size and is competitive to GPT-3.5.

4.5 Math Reasoning

Table 4 shows the GSM8K and MATH scores (Cobbe et al., 2021; Hendrycks et al., 2021b) for zero-shot chain-of-thought prompting for English and their translated version in Chinese, Vietnamese, Indonesian and Thai. As shown, SeaLLM-7B-v2.5

shows competitive English performance in math reasoning compared to open-source models, with 78.5 in GSM8K and 34.9 in MATH. It also exceeds GPT-3.5 in SEA languages. This is achieved by scaling supervised and preference data in math reasoning in multilingual settings.

4.6 Machine Translation

To benchmark the machine translation performance of our SeaLLMs, we evaluate 4-shot chrF++ scores on the test sets from Flores-200 (Costa-jussà et al., 2022). As can be seen from Figure 3, SeaLLM-13B exhibits clear superiority over ChatGPT-3.5 in low-resource languages, such as Lao and Khmer, while maintaining comparable performance with ChatGPT-3.5 in most higher resource languages (e.g., Vietnamese and Indonesian). We believe our SeaLLMs will play a key role in facilitating communication and cultural exchange across communities in Southeast Asia.

5 Conclusion

In conclusion, our research presents a substantial advance in the development of equitable and culturally aware AI with the creation of SeaLLMs, a specialized suite of language models attuned to the linguistic and cultural landscapes of Southeast Asia. Through rigorous pre-training enhancements and culturally tailored fine-tuning processes, SeaLLMs have demonstrated exceptional proficiency in language understanding and generation tasks, challenging the performance of dominant players such as ChatGPT-3.5, particularly in SEA languages. The models' attunement to local norms and legal stipulations—validated by human evaluations—establishes SeaLLMs as not only a technical breakthrough but a socially responsive innovation, poised to democratize access to high-quality AI language tools across linguistically diverse regions. This work lays a foundation for further research into language models that respect and uphold the rich tapestry of human languages and cultures, ulti-

mately driving the AI community towards a more inclusive future.

6 Limitations

SeaLLMs are among the most linguistically diverse multilingual large language models with remarkable abilities in languages beyond mainstream. However, they do not come without limitations. First, they only scratch the surface of the regionally linguistic diversity with 9 most common and representative languages, while there are hundreds other languages spoken in the Southeast Asia, such as Javanese and Tamil. Second, despite outperforming other popular models in non-Latin low-resource languages, SeaLLM models still suffer from considerable hallucination and degeneration under certain circumstances for languages such as Burmese and Lao. Mild hallucination is still inevitable for other common languages.

References

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: multilingual evaluation of generative AI. *CoRR*, abs/2303.12528.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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. *arXiv preprint arXiv:2110.14168*.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.

Wikimedia Foundation. Wikimedia downloads.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *CoRR*, abs/2304.05613.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, cheap, and powerful refinement of llms.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorca: An open dataset of gpt augmented flan reasoning traces. https://https://huggingface.co/Open-Orca/OpenOrca.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey

Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4.

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

OpenAI. 2023a. Chatgpt (june 2023 version.

OpenAI. 2023b. Gpt-4 technical report. *arXiv preprint*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

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. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston.

2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020.*

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *CoRR*, abs/2306.05179.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685.*

A Vocabulary Expansion

Algorithm 1 explains in details how we perform selective and recursive merger of tokens from target NLLB vocabulary into the original Llama vocabulary to enrich the linguistic coverage for new and low-resource languages. Specifically, given a small seed unlabeled dataset of a given new language, the algorithm first tokenizes a document with the current Llama tokenizer. The resulting tokens are then exhaustively merged into longer tokens that are supported by the target NLLB vocabulary. During this merger process, any intermediate sub-word is also added to the Llama tokenizer as long as they exist in the rich NLLB vocabulary.

The new set of collected tokens are then pruned to remove rarely appearing and low-quality tokens before being added to the final SeaLLM tokenizer. This frequency-based pruning process ensures the new language is sufficiently and efficiently encoded without introducing tokens from other existing languages (*e.g.,* English), which may disrupt the learned knowledge during the Llama-2 pre-training stage.

B Sea-bench Evaluation Details

Figure 4 breaks down the GPT-4 rated Sea-bench score-based evaluations of SeaLLM-13b and other baselines by both language and task category. As shown, our SeaLLM-13b model far exceeds ChatGPT-3.5 in most non-Latin languages, such as Burmese (Mya), Lao and Khmer, though it trails behind this formidable competitor in Latin-based languages, mostly in math reasoning skills.

Algorithm 1 Vocabulary Extension algorithm: V_i is Llama vocabulary, V_t is target NLLB vocabulary, D is unlabeled data and m is minimum frequency.

1: **function** EXHAUSTIVEMERGE(V_i, V_t, t_V)
2:      $T_{new} \leftarrow$ empty set $\emptyset$
3:      **repeat**
4:          **for** each consecutive token pair (prev, next) in t_V **do**
5:              $t_{merged} \leftarrow \langle\text{prev}\rangle\langle\text{next}\rangle$ ▷ Form a new token
6:              **if** t_{merged} exists in V_t **then**
7:                  Replace (prev, next) with t_{merged} in t_V ▷ Update t_V with new token
8:                  $T_{new} \leftarrow T_{new} \cup t_{merged}$
9:                  **break**
10:      **until** no new token added to T_{new}
11:      **return** T_{new}
12: **function** VOCABEXTEND(V_i, V_t, D, m)
13:      $V \leftarrow V_i$
14:      $F \leftarrow$ empty set $\emptyset$
15:      $T \leftarrow$ empty set $\emptyset$
16:      **for** document d in D **do**
17:          $t_V \leftarrow$ tokenize(V, d)   ▷ tokenize the document
18:          $T_{new} \leftarrow$ EXHAUSTIVEMERGE(V_i, V_t, t_V)   ▷ obtain new words from V_t based on d
19:          $V \leftarrow V \cup T_{new}$ ▷ update V with new words T_{new}
20:          $T \leftarrow T \cup T_{new}$
21:          $F \leftarrow$ Update frequencies of T_{new} to F ▷ update appearance frequencies of T_{new}
22:      $T \leftarrow$ Prune $t_i \in T$ with corresponding $f_t \in F$ where $f_t < m$ ▷ Remove rare words
23:      $V_{final} \leftarrow V_i \cup T$
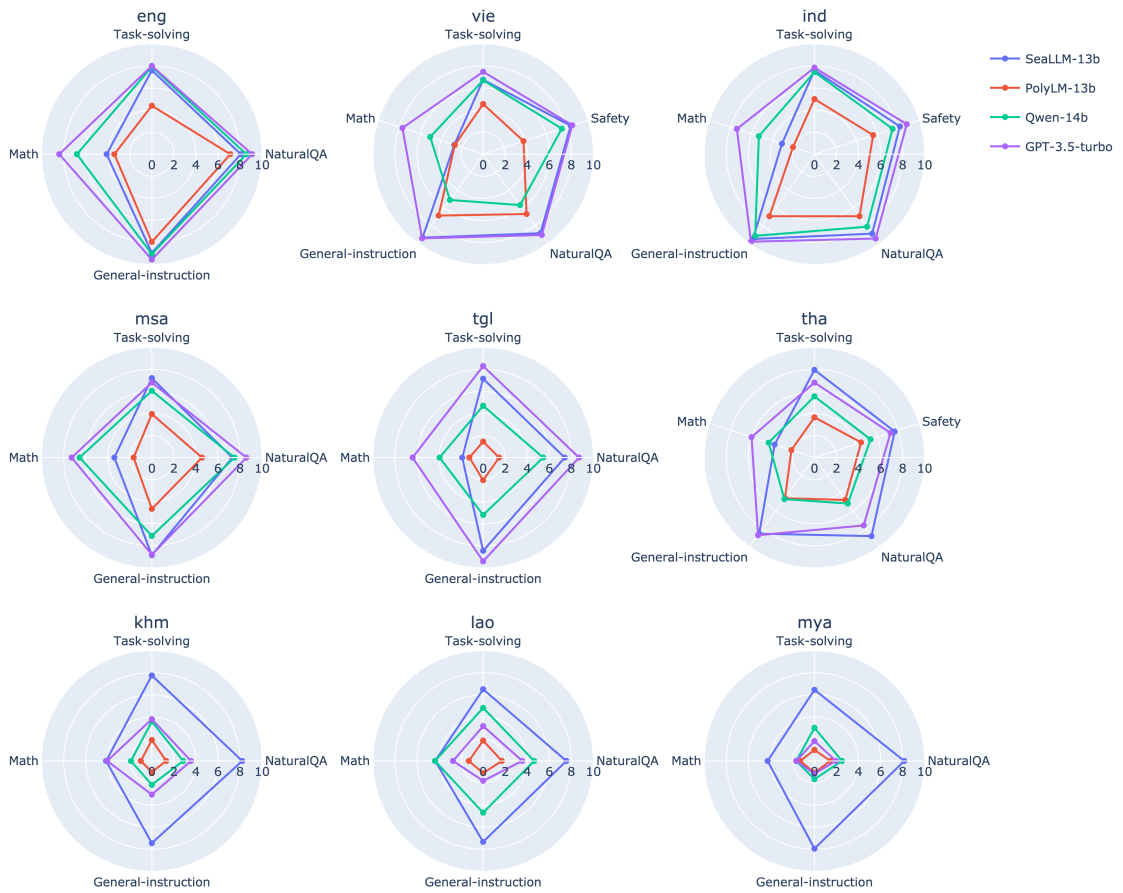24:   **return** V_{final}

Figure 4: Sea-bench scores as evaluated by GPT-4 for different models across 9 languages and 5 categories.