# SEA-LION: Southeast Asian Languages in One Network

**Anonymous ACL submission**

## Abstract

Recently, Large Language Models (LLMs) have dominated much of the artificial intelligence scene with their ability to process and generate natural languages. However, the majority of LLM research and development remains English-centric, leaving low-resource languages such as those in the SouthEast Asian (SEA) region under-represented. To address this representation gap, we introduce **Llama-SEA-LION-8B-IT** and **Gemma-SEA-LION-9B-IT**, two cutting-edge multilingual LLMs designed for SEA languages. The SEA-LION family of LLMs supports 11 SEA languages, namely English, Chinese, Indonesian, Vietnamese, Malay, Thai, Burmese, Lao, Filipino, Tamil, and Khmer. Our work leverages large-scale multilingual continued pre-training with a comprehensive post-training regime involving multiple stages of instruction fine-tuning, alignment, and model merging. Evaluation results on multilingual benchmarks show that our models achieve state-of-the-art performance across LLMs supporting SEA languages. We open-source the models[1] to benefit the wider SEA community.

## 1 Introduction

Large language models (LLMs) have significantly transformed the field of natural language processing, achieving remarkable performance in text generation, summarization and sentiment analysis (Brown et al., 2020; OpenAI, 2023; Dubey et al., 2024; Rivière et al., 2024; Zhang et al., 2024b; Yeo et al., 2024).

Despite their impressive capabilities, most LLMs remain heavily English-centric (Wendler et al., 2024; Zhong et al., 2024). Unfortunately, this situation has led LLMs in regions with many under-represented languages such as **S**outh**e**ast **A**sia (**SEA**) to suffer. Languages with lower resources, such as Filipino, Lao, Burmese and Khmer in the SEA region, are not supported by many open-source English-centric LLMs. This underscores the need to bridge the resource and representation gap between English and SEA languages.

Recently, there have been many attempts to create multilingual LLMs in an open-source manner, e.g., BLOOM (Scao et al., 2022), a project aimed at increasing multilingual presence in open-source LLMs by supporting 46 languages. Popular LLM families such as Llama (Dubey et al., 2024), Gemma (Rivière et al., 2024) and Qwen (Yang et al., 2024a) have also introduced multilingual LLMs for their latest iteration. During our evaluations, we found that the performance of these models is acceptable in the general case, i.e., if we consider evaluation benchmarks formulated from English datasets, but we observe that the performance degrades on SEA-specific benchmarks. Moreover, researchers have also introduced LLMs such as SeaLLMs (Nguyen et al., 2024; Zhang et al., 2024a) and Sailor (Dou et al., 2024) to specifically address the LLM gap in SEA languages. However, the performance of these models is less than ideal for languages such as Thai or Tamil[2] (10X et al., 2024; AI Products Team, 2024).

In this paper, we address the issues by proposing a robust open-source Southeast Asian model with data transparency for reproducibility, namely **SEA-LION** – a family of LLMs continued pre-trained (CPT) and fine-tuned on Llama-3.1-8B-Instruct for Llama-SEA-LION-8B-IT and Gemma-2-9B for Gemma-SEA-LION-9B-IT with a focus on SEA languages. To tackle the performance problem, we utilize 200 billion English, code and SEA languages tokens as well as 16.8 million English and SEA languages instruction and answer pairs for CPT and post-training steps respectively, to achieve a significant improvement in SEA languages. In

[1] [LINK_HIDDEN]

[2] Tamil is one of the official languages in Singapore. It is also spoken in other areas in the SEA region, such as Malaysia.

order to allow our models to be used by everyone without restrictions, we release our models under a fully open MIT [3]. We benchmark our models against the SEA-HELM(Susanto et al., 2025) and Open LLM Leaderboard[4] with other LLMs of similar sizes in Southeast Asia like Sailor 2 (Team, 2024) and SeaLLMs 3 (Zhang et al., 2024a), where our models achieve state-of-the-art performances. We summarize the contribution of our paper as follows.

- We released two LLMs, **Llama-SEA-LION-8B-IT** and **Gemma-SEA-LION-9B-IT**, that are meticulously trained to accurately represent the unique linguistic diversity of SEA languages.
- We also provide in-depth insights in this paper into our end-to-end training workflow to benefit the community developing multilingual LLMs.
- We present a reproducible dataset development process, covering sourcing and the model training process. Upon the acceptance of this paper, we will release our training artifacts, including the training dataset, training scripts, training checkpoints, and fine-tuned models, including **Llama-SEA-LION-8B-IT** and **Gemma-SEA-LION-9B-IT**, to provide strong baselines, promote reproducibility, and enable future research on applications that require SEA-specific knowledge.

# 2 Continued pre-training (CPT)

## 2.1 Pre-training data

The CPT data consists of a curated set of English, multilingual, and code corpora from several open source repositories like Dolma (Soldaini et al., 2024), FineWeb (Penedo et al., 2024), the-stackv2 (Lozhkov et al., 2024), SEA-LION-Pile (AI Singapore, 2023), SEA-LION-Pile-v2 (AI Singapore, 2025), as well as documents from CommonCrawl (CommonCrawl, 2024) and from the public domain, such as Wikipedia (Foundation, 2024). For SEA-LION-Pilev2, we filter CommonCrawl WARC data for documents in SEA languages (i.e., Burmese, Simplified Chinese, Indonesian, Khmer, Lao, Malay, Filipino, Tamil, Thai, and Vietnamese) using the pretrained fast-text language classifier (Joulin et al., 2017).

A document is retained if the language code reported in its metadata matches with that of one of the aforementioned SEA languages. Additionally, we further clean up the data with Trafilatura (Barbaresi, 2021). To determine the best dataset ratio between SEA languages, code and English for the CPT process, we perform a series of small-scale CPT experiments each with a training budget of 10B tokens and varying proportions of English, code and SEA language data. We settled on an optimal data mix ratio of 55% SEA languages, 25% English, and 20% code tokens for a budget of 200B tokens. For a detailed breakdown of the token count by languages, please refer to Table 6.

## 2.2 CPT process

**Model selection**. For the models to CPT from, we choose Llama-3.1-8B-Instruct (Dubey et al., 2024) and Gemma-2-9B (Rivière et al., 2024).

**Training setup**. Following previous works (Dou et al., 2024), we use BPE-Dropout (Provilkov et al., 2020) to increase the performance and robustness of the training. We use a Warmup-Stable-Decay (WSD) (Hu et al., 2024) scheduler with warm-up and cooldown phases each representing 10% of the entire training budget. We use the AdamW (Loshchilov and Hutter, 2019) optimiser with the maximum learning rate (LR) set to $1e^{-5}$ and the final LR after cooldown is $1e^{-7}$. Following Wortsman et al. (2024), we set epsilon to $1e^{-15}$. We use Composer (Team, 2021) and LLM Foundry (Team, 2022) for distributed training using Fully Sharded Data Parallel (Zhao et al., 2023) on a cluster of eight nodes of the p5.48xlarge instance from Amazon Web Services (AWS). The total training duration was approximately 6 days and 10 days for the Llama 3.1 and Gemma 2 models, respectively. In this paper, we refer to the post-CPT models as *Llama-SEA-LION-8B* and *Gemma-SEA-LION-9B* for the Llama 3.1 and Gemma 2 continued pre-trained models respectively.

# 3 Post-training

## 3.1 Post-training data

The post-training data consists of 3 subsets of data for Stage 1 IFT, Stage 2 IFT, and the Preference dataset for alignment, respectively. We describe the training data information of each step as follows.

**Stage 1 IFT**. In this step, we employ Infinity-Instruct [Foundation and Chat] (Beijing Academy of Artificial Intelligence, 2024) and OpenMath-Instruct 2 (Toshniwal et al., 2024) to improve the mathematical, reasoning, and

[3] https://mit-license.org/

[4] Open LLM Leaderboard

coding skills of the instruction model. The full details of the training data are shown in Appendix 7.

**Stage 2 IFT**. Then, in this step, we use generalized large-scale instructions on the previous instruction model. In particular, we employ 22 existing datasets (written in English, Thai, and Vietnamese) and formulate new 22 synthetic datasets using various models and techniques to create SEA instruction datasets (see Appendix A.3 for the full data generation details). As shown in Appendix 9, we use a total of 7,298,828 instruction samples that cover 11 languages.

**Helpfulness and preference alignment**. We also conduct an alignment learning on top of the instruction model using a feedback dataset called UltraFeedBack (Cui et al., 2024). In addition, we also synthesized the SEA version of the UltraFeedBack using NemoTron-70b with Gemma2 as a reward model, see Appendix A.4 for the full details.

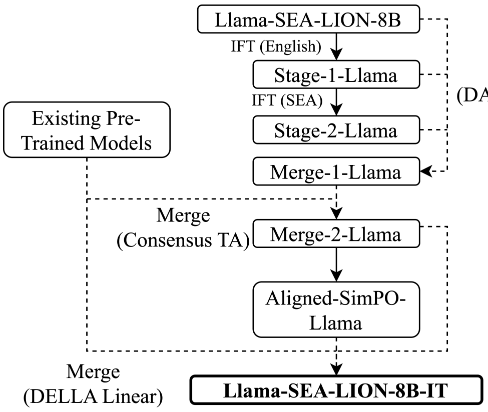

Figure 1: Training process of **Llama-SEA-LION-8B-IT** (Section 3.2.1). The post-training process consists of 2 stages of instruction fine-tuning, an alignment stage and multiple merge stages. Dotted lines denote a merge stage and solid lines denote an alignment stage.

## 3.2 Post-training process

We use LLaMaFactory (Zheng et al., 2024b) with DeepSpeed (Rasley et al., 2020) for all Instruction Fine Tuning (IFT) and alignment steps. All IFT stages are performed using full model fine-tuning where the models are from the previous step (Section 2.2) and existing models. We use MergeKit (Goddard et al., 2024) with a value of 1 for weight and density parameters for all merge steps. Models selected for merging are selected empirically, based on the openness of model licenses, the suitability for merging and performance.

### 3.2.1 Llama-SEA-LION-8B-IT

**Stage 1 IFT** As shown in Figure 1, we started off the post-training phase with IFT of *Llama-SEA-LION-8B* with the Infinity Instruct (Foundation) (Beijing Academy of Artificial Intelligence, 2024) and OpenMathInstruct2 (Toshniwal et al., 2024) datasets. Both datasets contain approximately 9.5 million instruction pairs, primarily in English and centered around reasoning, math, and code. We refer to the model at this stage as *Stage-1-Llama*.

**Stage 2 IFT** We performed a second round of IFT using the SEA-Instruct dataset, which consists of approximately 7.3 million instruction pairs, of which 5 million instruction pairs are generated using the Gemma-2-27B-Instruct (Rivière et al., 2024) model and the Qwen2.5-32B-Instruct model (Yang et al., 2024a) in SEA languages. The remaining are English language instruction pairs from the Infinity-Instruct (Chat) (Beijing Academy of Artificial Intelligence, 2024) dataset. We refer to the model at this stage as *Stage-2-Llama*.

**First merge** After finishing the IFT stages, we performed the first of a series of merges by merging *Stage-1-Llama* and *Stage-2-Llama* into the *Llama-SEA-LION-8B* using the DARE TIES (Yu et al., 2024; Ilharco et al., 2023) method. We refer to the model at this stage as *Merge-1-Llama*.

**Second merge** In order to mitigate catastrophic forgetting due to the fine-tuning process (Alexandrov et al., 2024), we performed the second round of merge by merging top-performing instruction-tuned models that share the Llama 3.1 lineage. We merge the original Llama-3.1-8B-Instruct, Llama3-8B-SEA-LION-v2.1-Instruct (SEA-LION Team, 2024), and SuperNova-Lite (Arcee-AI, 2024) into *Merge-1-Llama* using the Consensus TA (Wang et al., 2024b; Ilharco et al., 2023) merge method. We refer to the model at this stage as *Merge-2-Llama*.

**Helpfulness and preference alignment** We performed one round of alignment on *Merge-2-Llama* using SimPO (Meng et al., 2024) with the SEA-Preference dataset. We refer to the model at this stage as *Aligned-SimPO-Llama*.

**Final merge** Lastly, we perform a merge using the DELLA-Linear merge. With the original Llama-3.1-8B-Instruct model as the base for merging, we merge in *Merge-2-Llama* and *Aligned-SimPO-Llama* to produce the final model, ***Llama-SEA-***

***LION-9B-IT***.

### 3.2.2 Gemma-SEA-LION-9B-IT

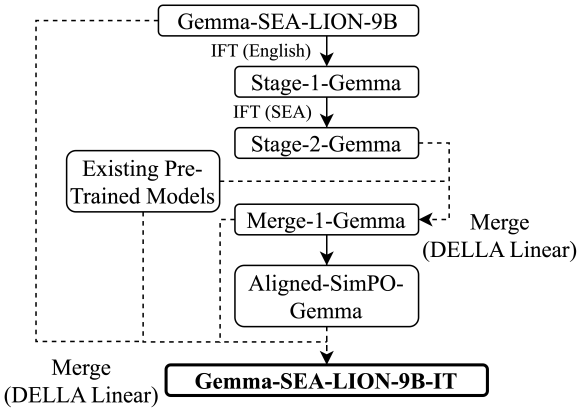

Figure 2: Training process of Gemma-SEA-LION-9B-IT (Section 3.2.2). The post-training process comprises two stages of instruction fine-tuning, an alignment stage, and multiple merge stages. Dotted lines denote a merge stage and solid lines denote an alignment stage.

**Stage 1 and Stage 2 IFT** Similar to the *Llama-SEA-LION-8B-IT*, we started off the post-training phase with both stages of IFT using the same datasets on the Gemma-2-9B model (Rivière et al., 2024). We refer to both models at stage 1 and stage 2 as *Stage-1-Gemma* and *Stage-2-Gemma*, respectively.
**First merge** We merge the Gemma-2-9B-IT (Rivière et al., 2024) and *Stage-2-Gemma* into Gemma-2-9B using the DELLA Linear method. We refer to the model at this stage as the *Merge-1-Gemma*.
**Helpfulness and preference alignment** Using the *Merge-1-Gemma* as the base model, we performed one round of alignment using SimPO with the SEA-Preference dataset. We refer to the model at this stage as the *Aligned-SimPO-Gemma*.
**Final merge** Finally, using the Gemma-2-9B model as the base model, we merged *Merge-1-Gemma*, FuseChat Gemma-2-9B-Instruct (Yang et al., 2024b), *Gemma-SEA-LION-9B*, and *Aligned-SimPO-Gemma* into it to produce the final model ***Gemma-SEA-LION-9B-IT***.

### 3.3 Discussion

This post-training workflow emphasizes the careful balance between general capabilities, SEA-specific linguistic fluency, and natural conversational abilities. Each step in the workflow is designed to progressively refine the model, ensuring it meets the diverse needs of users in the Southeast Asian region.

The entire post-training process for ***Gemma-SEA-LION-9B-IT*** and ***Llama-SEA-LION-8B-IT*** took approximately 1350 and 1024 GPU hours, respectively, on eight H100 GPUs. To make the training efficient, all post-training steps utilize Liger Kernel (Hsu et al., 2024) for substantial memory savings of approximately 60%.

## 4 Experimental Setup

### 4.1 Competitive methods

For the evaluation, we compared our models against well-known LLMs for multilingual and SEA languages, such as *SeaLLMsv3* (Zhang et al., 2024a), *Sailorv2* (Team, 2024), *Qwen 2.5* (Yang et al., 2024a), *Gemma 2* (Rivière et al., 2024) and *Llama 3.1* (Dubey et al., 2024), where the parameters of those models are less than 10 billion parameters, similar to our models.

### 4.2 Evaluation Benchmarks

To evaluate the robustness of our proposed models, we compare our models to competitors in three benchmarks.
**SEA Benchmarks**. We evaluated the multilingual performance of each LLM using the *SEA-HELM Leaderboard* (Leong et al., 2023; Susanto et al., 2025). We selected SEA-HELM because the design choice of this benchmark reflects the performance of SEA culture and knowledge the most compared with other existing benchmarks (DAMO-NLP-SG, 2024; Lovenia et al., 2024; Wang et al., 2024a). We also evaluate on a wide-range SEA coverage language benchmark called *SEACrowd* (Lovenia et al., 2024). This benchmark consists of all SEA languages for natural language understanding and generation datasets. However, due to maintenance reasons, we cannot reproduce the NLG benchmark of SEACrowd. Therefore, we experiment only with the NLU benchmark (zero-shot), which has 131 data subsets, 7 tasks, and 31 SEA indigenous languages.
**English performance**. We also evaluated the English performance of the models using the Open LLM Leaderboard (HuggingFace, 2024). This is because English is also widely used in SEA countries. Therefore, we need to evaluate the understanding and knowledge of LLMs in the English benchmark as well. The leaderboard consists of six benchmarks, IFEval (Zhou et al., 2023), Big Bench Hard (Suzgun et al., 2023), MATH (Hendrycks et al., 2021), GPQA (Rein et al.,

SEA-HELM

| | | NLU, NLG, NLR, NLI | | | | Instruction Following | | |
|---|---|---|---|---|---|---|---|---|
| **Models** | **Average** | **ID** | **VI** | **TH** | **TA** | **ID** | **VI** | **TH** |
| Meta-Llama-3.1-8B | 35.37 | 42.33 | 40.67 | 35.13 | 38.88 | 16.19 | 19.05 | 9.00 |
| SeaLLMs-v3-7B | 37.04 | 44.79 | 48.29 | 43.53 | 27.45 | 26.67 | 35.24 | 26.00 |
| Gemma-2-9B | 41.48 | 47.65 | 43.28 | 42.00 | 53.26 | 4.76 | 3.81 | 10.00 |
| Qwen2.5-7B | 41.98 | 51.63 | **52.17** | 46.55 | 36.60 | **31.43** | **36.19** | 30.00 |
| Sailor2-8B | 42.62 | 53.23 | 47.33 | 46.64 | 45.04 | 30.48 | 30.48 | **35.00** |
| Llama-SEA-LION-8B | 41.42 | 44.98 | 46.25 | 42.79 | 43.03 | 25.71 | 32.38 | 23.00 |
| Gemma-SEA-LION-9B | **48.67** | **57.16** | 49.39 | **47.16** | **60.56** | 25.71 | 20.00 | 27.00 |

Table 1: SEA-HELM multilingual benchmark on NLU, NLG, NLR, NLI and instruction following on base and continued pre-trained models of similar sizes.

Open LLM Leaderboard

| **Models** | **Average** | **MMLU-PRO** | **BBH** | **GPQA** | **MATH Lvl 5** | **IFEval (EN)** | **MUSR** |
|---|---|---|---|---|---|---|---|
| Meta-Llama-3.1-8B | 13.9 | 24.95 | 25.29 | 6.32 | 5.14 | 12.7 | 8.98 |
| Sailor2-8B | 17.71 | 25.74 | 27.62 | 4.87 | 7.02 | 21.95 | **19.03** |
| Gemma-2-9B | 21.15 | 34.48 | 34.1 | **10.51** | 13.14 | 20.4 | 14.3 |
| SeaLLMs-v3-7B | 24.00 | 35.71 | 34.57 | 9.28 | 18.81 | 32.94 | 12.68 |
| Qwen2.5-7B | **24.99** | **37.39** | 35.81 | 9.96 | **18.88** | **33.74** | 14.14 |
| Llama-SEA-LION-8B | 16.61 | 27.6 | 26.04 | 7.49 | 9.89 | 16.56 | 12.07 |
| Gemma-SEA-LION-9B | 22.41 | 32.78 | **37.24** | 10.29 | 9.89 | 30.12 | 14.11 |

Table 2: Open LLM Leaderboard benchmarks across different continued pre-trained models of similar sizes.

2023), MuSR (Sprague et al., 2024) and MMLU-PRO (Wang et al., 2024c). Moreover, we also evaluate the CPT models on SEA-HELM and the Open LLM Leaderboard since these benchmarks support the CPT evaluation.

## 5 Experimental Results

To understand the robustness and generalization of our proposed models, we conduct three studies as follows. Section 5.1 evaluates the robustness of continual pre-training models using SEA-HELM and the Open LLM leaderboard. In Section 5.2, we compare our instruction fine-tuning models with competitors in three benchmarks to demonstrate the generalization of our models. Lastly, we discuss the design choice of our models in Section 5.3.

### 5.1 Continued Pre-Training Results

**SEA performance**. The CPT stage is primarily focused on gaining SEA language capabilities and knowledge. For the purpose of comparison against base and CPT models, as shown in Table 1, we observed a 6.05 and 7.19 average SEA-HELM performance increase over the *Meta-Llama-3.1-8B* and *Gemma-2-9B* for *Llama-SEA-LION-8B* and *Gemma-SEA-LION-9B*, respectively. We observed a much larger average increase with instruction following capabilities in particular, which we attribute to the fact that our CPT models are trained from the instruction models rather than from the base models. Moreover, in the average performance, we found that our Gemma-SEA-LION-9B models perform the best compared to other models. This emphasizes a strong reason to perform CPT for improving the performance of SEA languages, rather than skipping the CPT and performing SFT directly.

**English performance**. For the English performance, as shown in Table 2, both CPT models also managed to perform competitively against the *Meta-Llama-3.1-8B* and *Gemma-2-9B base* models on the Open LLM Leaderboard benchmarks. This indicates that our choice of retraining with a proportion of 25% English tokens has been beneficial in mitigating catastrophic forgetting, which has been shown to stem from CPT (Zheng et al., 2024a). Although our CPT models perform lower than Qwen and SeaLLMs on this benchmark, we outperform them on the SEA language instead, which is the main focus of this work.

| SEA-HELM | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NLU, NLG, NLR, NLI | | | | Instruction Following | | | MTBench | | |
| **Models** | **Average** | **ID** | **VI** | **TH** | **TA** | **ID** | **VI** | **TH** | **ID** | **VI** | **TH** |
| SeaLLMs-v3-7B-Chat | 39.19 | 42.72 | 48.50 | 42.59 | 12.06 | 57.14 | 53.33 | 47.00 | 59.81 | 65.24 | 56.59 |
| Llama-3.1-8B-Instruct | 41.48 | 51.50 | 51.31 | 45.32 | 15.40 | 77.14 | 75.24 | 63.00 | 56.38 | 57.59 | 54.34 |
| Sailor2-8B-Chat | 43.13 | 48.98 | 48.01 | 45.44 | 28.29 | 49.52 | 45.71 | 40.00 | **69.76** | 66.97 | **73.94** |
| Qwen2.5-7B-Instruct | 44.58 | 60.28 | 53.46 | 53.43 | 21.03 | 81.90 | 69.52 | 66.00 | 65.66 | 66.80 | 68.71 |
| Gemma-2-9B-IT | 55.33 | 64.04 | 59.86 | 57.22 | 52.28 | 88.57 | 78.10 | 71.00 | 68.78 | 68.37 | 73.51 |
| Stage-1-Llama | 50.76 | 51.84 | 51.83 | 46.23 | 27.53 | 69.52 | 73.33 | 59.00 | 42.74 | 46.41 | 46.46 |
| Stage-2-Llama | 59.49 | 53.87 | 55.18 | 50.92 | 44.80 | 77.14 | 76.19 | 67.00 | 50.90 | 53.72 | 46.97 |
| Merge-1-Llama | 59.36 | 56.73 | 56.82 | 51.71 | 46.63 | 81.90 | 82.86 | 67.00 | 57.04 | 54.01 | 50.28 |
| Merge-2-Llama | 58.01 | 59.19 | 52.63 | 51.89 | 35.40 | 87.62 | 80.95 | 78.00 | 56.38 | 59.32 | 58.86 |
| Aligned-SimPO-Llama | 51.30 | 54.86 | 51.69 | 46.77 | 26.40 | 82.86 | 80.00 | 68.00 | 68.20 | 64.68 | 64.92 |
| Llama-SEA-LION-8B-IT | 61.84 | 60.50 | 61.48 | 55.92 | 43.61 | 84.76 | 85.71 | 76.00 | 62.65 | 68.32 | 65.13 |
| Stage-1-Gemma | 56.56 | 55.06 | 54.51 | 51.96 | 42.74 | 66.67 | 74.29 | 61.00 | 47.35 | 47.26 | 55.05 |
| Stage-2-Gemma | 66.66 | 64.10 | 61.76 | 56.90 | 57.85 | 89.52 | 82.86 | 76.00 | 60.54 | 58.93 | 58.76 |
| Merge-1-Gemma | 69.26 | 66.25 | 64.95 | **59.74** | **60.41** | 89.52 | **91.43** | **82.00** | 66.45 | 64.47 | 65.00 |
| Aligned-SimPO-Gemma | **69.37** | 65.69 | **65.47** | 59.51 | 57.38 | 86.67 | 88.57 | 78.00 | 68.89 | **73.67** | 73.51 |
| Gemma-SEA-LION-9B-IT | 69.35 | **66.26** | 64.93 | 59.23 | 58.82 | **94.29** | 88.57 | 78.00 | 65.85 | 73.27 | 69.07 |

Table 3: SEA-HELM multilingual benchmark on NLU, NLG, NLR, NLI, instruction following and multi-turn chat on instruct models of similar sizes.

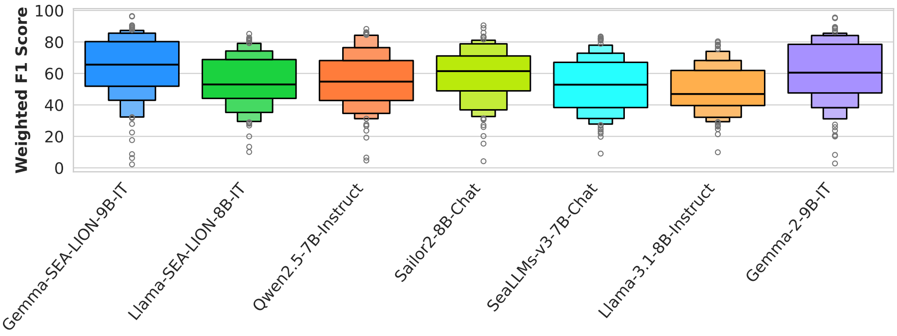

Figure 3: Zero-shot model performance across NLU tasks in SEA languages.

| **Model** | **NLU Score** |
|---|---|
| SeaLLMs-v3-7B-chat | 52.68 |
| Llama-3.1-8B-Instruct | 49.94 |
| Sailor2-8B-Chat | 60.21 |
| Qwen2.5-7B-Instruct | 54.51 |
| Gemma-2-9B-IT | 60.21 |
| Llama-SEA-LION-8B-IT | 55.10 |
| Gemma-SEA-LION-9B-IT | **64.13** |

Table 4: The average NLU performance across 131 data subsets and 31 indigenous languages.

### 5.2 Instruction Fine-tuning Results

In this study, we compare our models with competitors on SEA-HELM, SEACrowd, and the Open LLM Leaderboard as follows.

**SEA-HELM**. As shown in Table 3, the SEA-HELM benchmark performance demonstrates that our instruct models, *Llama-SEA-LION-8B-IT* and *Gemma-SEA-LION-9B-IT*, attain competitive performance in SEA languages, with *Gemma-SEA-LION-9B-IT* achieving one of the highest average performances. Moreover, we significantly improve the performance of Llama-3.1-8B-Instruct from 41.48 to 61.84 using *Llama-SEA-LION-8B-IT*, while *Gemma-SEA-LION-9B-IT* achieves 14.02 improvement points compared to Gemma-2-9B-IT. Both *Llama-SEA-LION-8B-IT* and *Gemma-SEA-LION-9B-IT* outperform other SEA languages-focused LLMs, such as *Sailor2-8B-Chat* and *SEALLMs-v3-7B-Chat*, with an average score of 69.35 across all the languages covered by the SEA-HELM benchmark, apart from the SEA-MTBench tasks. This conforms with the previous results on the CPT models (Section 5.1) that our CPT model performs the best on SEA languages, resulting in the best performer in this experiment.

**SEACrowd**. Other than evaluating on some SEA languages like SEA-HELM, we also evaluated our model compared to competitors on 31 SEA indigenous languages using SEACrowd-NLU. Note that, for this study, we use only the best settings of our models from the previous experiment (Table 3). As shown in Table 4, we observe a state-of-the-art result from *Gemma-SEA-LION-9B-IT* by achieving 64.13 points on the NLU benchmark, while *Llama-SEA-LION-8B-IT* improves its baseline from 49.94 to 55.10 points. Moreover, the results from Figure 3 also emphasize the robustness of our model by reaching more than 80 points on this benchmark, while SeaLLMs and Llama-3.1 have only a few cases where the performance exceeds 80 points.

| | | | Open LLM Leaderboard | | | | |
|---|---|---|---|---|---|---|---|
| **Models** | **Average** | **MMLU-PRO** | **BBH** | **GPQA** | **MATH Lvl 5** | **IFEval (EN)** | **MUSR** |
| Sailor2-8B-Chat | 16.37 | 27.93 | 27.15 | 3.47 | 0.00 | 37.49 | 2.19 |
| SeaLLMs-v3-7B-Chat | 22.49 | 33.93 | 24.37 | 7.27 | 15.86 | 44.10 | 9.38 |
| Llama-3.1-8B-Instruct | 27.88 | 29.36 | 26.10 | 10.63 | 17.45 | 77.03 | 6.75 |
| Qwen2.5-7B-Instruct | 27.93 | 37.00 | 34.72 | 10.18 | 0.00 | 76.34 | 9.34 |
| Gemma-2-9B-IT | 28.86 | 31.95 | 42.14 | 14.77 | 0.23 | 74.36 | 9.74 |
| Stage-1-Llama | 24.51 | 25.87 | 26.32 | 7.83 | 19.26 | 62.89 | 4.88 |
| Stage-2-Llama | 27.75 | 28.10 | 24.64 | 7.72 | 19.56 | 78.78 | 7.74 |
| Merge-1-Llama | 27.49 | 27.47 | 26.22 | 8.28 | 19.79 | 76.16 | 7.04 |
| Merge-2-Llama | 29.96 | 29.92 | 28.78 | 9.96 | 19.94 | 82.61 | 8.54 |
| Aligned-SimPO-Llama | 30.58 | 30.84 | 34.31 | 8.39 | 26.59 | 75.76 | 7.61 |
| Llama-SEA-LION-8B-IT | 30.39 | 31.01 | 29.47 | 10.40 | 22.58 | 80.35 | 8.54 |
| Stage-1-Gemma | 29.88 | 33.34 | 38.51 | 10.74 | 24.17 | 56.87 | **15.66** |
| Stage-2-Gemma | 33.48 | 34.67 | 36.06 | 11.74 | 20.77 | **83.00** | 14.61 |
| Merge-1-Gemma | 35.15 | 36.22 | 41.42 | **15.32** | 26.28 | 82.09 | 9.59 |
| Aligned-SimPO-Gemma | 35.31 | **37.65** | 42.38 | 14.99 | **27.79** | 80.23 | 8.82 |
| Gemma-SEA-LION-9B-IT | **35.43** | 36.94 | **43.39** | 15.10 | 24.24 | 81.85 | 11.07 |

Table 5: Open LLM Leaderboard benchmarks across different instruct models of similar sizes.

These results emphasize the robustness of our models by achieving the state-of-the-art with a model parameter less than 10B on SEA benchmarks, including both traditional classical NLP benchmark (SEACrowd-NLU) and modern LLM benchmark (SEA-HELM).

**English performance**. We also evaluate the performance of a widely used language, English, to observe a difference between the results of SEA and English. The Open LLM Leaderboard performance is shown in Table 5. Both *Llama-SEA-LION-8B-IT* and *Gemma-SEA-LION-9B-IT* performed competitively in English language, math, and reasoning tasks, with *Gemma-SEA-LION-9B-IT* achieving the highest average score of 35.43. Moreover, we notice that the SEA models (Sailor and SeaLLMs) failed to perform on the English dataset. This might be because these models are optimized for SEA languages during supervised fine-tuning, and English performance decreased as a result. In contrast, our models balance the performance between SEA and English knowledge, resulting in a high score for all benchmarks.

### 5.3 Performance Analysis

In this study, we discuss the performance improvement in each design decision of our models (Tables 3 and 5) as follows.

**Stage 1: English instruction fine tuning** In Stage 1 IFT, the focus is predominantly on gaining general capabilities in math, code and general instruction following in the English language. Although our CPT models are based off of the instruct versions of *Llama-3.1-8B*, the CPT process has eroded the instruction following capabilities (See Table 5). We observe an increase of 3.86 and 9.72 for *Stage-1-Llama* and *Stage-1-Gemma* respectively in English instruction following capabilities on the IFEval benchmark. We also observe an average increase of 7.9 for *Stage-1-Llama* and 7.47 for *Stage-1-Gemma* for the SEA-HELM benchmark.

**Stage 2: Multilingual instruction fine tuning** In Stage 2 IFT, the focus is on multilingual and reasoning capabilities. By instruction fine tuning on SEA languages and higher complexity English instruction pairs, the Stage 2 models saw an average increase of 8.73 for *Stage-2-Llama* and 10.1 for *Stage-2-Gemma* over Stage 1 models on the SEA-HELM benchmark.

**Merge 1: Combining Stage 1 and Stage 2** Despite the significant gains observed in Stage 1 and 2, we observed that the effects of catastrophic forgetting from earlier stages could still be observed after Stage 2. In order to mitigate this, we merge Stage 1 and Stage 2 models into the CPT model, after which we we observed an average increase of 2.6 for *Merge-1-Gemma*. We also observed an increase across all SEA-HELM benchmark tasks for *Merge-1-Llama*.

**Merge 2: Incorporating instruct models** To reintroduce helpfulness, relevance and informativeness of responses observed in Llama 3.1 and Gemma 2 models, we perform further merges of open-source instruct models. While we observed significant increases in MT-Bench benchmark scores for Vietnamese and Thai, we also observed a slight degradation of average SEA-HELM performance as well as a slight degradation of Indonesian MT-Bench scores, which we view as acceptable trade-offs for the significant performance increases in Vietnamese and Thai.

**Alignment steps** In the alignment step to align the models to human preference, we prioritize the SEA MTBench performance over the other SEA-HELM benchmark tasks. We observed a broad increase in SEA MTBench performances across all languages for both models. However, this comes with minor degradation of instruction following capabilities and overall Indonesian SEA-HELM performance. The alignment step encourages longer, more helpful and sensitive responses but hurts performance on task-specific benchmarks and instruction following in some languages – an issue we address in the next step.

**Final merge: Combining aligned models** To compensate for the capability degradation in the previous steps, we merge *Merge-2-Llama* and *Merge-1-Gemma* with *Aligned-SimPO-Llama* and *Aligned-SimPO-Gemma* and various open sourced pre-trained models describe in sections 3.2.1 and 3.2.2 for their respective model families. For *Llama-SEA-LION-8B-IT*, we observed a significant increase in average SEA-HELM performance (61.84) from the alignment stage (51.30), mainly from the increase in performance for the core tasks in SEA-HELM. This performance increase demonstrates the value of empirical selection of pre-trained models to be merged in based on each model's strengths and weaknesses to produce a far superior model. For *Gemma-SEA-LION-9B-IT*, it easily achieves higher performance compared to the *Llama-SEA-LION-8B-IT* with fewer post training steps. We attribute this performance to the high performance of the base Gemma 2 model and also to the larger vocabulary size which have been demonstrated (Takase et al., 2024) to produce better models.

## 6 Related Works

Recently, researchers have proposed large language models that support multilingual settings. Llama (Dubey et al., 2024) is the prior effort to release an open-source large language model for the research community to develop their own models. Then, Qwen (Yang et al., 2024a) and Gemma (Rivière et al., 2024) introduced open-source LLMs that perform comparably or better than Llama with a larger amount of training data and many supported languages for these recent models. Massively multilingual open-source models like Bloom (Scao et al., 2022) and Aya (Üstün et al., 2024) also support a very wide range of languages, including some SEA languages. Although these models demonstrate a robust performance in English benchmarks, they mostly underperformed on SEA benchmarks that tested for SEA languages, SEA knowledge and cultural understanding (Lovenia et al., 2024; Susanto et al., 2025), presumably due to a lack of language support for certain SEA languages or cultures.

In the SEA community, many works propose a large language model that is designed specifically for SEA languages by adding more SEA tokens in the training process, such as SeaLLMs (Nguyen et al., 2024) and Sailor (Sailor2 Team, 2024). However, the performance of these models is robust only on in-domain datasets or favors only some tasks (i.e., classical NLP datasets). This is because the design choice in the pre-training or fine-tuning of these models is not well studied, e.g., performing a single SFT step with low-quality datasets written in some SEA languages, resulting in a slight improvement on SEA benchmarks. To create a robust SEA LLM, we need to carefully balance language representation and design both pre-training and post-training (i.e., SFT, alignment, and model merging) for SEA contexts.

## 7 Conclusion

Despite the sizable population and language diversity in Southeast Asia, there remains a scarcity of resources and accurate linguistic and cultural representation with open-source LLMs. In this paper, we introduce *Llama-SEA-LION-8B-IT* and *Gemma-SEA-LION-9B-IT*, two multilingual LLMs comprehensively trained to achieve state-of-the-art performances in SEA languages, based on the Llama and Gemma family of LLMs. SEA-LION represents the next advancement in the development of LLMs that explicitly supports SEA languages. Both models are fully open-source and available for commercial use to increase accessibility and innovation in multilingual LLMs in Southeast Asia. We will make our resources publicly available — including the dataset, training scripts, training checkpoints, and all fine-tuned models, even those that achieve state-of-the-art performance on the benchmarks — to establish solid baselines, ensure reproducibility, and support future research focused on culturally and professionally relevant SEA applications.

## Limitation

Although we propose the state-of-the-art SEA LLMs, we found that the benchmark might not cover all the properties and languages we want to evaluate. For example, SEA-HELM is a robustness benchmark, but only covers four languages. SEACrowd is a benchmark that covers all SEA languages, but it is only classical NLP datasets (no chat or instruction following datasets). We require a more holistic SEA benchmark that covers LLM-specific tasks written in all SEA languages. However, with the current evaluation design choice, these benchmarks are the best design choice for current SEA research works.

Moreover, we conduct experiments using only 8 and 9 billion parameter models. We argue that this is the most commonly used model size in real-world scenarios. In addition, our method can and should also work with a higher or smaller model size since our proposed technique does not rely on the model size, as we demonstrated by applying the SFT and alignment techniques on both Llama and Gemma models.

## References

SCB 10X, VISTEC, and SEACrowd. 2024. Thai llm leaderboard.

AI Singapore AI Products Team. 2024. Sea-helm.

AISG AI Singapore. 2023. Sea-lion-pile.

AISG AI Singapore. 2025. Sea-lion-pile-v2.

Anton Alexandrov, Veselin Raychev, Mark Niklas Mueller, Ce Zhang, Martin Vechev, and Kristina Toutanova. 2024. Mitigating catastrophic forgetting in language transfer via model merging. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17167–17186, Miami, Florida, USA. Association for Computational Linguistics.

Arcee-AI. 2024. Llama-3.1-supernova-lite.

Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131. Association for Computational Linguistics.

BAAI Beijing Academy of Artificial Intelligence. 2024. Infinity instruct.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

CommonCrawl. 2024. Commoncrawl.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. ULTRAFEEDBACK: boosting language models with scaled AI feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

DAMO-NLP-SG. 2024. Seaexam.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Wikimedia Foundation. 2024. Wikimedia enterprise html dumps downloads.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's mergekit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: EMNLP 2024 - Industry Track, Miami, Florida, USA, November 12-16, 2024*, pages 477–485. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, and Yanning Chen. 2024. Liger kernel: Efficient triton kernels for llm training. *arXiv preprint arXiv:2410.10989*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhen Leng Thai, Kai Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *CoRR*, abs/2404.06395.

HuggingFace. 2024. Open llm leaderboard.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William-Chandra Tjhi. 2023. BHASA: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *CoRR*, abs/2309.06085.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Jann Montalan, Ryan Hadiwijaya, Joanito Agili Lopo, William Nixon, Börje Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus Irawan, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Ngee Tai Chia, Ayu Purwarianti, Sebastian Ruder, William-Chandra Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng Xin Yong, and Samuel Cahyawijaya. 2024. Seacrowd: A multilingual multimodal data hub and benchmark suite for southeast asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5155–5203. Association for Computational Linguistics.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian J. McAuley, Han Hu, Torsten Scholak, Sébastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, and et al. 2024. Starcoder 2 and the stack v2: The next generation. *CoRR*, abs/2402.19173.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *CoRR*, abs/2405.14734.

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, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia.

In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

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

Guilherme Penedo, Hynek Kydlícek, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *CoRR*, abs/2406.17557.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. Bpe-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1882–1892. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Sailor2 Team. 2024. Sailor2: Sailing in south-east asia with inclusive multilingual llm.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

AI Singapore SEA-LION Team. 2024. Llama3 8b cpt sea-lionv2.1 instruct.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xian Bin Yong, Weiqi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. Sea-helm: Southeast asian holistic evaluation of language models. *Preprint*, arXiv:2502.14301.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13003–13051. Association for Computational Linguistics.

Sho Takase, Ryokan Ri, Shun Kiyono, and Takuya Kato. 2024. Large vocabulary size improves large language models. *CoRR*, abs/2406.16508.

Sailor Team. 2024. Sailor2: Sailing in south-east asia with inclusive multilingual llms.

The Mosaic ML Team. 2021. composer. `https://github.com/mosaicml/composer/`.

The Mosaic ML Team. 2022. Llm foundry. `https://github.com/mosaicml/llm-foundry`.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. 2024. Openmathinstruct-2: Accelerating AI for math with massive open-source instruction data. *CoRR*, abs/2410.01560.

Ahmet Üstün, Viraat Aryabumi, Zheng Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15894–15939. Association for Computational Linguistics.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024a. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 370–390. Association for Computational Linguistics.

Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jiménez, François Fleuret, and Pascal Frossard. 2024b. Localizing task information for improved model merging and compression. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024c. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *CoRR*, abs/2406.01574.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15366–15394. Association for Computational Linguistics.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie E. Everett, Alexander A. Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. 2024. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *CoRR*, abs/2406.08464.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 technical report. *CoRR*, abs/2407.10671.

Ziyi Yang, Fanqi Wan, Longguang Zhong, Tianyuan Shi, and Xiaojun Quan. 2024b. Weighted-reward preference optimization for implicit model fusion. *CoRR*, abs/2412.03187.

Wei Jie Yeo, Teddy Ferdinan, Przemyslaw Kazienko, Ranjan Satapathy, and Erik Cambria. 2024. Self-training large language models through knowledge detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 15033–15045. Association for Computational Linguistics.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024a. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. *CoRR*, abs/2407.19672.

Xulang Zhang, Rui Mao, and Erik Cambria. 2024b. Multilingual emotion recognition: Discovering the variations of lexical semantics between languages. In *International Joint Conference on Neural Networks, IJCNN 2024, Yokohama, Japan, June 30 - July 5, 2024*, pages 1–9. IEEE.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch FSDP: experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860.

Wenzhen Zheng, Wenbo Pan, Xu Xu, Libo Qin, Li Yue, and Ming Zhou. 2024a. Breaking language barriers: Cross-lingual continual pre-training at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7725–7738. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024b. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in? *CoRR*, abs/2408.10811.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911.

# A Appendix

## A.1 Continued pre-training (CPT) data

**Existing data**: We utilize existing datasets as shown in Table 6 (HuggingFace Datasets).
**Other data**: As shown in Table 6 (the other data section), the listed datasets contain data from a diverse range of domains, including news, books, articles, poems, etc.

| Continued Pre-training Data | | |
|---|---|---|
| **Source (HuggingFace Datasets)** | **Languages** | **Size (Billions of Tokens)** |
| bigcode/the-stack-v2-dedup | CODE | 40 |
| allenai/dolma | EN | 37.5 |
| HuggingFaceFW/fineweb-edu | EN | 7.5 |
| aisingapore/SEA-PILE-v1 | SEA | 47.58 |
| aisingapore/SEA-PILE-v2 | ID | 7 |
| **Source (Others)** | **Languages** | **Size (Billions of Tokens)** |
| VinBigData | VI | 16 |
| WangChanBERTa | TH | 8.5 |
| Others - EN | EN | 5 |
| Others - SEA | SEA | 30.92 |

Table 6: List of datasets for the continued pre-training stage.

## A.2 Stage 1 IFT data

| Stage 1 IFT Datasets | | |
|---|---|---|
| **Source (HuggingFace Datasets)** | **Languages** | **Size** |
| BAAI/Infinity-Instruct | EN | 7,449,106 |
| nvidia/OpenMathInstruct-2 | EN | 2,000,000 |

Table 7: List of datasets for Stage-1-IFT. For BAAI/Infinity-Instruct dataset, any conversation that originally ended with a user turn has had that last turn removed.

## A.3 Stage 2 IFT data

**Existing data**: We utilize existing datasets as shown in Table 9 (HuggingFace Datasets).
**Synthetic data**: As shown in Table 9 (the generated part), we describe how to formulate synthetic data as follows

- qwen_gemma_synthetic datasets are generated first in English with Qwen 32B, utilizing an approach similar to Magpie. Instructions are then translated into the target language with Gemma 2 27B.
- llama_gemma_synthetic datasets are generated first in English with Llama 3.1 70B, utilizing an approach similar to Magpie (Xu et al., 2024). Instructions are then translated into the target language with Gemma 2 27B.
- gemma_synthetic datasets are generated directly with Gemma 2 27B using Magpie (Xu et al., 2024).
- sea_multilingual_systemchat is a synthetic dataset translated with Gemma 2 27B from the English systemchat dataset.
- rewritten_oasst is a dataset rewritten with Gemma 2 27B based on the English OASST dataset.
- rewritten_helpsteer is a dataset rewritten with Gemma 2 27B based on the English Helpsteer dataset.

### A.4 Helpfulness and preference alignment data

As shown in Table 8, we use the princeton-nlp/gemma2-ultrafeedback-armorm as the source of the alignment data. We then further re-scored with the reward model, nvidia/Llama-3.1-Nemotron-70B-Reward to create the SEA version. In particular, generated-gemma2-27b-seapref-nemotron-70b takes prompts from seald, wangchan_thaiinstruct, and additional hand-written Southeast Asian cultural prompts collected from native speakers and then generates responses (with a varying temperature) from them with Gemma 2 27B. The responses are then scored with nvidia/Llama-3.1-Nemotron-70B-Reward, with the top-scoring response selected as chosen and vice versa, similar to princeton-nlp/gemma2-ultrafeedback-armorm.

**Preference Data**

| **Source (HuggingFace Datasets)** | **Languages** | **Size** |
|---|---|---|
| princeton-nlp/gemma2-ultrafeedback-armorm | EN | 61,510 |
| **Source (Generated)** | **Languages** | **Size** |
| generated-gemma2-27b-seapref-nemotron-70b | SEA | 5,511 |

Table 8: List of preference datasets used for the alignment stage.

**Stage 2 IFT Datasets**

| **Source (HuggingFace Datasets)** | **Languages** | **Size** |
|---|---|---|
| BAAI/Infinity-Instruct^* | EN | 1,456,927 |
| HuggingFaceTB/smoltalk | EN | 409,537 |
| allenai/tulu-3-sft-personas-math | EN | 149,960 |
| parinzee/seed-free-synthetic-instruct-thai-v1 | TH | 118,898 |
| HuggingFaceTB/smoltalk | EN | 96,356 |
| HuggingFaceTB/smoltalk | EN | 83,144 |
| arcee-ai/EvolKit-75K | EN | 74,174 |
| AI-MO/NuminaMath-TIR | EN | 72,441 |
| Post-training-Data-Flywheel/AutoIF-instruct-61k | EN | 61,492 |
| argilla/ifeval-like-data | EN | 56,339 |
| HuggingFaceTB/smoltalk | EN | 53,342 |
| ai2-adapt-dev/tulu_v3.9_wildjailbreak_decontaminated_50k | EN | 50,000 |
| ai2-adapt-dev/tulu_v3.9_synthetic_finalresp_wildguardmixtrain_decontaminated_50k | EN | 50,000 |
| allenai/tulu-3-sft-personas-math-grade | EN | 49,980 |
| allenai/tulu-3-sft-personas-code | EN | 34,999 |
| HuggingFaceTB/smoltalk | EN | 34,424 |
| allenai/tulu-3-sft-personas-instruction-following | EN | 29,980 |
| airesearch/WangchanThaiInstruct | TH | 25,014 |
| allenai/tulu-3-sft-personas-algebra | EN | 20,000 |
| arcee-ai/EvolKit-20k-vi | VI | 15,378 |
| allenai/coconot | EN | 10,983 |
| ai2-adapt-dev/tulu_v3.9_sciriff_10k | EN | 10,000 |
| **Source (Generated)** | **Languages** | **Size** |
| qwen_gemma_synthetic_tamil | TA | 480,000 |
| qwen_gemma_synthetic_thai | TH | 480,000 |
| qwen_gemma_synthetic_indonesian | ID | 465,019 |
| qwen_gemma_synthetic_vietnamese | VI | 465,019 |
| gemma_synthetic_indonesian | ID | 458,149 |
| gemma_synthetic_filipino | TL | 455,093 |
| gemma_synthetic_viet | VI | 291,576 |
| gemma_synthetic_tamil | TA | 276,314 |
| gemma_synthetic_thai | TH | 186,339 |
| gemma_synthetic_javanese | JV | 110,000 |
| gemma_synthetic_sudanese | SU | 110,000 |
| llama_gemma_synthetic_thai | TH | 88,920 |
| llama_gemma_synthetic_tamil | TA | 88,920 |
| llama_gemma_synthetic_vietnamese | VI | 88,920 |
| llama_gemma_synthetic_javanese | JV | 88,920 |
| llama_gemma_synthetic_indonesian | ID | 88,920 |
| llama_gemma_synthetic_filipino | TL | 80,000 |
| enrich_27k | SEA | 27,463 |
| sea_multilingual_systemchat | SEA | 1,903 |
| rewritten_oasst | SEA | 841 |
| rewritten_helpsteer | SEA | 838 |

Table 9: List of datasets for Stage-2-IFT.