sea | AI Lab

connecting the dots

Sailor: Open Language Models for South-East Asia

Presenter: Longxu Dou

Team: Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Ziqi Jin, Xin Mao, Wei Lu, Min Lin

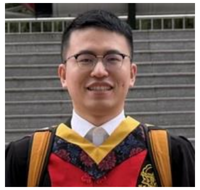
Introduction

Sailor is a suit of LLMs that perform well across the SEA region, encompassing a range of languages including **Vietnamese, Thai, Indonesian, Malay**, and **Lao**.

SAILOR: Open Language Models for South-East Asia

built by

sea connecting the dots AI Lab × SUTD SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

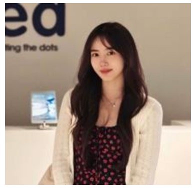


Longxu Dou
Sea AI Lab

Qian Liu
Sea AI Lab

Guangtao Zeng
SUTD

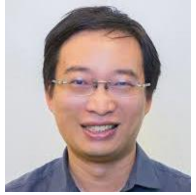
Jia Guo
NUS



Jiahui Zhou
Sea AI Lab

Ziqi Jin
SUTD

Xin Mao
NTU

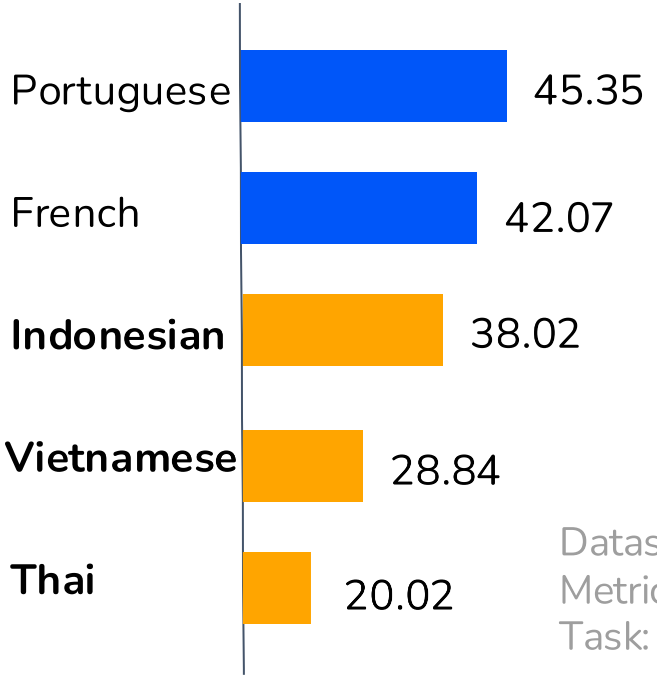


Wei Lu
SUTD

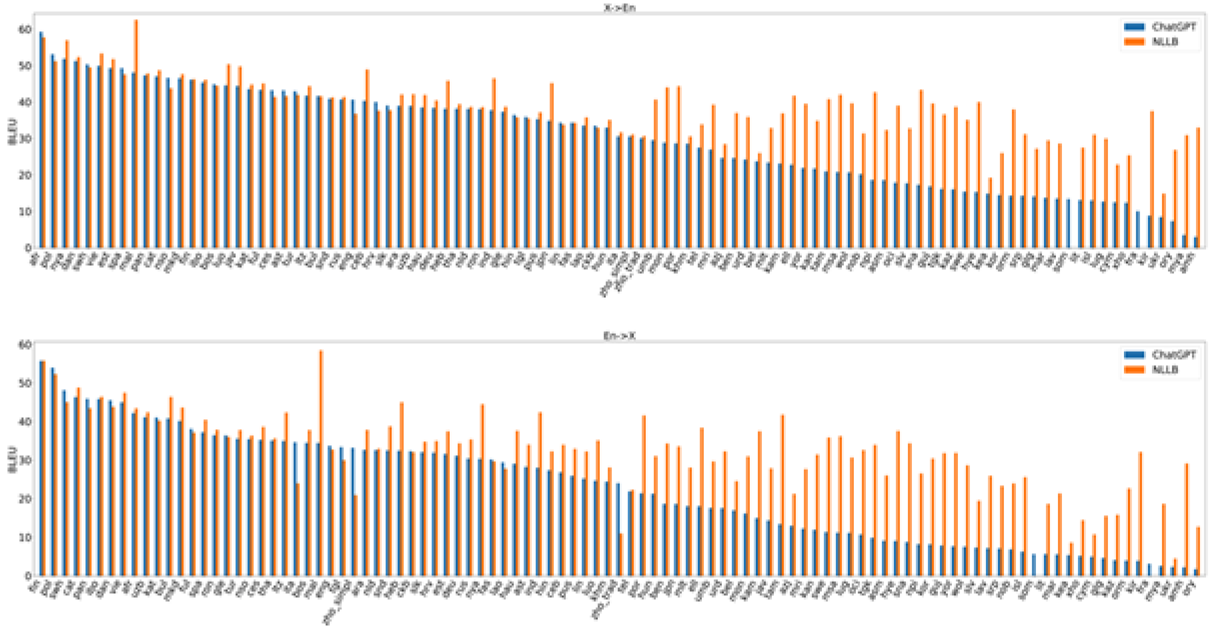
Min Lin
Sea AI Lab

Background

Even for GPT-4, it struggles with SEA languages compared to European languages and lags the supervised baseline (NLLB) a lot.



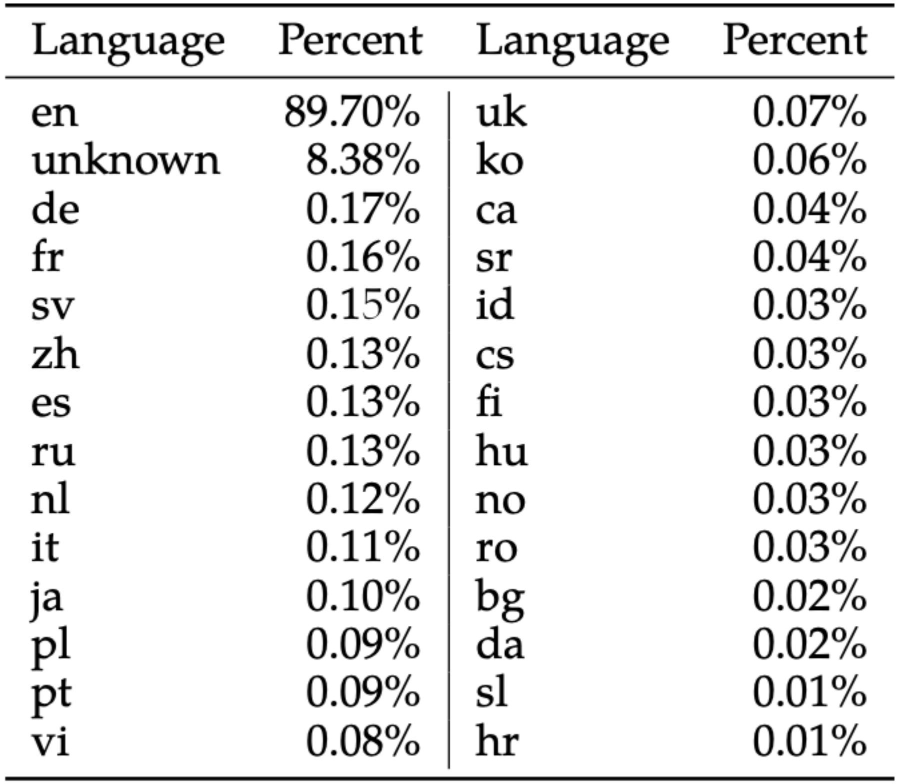
Dataset: FLORES-200
Metric: chrF++
Task: Machine Translation



Chain-of-Dictionary Prompting Elicits Translation in Large Language Models. Lu et al.

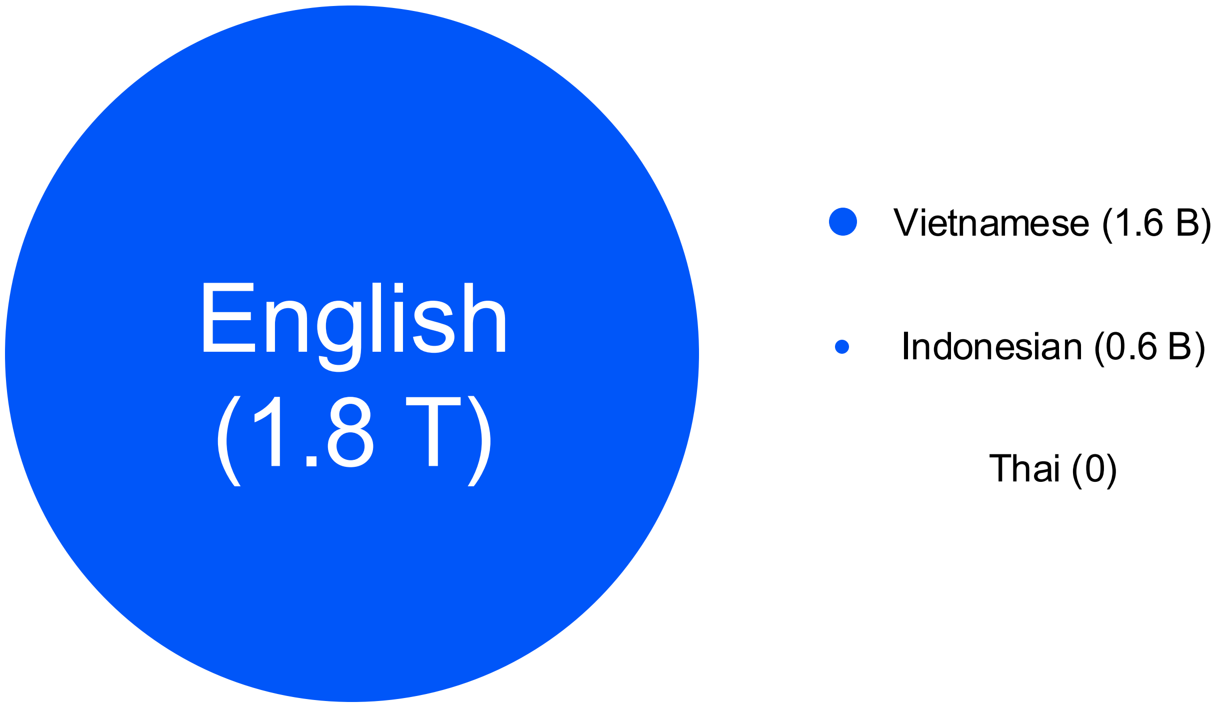
Background

Developing high-quality models crucially depends on access to a large-scale and high-quality dataset, where **English content** has established it as a preeminent source.



Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

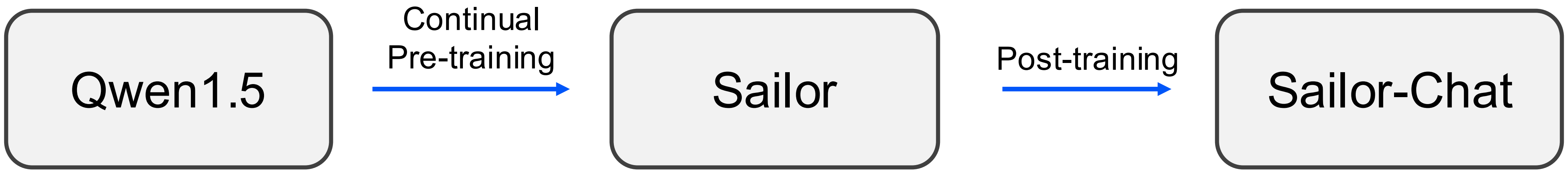
The language proportion of Llama-2 pre-training



Overview

- **Team**: 7 members x 4 months

- **Roadmap**: continual pre-training from Qwen 1.5
  - Qwen has a multilingual friendly vocabulary.
  - Reuse the flops used for the base model.



Qwen1.5

Continual Pre-training

Sailor

Post-training

Sailor-Chat

Overview

- **Model**: 0.5B, 1.8B, 4B, 7B, 14B (**150K Downloads** since March 2024)

- **Data**: **200B clean tokens** (some repeated more than 1 epoch), including **140B SEA related tokens**, all from publicly available corpus.

- **Total Computation**: **128** A100-40G GPUs for **28 days**

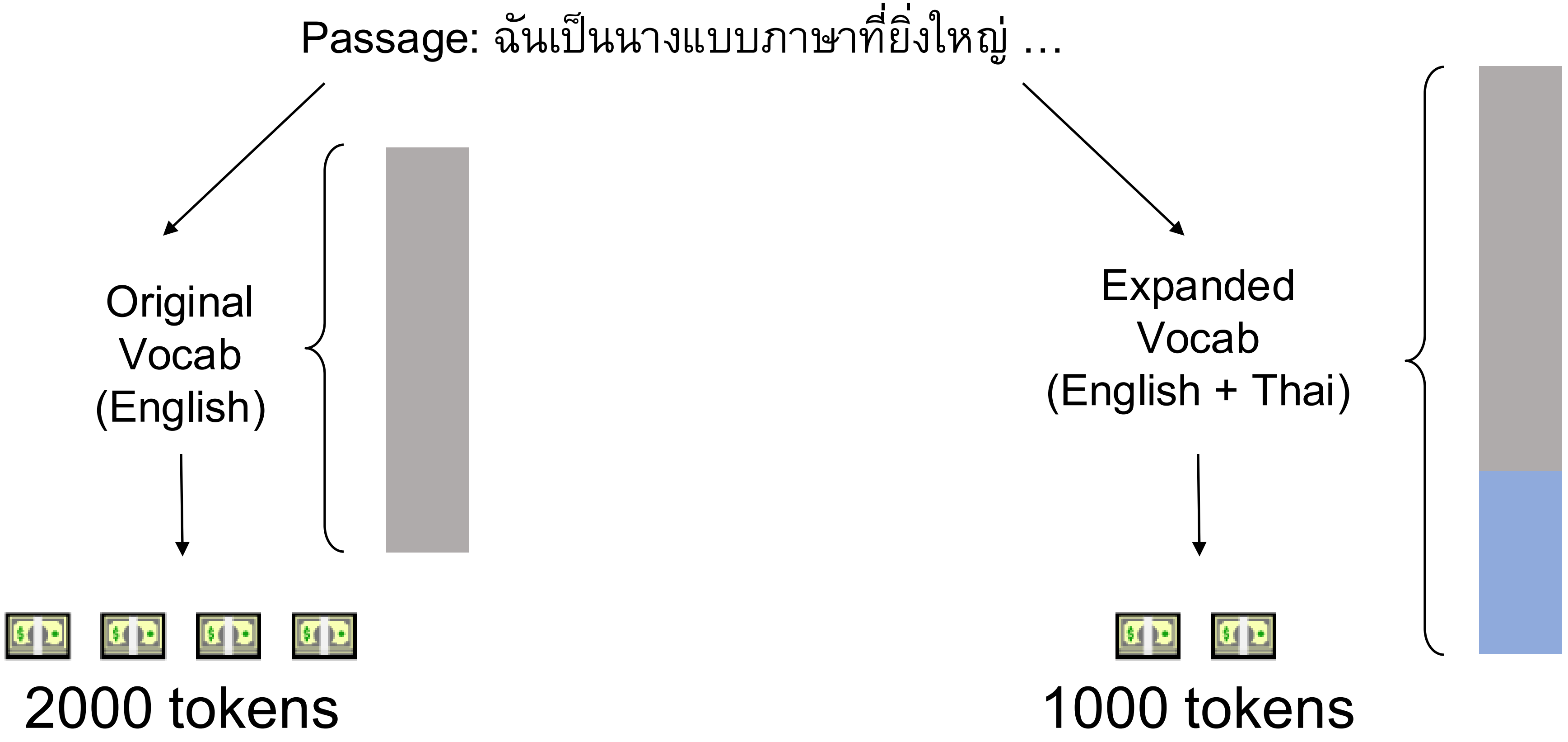
- **Open Language Model**:
  - Release the **base model** and the **chat model**.
  - Open-source the data processing pipeline, the fine-tuning code, and the evaluation code.

The Journey of Sailor

- The Lesson of Vocabulary Expansion
- Data Curation and Processing
- Data Mixture for Language Balance
- Continual Pre-training and Forgetting
- Other Training Tricks

The Lesson of Vocabulary Expansion

We have tried our best to do vocabulary expansion on models like Mistral. However, **it is challenging** to expand the vocabulary with maintaining the original performance.



Passage: ฉันเป็นนางแบบภาษาที่ยิ่งใหญ่ …

Original Vocab (English)

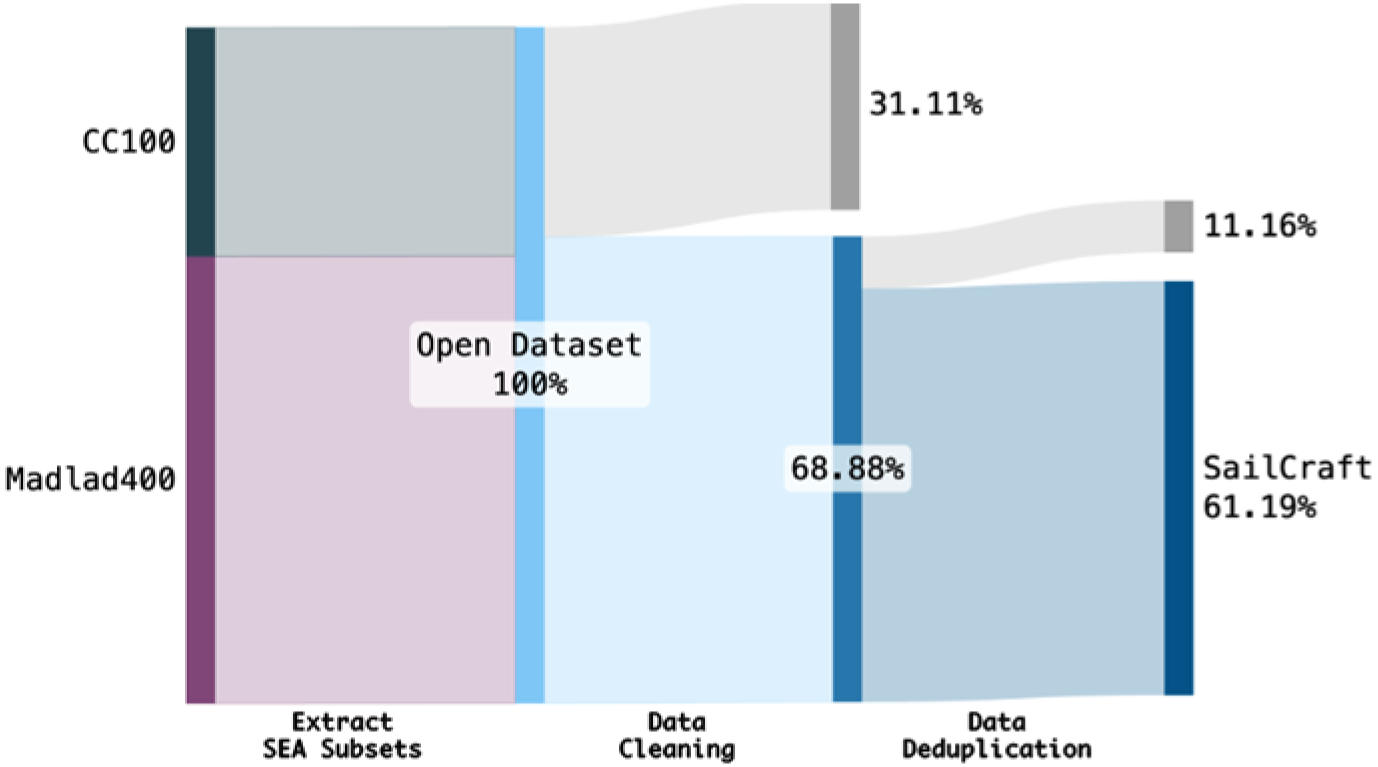
2000 tokens

Expanded Vocab (English + Thai)

1000 tokens

Data Curation and Processing

Pre-training requires a high-quality corpus, which is even more crucial for continual pre-training. Despite efforts, we found that publicly available corpora still **exhibit problems**.



31.11% are removed by Cleaning

11.16% are removed by Deduplication

Data Mixture for Language Balance

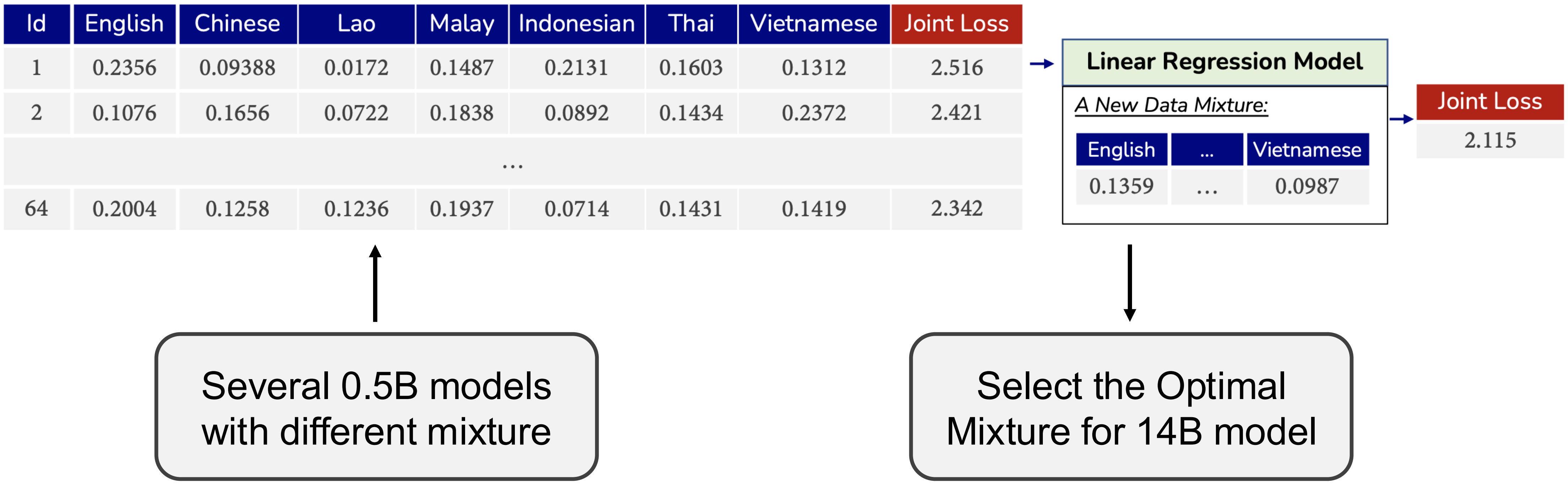
We aim to develop an improved LLM tailored for the entire SEA region, with a focus on ensuring balanced representation across all target languages.

Language	Effective Tokens (Billion)	Epoch
Indonesian (id)	51.56	0.74
Malay (ms)	7.91	1.44
Thai (th)	38.24	1.28
Vietnamese (vi)	41.50	0.66
Lao (lo)	0.34	0.97
English (en)	37.2	-
Chinese (zh)	22.64	-

While there are more corpus available, we choose not to utilize them all to balance the language performance

Data Mixture for Language Balance

We have developed a novel algorithm that determines **balanced mixture for different languages** during continual pre-training by conducting randomized data mixture experiments at a fixed learning rate, aiming to identify the most effective data mixture.



Id	English	Chinese	Lao	Malay	Indonesian	Thai	Vietnamese	Joint Loss
1	0.2356	0.09388	0.0172	0.1487	0.2131	0.1603	0.1312	2.516
2	0.1076	0.1656	0.0722	0.1838	0.0892	0.1434	0.2372	2.421
…								
64	0.2004	0.1258	0.1236	0.1937	0.0714	0.1431	0.1419	2.342

Linear Regression Model

A New Data Mixture:

English	...	Vietnamese
0.1359	…	0.0987

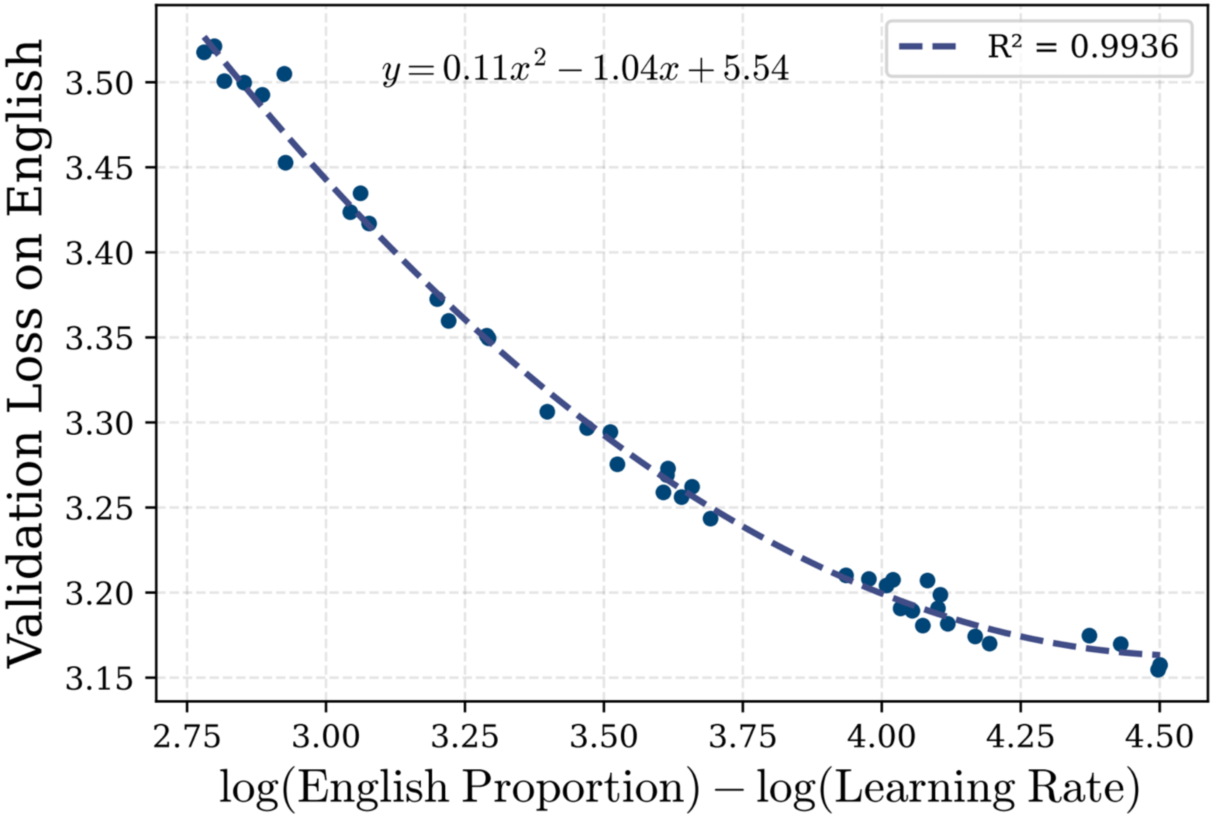
Joint Loss
2.115

Several 0.5B models with different mixture

Select the Optimal Mixture for 14B model

Continual Pre-training and Forgetting

Under the same token budget, we observe the validation loss on English can be modelled as a **quadratic function** of log(English Proportion) − log(Learning Rate).



Other Training Tricks: Code Switching

Code-switching refers to the phenomenon where different languages are used within the same context. We found that **doc-level code switching** is rather useful.

Word-Level Code Switching

Sea AI Lab (SAIL) เป็นองค์กรที่ตั้งใจจะ
พัฒนาเทคโนโลยีเพื่อช่วยขับเคลื่อน
เศรษฐกิจดิจิทัลในภูมิภาคนี้ มัน (It) เน้นที่
การสำรวจและพัฒนามุมมองและเทคโนโลยี
ระยะยาวที่เกี่ยวข้องกับธุรกิจปัจจุบันของ
Sea และโอกาสใหม่ ๆdนอกเหนือจากนี้

✗

Doc-Level Code Switching

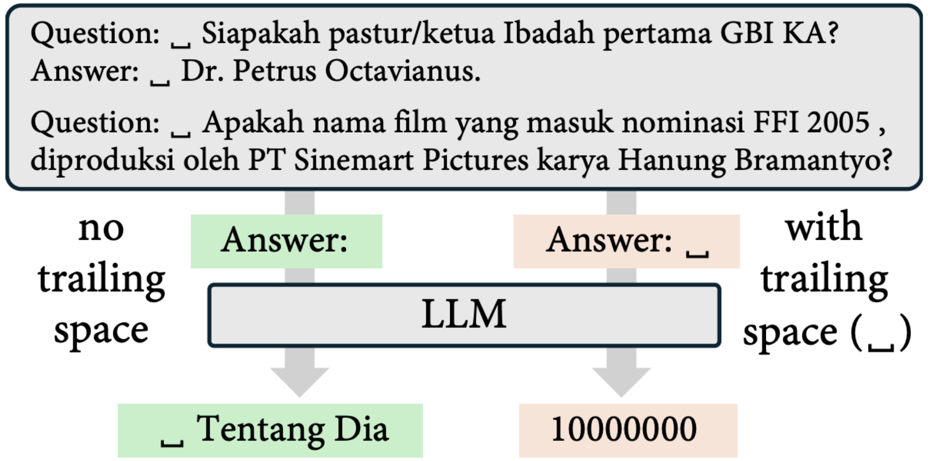
Sea AI Lab (SAIL) is committed to advancing technology to drive the development of the digital economy across our regions.

มุ่งเน้นไปที่การสำรวจและพัฒนาข้อมูลเชิงลึกและเทคโนโลยีในระยะยาวที่เกี่ยวข้อง
กับธุรกิจที่มีอยู่ของ Sea และโอกาสใหม่ ๆ นอกเหนือจากนั้น นอกจากนี้ยังมี
จุดมุ่งหมายเพื่อดึงดูดและร่วมมือกับผู้ที่มีความสามารถระดับสูงในด้าน
ปัญญาประดิษฐ์

✓

Other Training Tricks: BPE Dropout

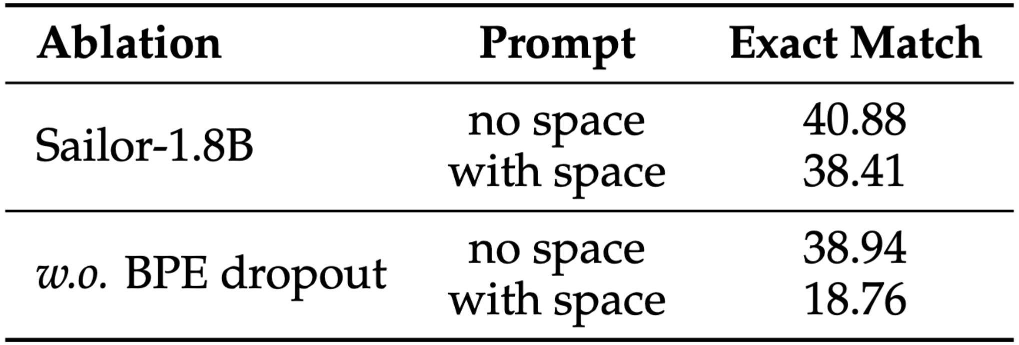
The initial Sailor models were trained on 200B tokens using the greedy tokenization strategy. Subsequently, we train them using BPE dropout on another 2B tokens, and the BPE dropout technique improved the models' robustness a lot.



Question: ␣ Siapakah pastur/ketua Ibadah pertama GBI KA?
Answer: ␣ Dr. Petrus Octavianus.

Question: ␣ Apakah nama film yang masuk nominasi FFI 2005 , diproduksi oleh PT Sinemart Pictures karya Hanung Bramantyo?

(a) Minor variations in prompts such as a trailing space visualized by ␣ can drastically change the prediction of LLMs.

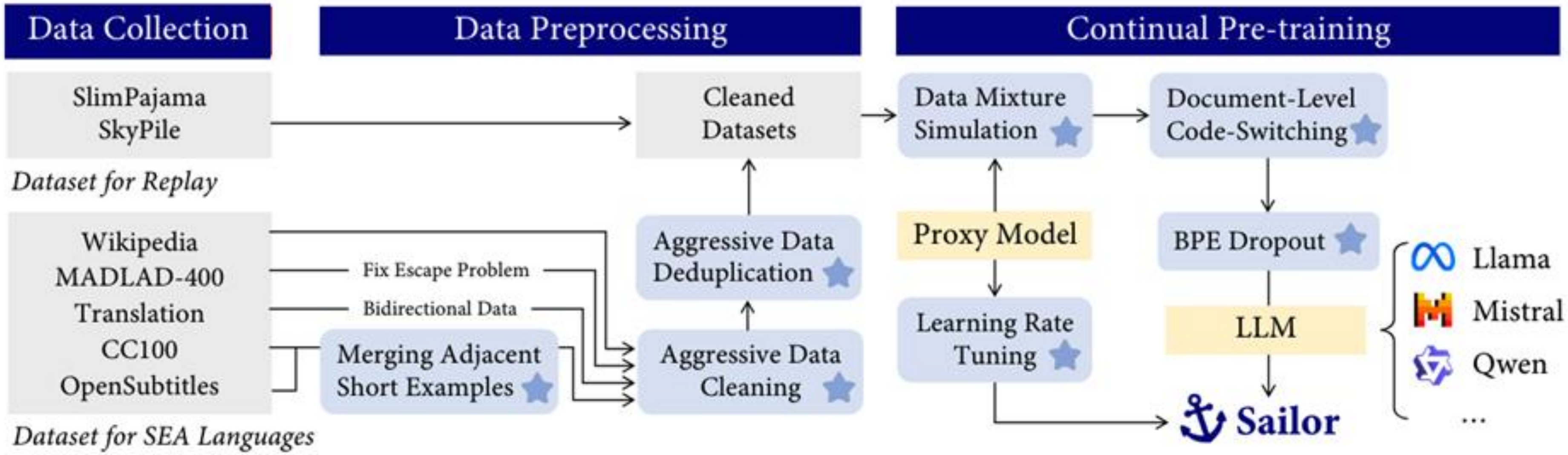


Ablation	Prompt	Exact Match
Sailor-1.8B	no space with space	40.88 38.41
w.o. BPE dropout	no space with space	38.94 18.76

(b) Experimental results on the TydiQA dataset indicate that applying BPE dropout significantly enhances the robustness of the Sailor-1.8B model when handling trailing spaces.

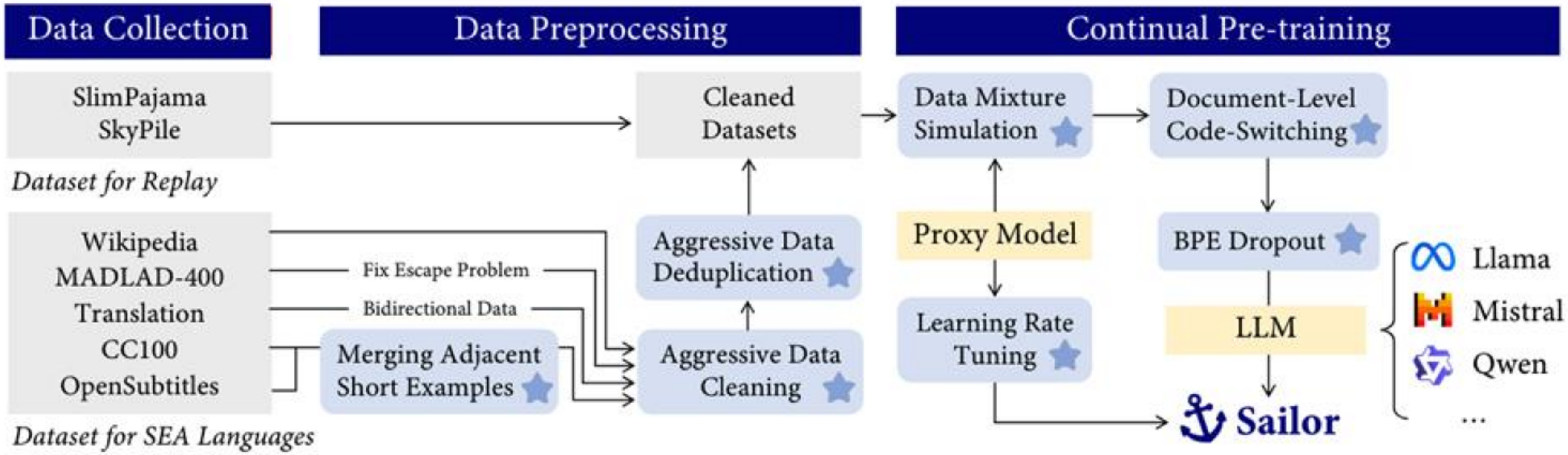
Takeaway

- **Data**:
  - **Aggressive Data Deduplication and Cleaning**: high-quality dataset yields a good model, and we have massive efforts to do aggressive data deduplication and data cleaning (https://github.com/sail-sg/sailcraft).
  - **Merging Adjacent Short Examples**: we also built a website-friendly crawler network specifically tailored for Southeast Asia, actively gathering data focused on the region.



Takeaway

- **Training**:
  - **BPE dropout**: we found that the model is sensitive to the trailing space, which is also observed by other multilingual language models. We apply BPE dropout to alleviate the problem.
  - **Learning curve**: to mitigate catastrophic forgetting, we analyse the learning curve about the English ratio and learning rate, predicting the optimal setting for our pre-training.
  - **Language balance**: to amalgamate data sources from diverse languages, we have devised a novel algorithm designed to automatically determine the optimal mixture for training.



Thank You