Sailor2: Sailing in South-East Asia with Inclusive Multilingual LLMs

Presenter: Longxu Dou

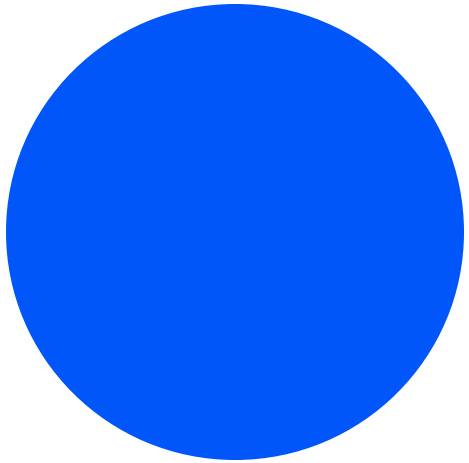
Team: Qian Liu, Fan Zhou, Changyu Chen, Ziqi Jin, Zichen Liu, Sailor2 Community Members

Sailor2: More Data, More Language

We have extracted SEA language corpora from 92 snapshots of CommonCrawl, initially consuming 250TB of disk, which may contain **800B high-quality** tokens at most.

Sailor
(140B SEA Tokens)

Sailor2
(800B SEA Tokens)



Sailor2: Team Member (SAIL and Community)

sea AI Lab | Hugging Face

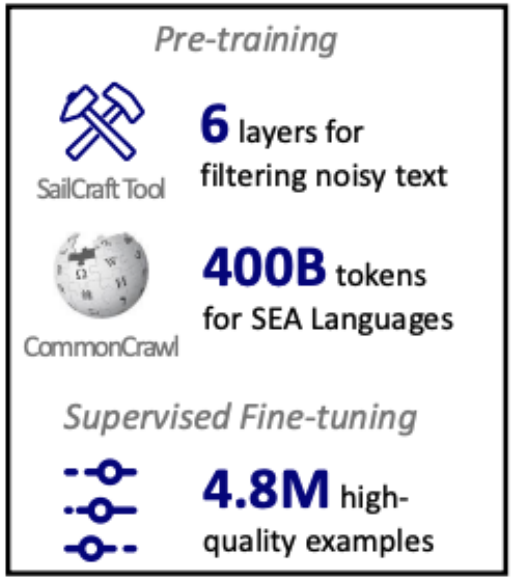
SCB 10X | WISESIGHT

Sailor2 Community

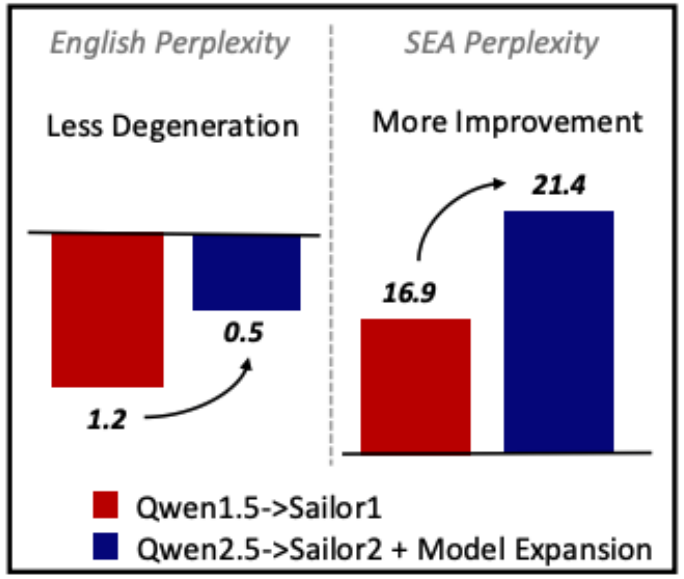
SAILOR2: Sailing in South-East Asia with Inclusive Multilingual LLMs

Sailor2: New Milestone of Open SEA Language Model

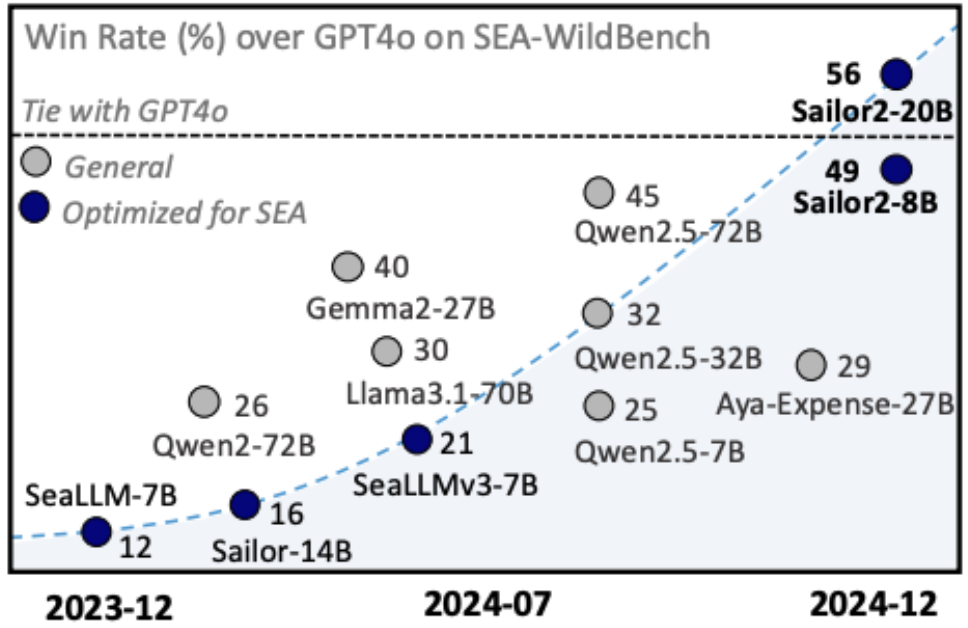
Data Curation



Model Expansion



Open Models for SEA Languages



Sailor2-20B achieves the **50-50 win rate over GPT4o** on SEA languages.

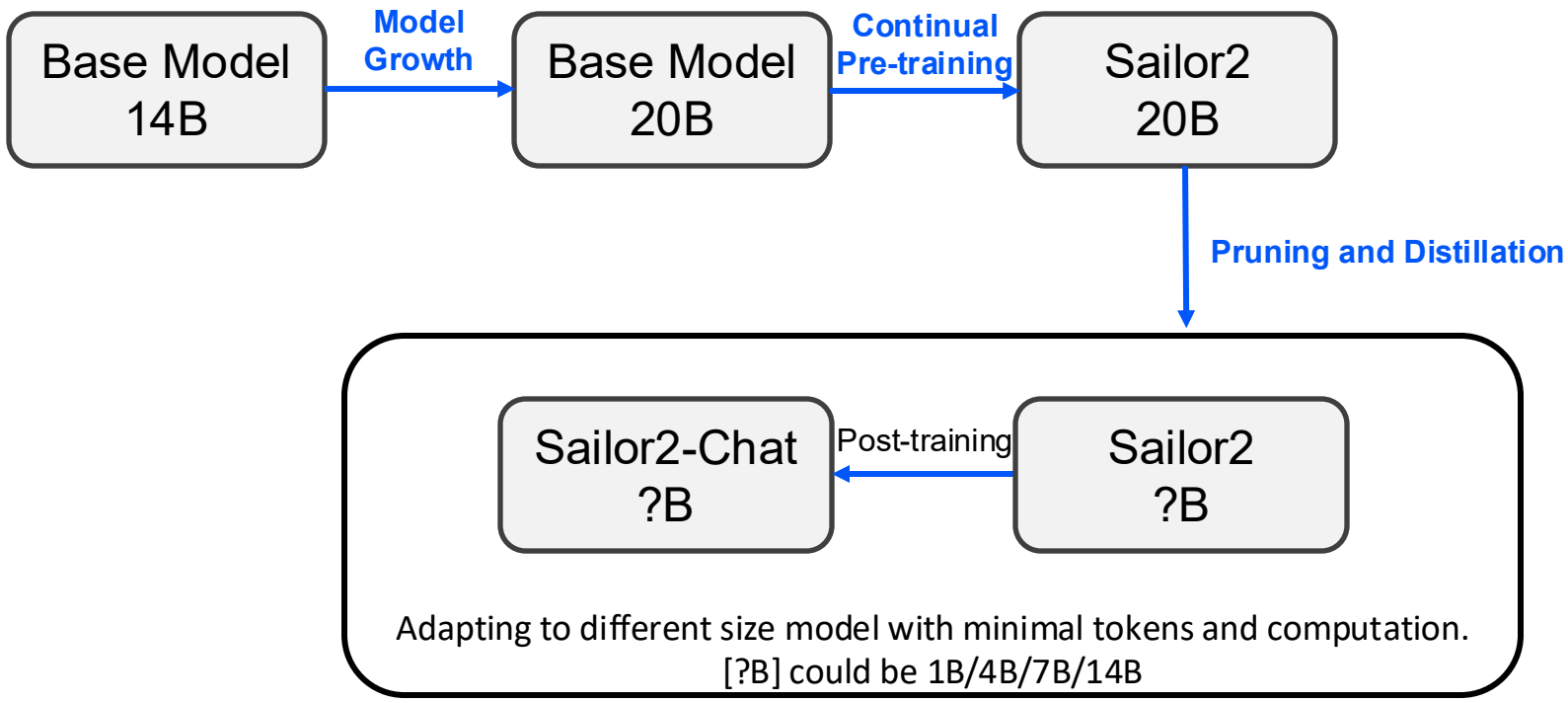
Sailor2: More Supported Languages

Language	ISO Code	Country/Region	No. of Speakers
Indonesian	ind	Indonesia	268 million
Vietnamese	vie	Vietnam	96 million
Javanese	jav	Indonesia (Java island)	82 million
Thai	tha	Thailand	70 million
Burmese	mya	Myanmar	54 million
Sundanese	sun	Indonesia (West Java)	42 million
Malay	zsm	Malaysia, Brunei, Singapore	33 million
Tagalog	tgl	Philippines (Luzon)	28 million
Cebuano	ceb	Philippines (Cebu, Mindanao)	21 million
Khmer	khm	Cambodia	16 million
Ilocano	ilo	Philippines (Northern Luzon)	8 million
Lao	lao	Laos	7 million
Waray	war	Philippines (Eastern Visayas)	3 million

Sailor2 support **13 SEA languages, Chinese and English**.

Sailor2: Training Roadmap

Goal: Training a very large model once to derive multiple smaller models efficiently.



Sailor2: Overview

- **Team**: 16 core members x 6 months
- **Model**: 1B, 8B, 20B

- **Data**: **500B clean tokens** (some repeated more than 1 epoch), including **400B SEA related tokens**, from CommonCrawl and Open Source.

- **Total Computation in Continual Pre-training**: **256** H100 GPUs for **30 days (Community)**
- **Total Computation in Post-training**: **64** A100-80G GPUs for **7 days (SAIL)**

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

Sailor2 Base Model Evaluation

Compared to other advanced multilingual models, Sailor2 demonstrates **comparable or better performance**, especially in **extremely low-resource languages** like Javanese.

Language	Benchmark$_{(eval)}$	Sailor2-8B	Qwen2.5-7B	Gemma2-9B	Lllama3.1-8B	SeaLLM-v3-7B	Sailor2-20B	Qwen2.5-32B	Gemma2-27B	Llama3.1-70B	Aya-Expanse-32B
	Avg.	**57.6**	52.8	52.5	47.2	43.4	**62.8**	59.1	61.8	61.2	51.1
Indonesian	IndoCulture$_{(0\ shot)}$	**73.4**	58.7	65.6	56.7	53.0	**76.4**	68.9	66.1	72.7	70.6
	TydiQA$_{(3\ shot)}$	**66.4**	63.5	65.5	63.4	65.5	**71.7**	63.9	65.1	69.9	58.2
	Belebele$_{(3\ shot)}$	48.9	49.3	**50.7**	46.8	30.6	52.1	54.1	53.3	56.4	**60.3**
Thai	MMLU$_{(5\ shot)}$	55.4	52.8	**57.8**	44.1	50.8	66.3	**70.7**	62.5	67.1	39.6
	M3Exam$_{(5\ shot)}$	**57.0**	51.7	52.7	43.7	51.3	**69.3**	69.2	57.0	63.7	38.6
	Belebele$_{(3\ shot)}$	43.2	44.1	40.6	43.1	43.0	47.4	49.4	46.0	**52.3**	45.3
Vietnamese	VMLU$_{(3\ shot)}$	56.2	52.6	51.7	48.9	**56.8**	**65.9**	64.9	59.1	63.9	65.9
	M3Exam$_{(3\ shot)}$	65.6	**66.4**	65.5	54.4	63.1	74.6	**77.3**	68.6	68.9	63.2
	Belebele$_{(3\ shot)}$	48.7	**50.8**	49.0	46.0	48.6	53.8	54.6	52.0	**61.8**	58.3
Malay	Tatabahasa$_{(3\ shot)}$	**67.3**	41.5	53.6	42.9	37.4	**67.3**	50.4	58.6	58.3	48.1
Javanese	M3Exam$_{(3\ shot)}$	**57.1**	35.9	45.3	40.4	38.5	**62.3**	47.7	49.1	53.4	46.1
Multiple	FLORES-200$_{(3\ shot)}$	35.4	30.6	**35.8**	31.7	29.6	35.8	34.3	**36.6**	36.5	35.7
	XCOPA$_{(3\ shot)}$	**74.1**	71.8	73.0	69.4	70.4	77.5	77.3	75.3	**79.8**	72.1

Sailor2 Chat Model Evaluation

The win rate of Sailor2-20B-Chat against GPT-4o-0806 on Sea-WildBench is nearly 50%, demonstrating it performs at **the GPT-4o level for local chat scenarios**.

Model	SWB Score	Coding	Creative Tasks	Info Seeking	Reasoning	Math	Length
Sailor2-20B-Chat	0.56	0.62	0.56	0.58	0.57	0.54	2814.74
Sailor2-8B-Chat	0.49	0.42	0.57	0.53	0.50	0.42	2849.41
Qwen2.5-72B-Instruct	0.45	0.50	0.39	0.44	0.45	0.49	3026.82
SEA-LIONv3-70B-Instruct	0.40	0.42	0.38	0.40	0.39	0.39	2340.65
Gemma-2-27B-Instruct	0.40	0.38	0.41	0.39	0.39	0.37	2288.33
Qwen2.5-32B-Instruct	0.32	0.39	0.28	0.29	0.32	0.33	2090.61
Gemma-2-9B-Instruct	0.31	0.26	0.36	0.33	0.30	0.26	2163.03
Qwen2.5-14B-Instruct	0.30	0.33	0.25	0.28	0.28	0.30	2267.94
Llama-3.1-70B-Instruct	0.30	0.37	0.26	0.28	0.28	0.28	2543.06
SEA-LIONv3-8B-Instruct	0.30	0.32	0.32	0.30	0.28	0.22	2357.14
Aya-Expanse-32B	0.29	0.29	0.28	0.28	0.27	0.24	2495.47
Qwen2-72B-Instruct	0.26	0.22	0.27	0.28	0.25	0.23	1546.21
Qwen2.5-7B-Instruct	0.25	0.28	0.20	0.23	0.22	0.22	2415.08
SEA-LIONv2.1-8B-Instruct	0.23	0.23	0.24	0.24	0.20	0.18	1735.26
SeaLLMs-v3-7B-Chat	0.21	0.21	0.19	0.19	0.18	0.15	2298.47
Llama-3.1-8B-Instruct	0.19	0.18	0.15	0.16	0.15	0.13	2356.67
SeaLLM-7B-v2	0.18	0.14	0.16	0.17	0.14	0.12	2298.15
SeaLLM-7B-v2.5	0.17	0.14	0.14	0.15	0.13	0.11	2184.55
Qwen2.5-3B-Instruct	0.16	0.14	0.10	0.12	0.12	0.13	2324.08
Sailor-14B-Chat	0.16	0.07	0.11	0.13	0.10	0.09	2465.85
SeaLLM-7B-v1	0.12	0.03	0.07	0.09	0.07	0.06	2585.40
Mistral-7B-Instruct-v0.3	0.10	0.11	0.03	0.07	0.06	0.07	2336.51
Sailor-7B-Chat	0.09	0.02	0.04	0.06	0.04	0.03	1404.60
Llama-2-70B-Chat	0.08	0.07	0.05	0.06	0.05	0.05	2354.30
Llama-2-13B-Chat	0.06	0.04	0.04	0.05	0.03	0.03	2317.36
Llama-2-7B-Chat	0.05	0.03	0.02	0.04	0.02	0.03	2330.50

Task-Level Evaluation

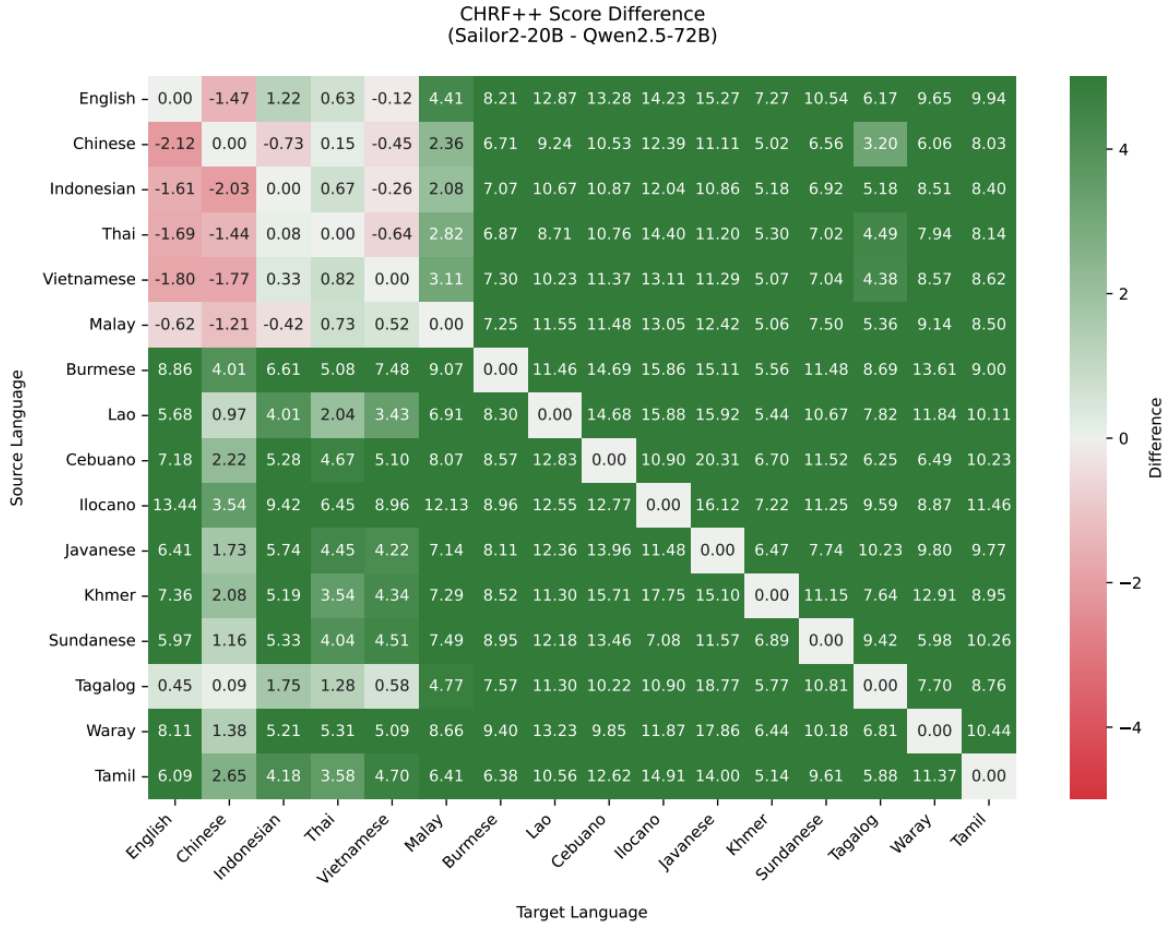
Model	SWB Score	tha	vie	ind	tgl	zsm	khm	lao	mya	Length
Sailor2-20B-Chat	0.56	0.53	0.50	0.54	0.50	0.49	0.63	0.69	0.64	2814.74
Sailor2-8B-Chat	0.49	0.48	0.46	0.46	0.42	0.43	0.50	0.66	0.55	2849.41
Qwen2.5-72B-Instruct	0.45	0.54	0.51	0.51	0.42	0.48	0.33	0.41	0.31	3026.82
SEA-LIONv3-70B-Instruct	0.40	0.45	0.45	0.48	0.40	0.41	0.32	0.28	0.32	2340.65
Gemma-2-27B-Instruct	0.40	0.43	0.40	0.46	0.40	0.39	0.34	0.38	0.31	2288.33
Qwen2.5-32B-Instruct	0.32	0.37	0.42	0.42	0.26	0.38	0.24	0.19	0.16	2090.61
Gemma-2-9B-Instruct	0.31	0.36	0.40	0.39	0.30	0.38	0.19	0.19	0.19	2163.03
Qwen2.5-14B-Instruct	0.30	0.40	0.40	0.23	0.35	0.20	0.21	0.12	0.30	2267.94
Llama-3.1-70B-Instruct	0.30	0.33	0.37	0.37	0.28	0.35	0.18	0.15	0.19	2543.06
SEA-LIONv3-8B-Instruct	0.30	0.38	0.40	0.38	0.34	0.35	0.12	0.08	0.14	2357.14
Aya-expanse-32B	0.29	0.25	0.45	0.46	0.27	0.35	0.06	0.12	0.13	2495.47
Qwen2-72B-Instruct	0.26	0.26	0.30	0.33	0.29	0.32	0.20	0.20	0.16	1546.21
Qwen2.5-7B-Instruct	0.25	0.30	0.35	0.36	0.12	0.29	0.09	0.09	0.08	2415.08
Sealionv2.1-8B-Instruct	0.23	0.30	0.33	0.31	0.16	0.28	0.07	0.08	0.10	1735.26
SeaLLMs-v3-7B-Chat	0.21	0.23	0.22	0.21	0.19	0.16	0.15	0.16	0.09	2298.47
Llama-3.1-8B-Instruct	0.19	0.19	0.26	0.21	0.15	0.18	0.06	0.07	0.07	2356.67
SeaLLM-7B-v2	0.18	0.18	0.18	0.19	0.09	0.13	0.10	0.12	0.09	2298.15
SeaLLM-7B-v2.5	0.17	0.18	0.19	0.18	0.10	0.14	0.08	0.11	0.06	2184.55
Qwen2.5-3B-Instruct	0.16	0.14	0.21	0.18	0.08	0.16	0.06	0.06	0.04	2324.08
Sailor-14B-Chat	0.16	0.11	0.17	0.14	0.04	0.14	0.02	0.12	0.06	2465.85
SeaLLM-7B-v1	0.12	0.05	0.07	0.07	0.04	0.05	0.10	0.11	0.09	2585.40
Mistral-7B-Instruct-v0.3	0.10	0.07	0.11	0.07	0.08	0.11	0.02	0.03	0.02	2336.51
Sailor-7B-Chat	0.09	0.04	0.07	0.05	0.02	0.06	0.02	0.07	0.03	1404.60
Llama-2-70B-Chat	0.08	0.02	0.05	0.11	0.06	0.13	0.03	0.01	0.03	2354.30
Llama-2-13B-Chat	0.06	0.01	0.05	0.08	0.02	0.03	0.01	0.01	0.03	2317.36
Llama-2-7B-Chat	0.05	0.01	0.02	0.05	0.04	0.03	0.01	0.02	0.04	2330.50

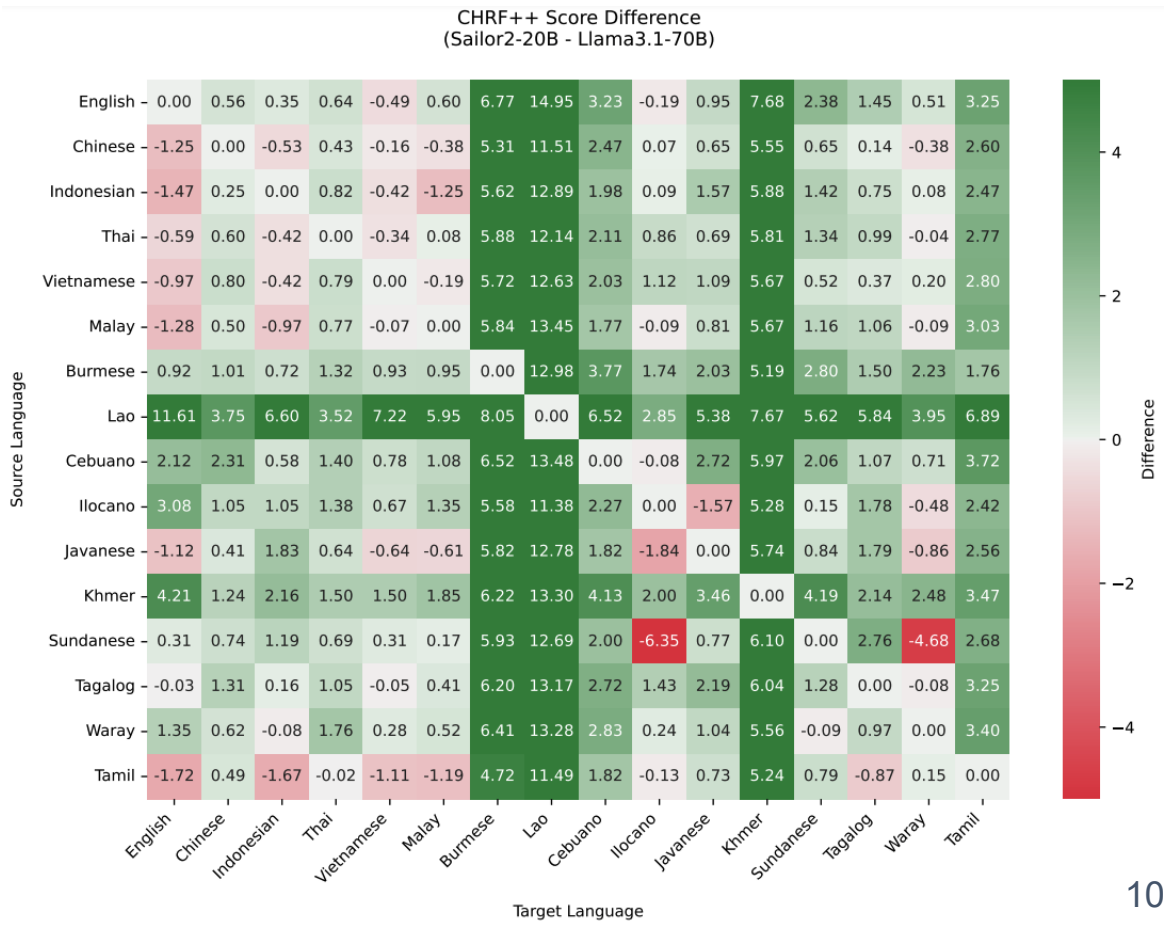
Language-Level Evaluation

SWB Score is the win rate against GPT-4o, which also serves as the evaluator.

Sailor2: Strong Translation Performance

Compare with models **3x larger** (Qwen2.5-72B and Llama3.1-70B), **Sailor2 performs significantly better in SEA languages** while remaining comparable in high-resource languages.





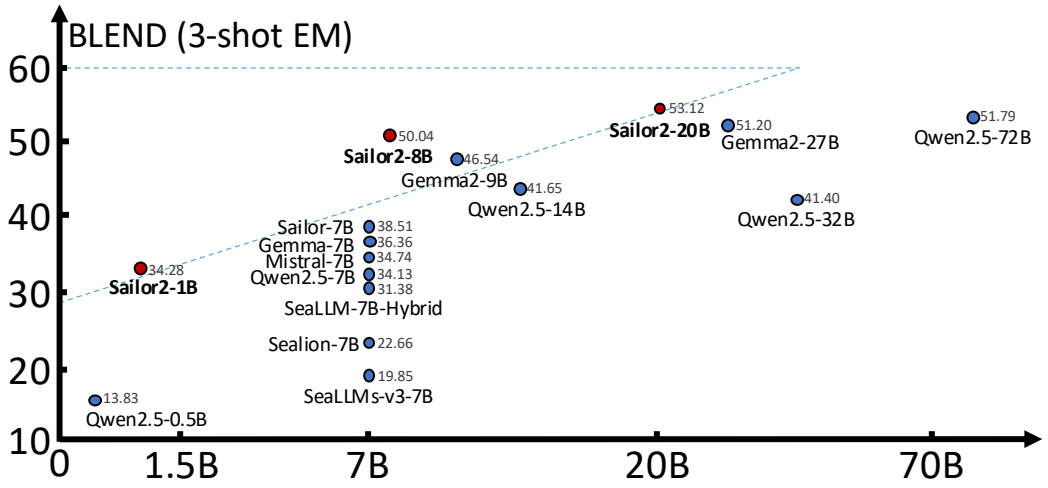
Sailor2: Long Context Training

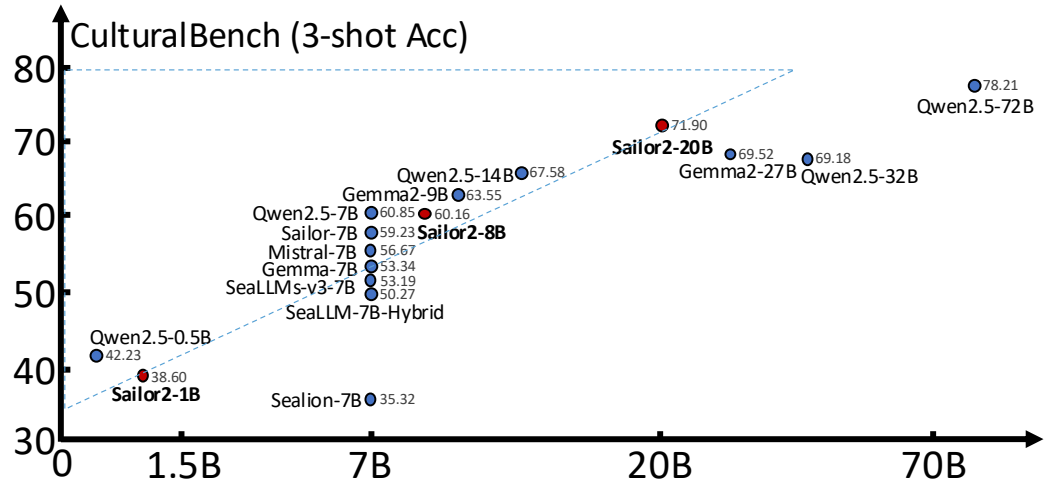
Sailor2 supports **32K-128K context input**, enabling more use cases such as **DocumentQA**.

Model	128K	64K	32K	16K	8K	4K
Qwen2.5-0.5B	0.00	0.00	46.50	52.65	55.95	64.42
Sailor2-1B	0.00	0.00	0.62	3.99	35.81	55.93
Sailor2-1B-32K	0.00	0.00	36.52	49.63	55.50	56.84
Qwen2.5-7B	20.67	61.70	78.58	81.72	83.58	86.72
Sailor2-8B	0.00	2.17	9.59	23.08	49.13	69.38
Sailor2-8B-128K	19.94	41.57	54.61	64.32	75.73	80.04
Qwen2.5-14B	32.93	66.68	85.09	86.96	87.40	87.56
Sailor2-20B	0.55	14.08	46.60	67.76	79.62	87.86
Sailor2-20B-128K	47.46	66.70	79.52	85.24	86.63	88.21

Sailor2: Better Culture Understanding

Among models of similar size, **Sailor2 understands SEA culture better**, including its food, traditions and geography, making it ideal for **locally-relevant chat applications.**





Paano madalas na bumabati ang mga tao sa Pilipinas nang hindi gumagamit ng mga salita? (English Question: How do people in the Philippines often greet each other without using words?) A. Sa pagtaas ng kilay nila B. Sa pamamagitan ng pagbibigay ng mahigpit na pagkakamay C. Sa pagtango ng kanilang ulo D. Sa pamamagitan ng pagbibigay ng high five.		
Prediction:		
Golden: C	Qwen2.5-32B: B ✗	Sailor2-20B: C ✓

Masakan Indonesia populer apa yang menggunakan daging yang ditusuk dan dipanggang, disajikan dengan saus kental, pedas, dan berbahan dasar kunyit? (English Question: What popular Indonesian dish uses skewered and grilled meat, served with a thick, spicy, and turmeric-based sauce?) A. Sate Padang. B. Sate Madura. C. Steak Wagyu. D. Iga BBQ.		
Prediction:		
Golden: A	Qwen2.5-72B: B ✗	Sailor2-20B: A ✓

Sailor2: Accelerate the Multilingual LLM Research with Open Cookbook

Data Curation	Continual Pre-Training	Post-Training	Model Customization	Evaluation
Fineweb-Pro Chinese-Fineweb-Edu Open-Web-Math-Pro	Model Expansion	Data Selection for Instruction Tuning	Long-Context Training	SailCompass (Generation, Classification)
Dataset for Replay	Two-Stage Continual Training	Two-Stage Instruction Tuning	Speculative Decoding	FLoRes (Translation)
Wikipedia Open Subtitles Translation CommonCrawl Public PDF	Model Parallel Optimization	Two-Stage Preference Tuning	Model Pruning	CultureBench (Culture)
Dataset for SEA Languages				BLEnD (Culture)
SEA-UltraFeedback SEA-UltraChat				Global MMLU (Culture, Knowledge)
Dataset for Post-Training				RULER (Long-Context)
Data Recall for Low-resource Language				SEA-WildBench (Chat)
Data Mixture with Tiny Model Simulation				

Key Insights in Building Sailor2

Data

- **Pre-training Data:** Rigorous multi-level deduplication (document, sentence, URL) to mitigate the redundancy of Common Crawl sources.
- **Post-training Data:** Popular translation models (e.g., GPT4o/NLLB) often underperform in low-resource languages and require careful refinement.

Training

- **Model Expansion:** Improve model capacity to better absorb multilingual knowledge, especially in over-trained models.
- **Two-stage Training:** Ensure steady learning from data of varying quality.

Evaluation

- **Comprehensive Metrics:** Assess basic language understanding, cultural context awareness, and conversational proficiency.
- **Probe Data:** Use the selected probe data to accurately and promptly check the model's capabilities beyond standard quantitative metrics.

Sailor2: Open Model for Better Developments

All **models, resources, and code** associated with Sailor2 are released under the Apache 2.0 License, which allows **commercial usage**.

Model Checkpoints

Stage	Sailor2-1B	Sailor2-8B	Sailor2-20B
Pre-Annealing	sail/Sailor2-1B-Pre	sail/Sailor2-8B-Pre	sail/Sailor2-20B-Pre
Base	sail/Sailor2-1B	sail/Sailor2-8B	sail/Sailor2-20B
SFT	sail/Sailor2-1B-SFT	sail/Sailor2-8B-SFT	sail/Sailor2-20B-SFT
Chat	sail/Sailor2-1B-Chat	sail/Sailor2-8B-Chat	sail/Sailor2-20B-Chat

Codebases / Tools

Type	Link
Data Cleaning	sail-sg/sailcraft
Data Mixture	sail-sg/regmix
Pre-training	sail-sg/Megatron-Sailor2
Post-training	sail-sg/oat
Evaluation	sail-sg/sailcompass

Post-Training Dataset

Domain	Link
SFT-Stage1	sailor2/sailor2-sft-stage1
SFT-Stage2	sailor2/sailor2-sft-stage2
Off-policy DPO	sailor2/sea-ultrafeedback
On-policy DPO	sailor2/sea-ultrafeedback-onpolicy

Evaluation Dataset

Domain	Link
SailCompass	sail/Sailcompass_data
SEA-WildBench	sailor2/sea-wildbench

Model Checkpoints (via Long-Context Training)

Stage	Sailor2-1B	Sailor2-8B	Sailor2-20B
Base	sail/Sailor2-L-1B	sail/Sailor2-L-8B	sail/Sailor2-L-20B
SFT	sail/Sailor2-L-1B-SFT	sail/Sailor2-L-8B-SFT	sail/Sailor2-L-20B-SFT
Chat	sail/Sailor2-L-1B-Chat	sail/Sailor2-L-8B-Chat	sail/Sailor2-L-20B-Chat

Model Checkpoints (via Speculative Decoding)

Stage	Sailor2-8B	Sailor2-20B
Base Model	sail/Sailor2-8B-Chat-Glide	sail/Sailor2-20B-Chat-Glide

Model Checkpoints (via Model Pruning)

Stage	Sailor2-3B (Pruning via Sailor2-8B)	Sailor2-14B (Pruning via Sailor2-20B)
Base Model	sail/Sailor2-3B	sail/Sailor2-14B
SFT	sail/Sailor2-3B-SFT	sail/Sailor2-14B-SFT
Chat	sail/Sailor2-3B-Chat	sail/Sailor2-14B-Chat

What's Next? More Low-Resource Languages.

(1) Synthetic **Data** for Low-resource languages

- Current Common-crawl only sources a **small set of data** in low-resource languages.
- TODO: Global data mining and translate high-quality corpus from English.

(2) Tokenizer-Free **Model** for Open-Vocabulary Learning

- Current Tokenizer is **unfriendly to morphological-rich languages** like Thai and Khmer.
- TODO: Pixel-based LLM or Byte-level LLM.

(3) Efficient **Continual Pre-training** for Multilingual model

- Existing LLMs are **overfitting to certain languages**. Continual training is tricky and costly.
- TODO: Improve the model plasticity and partially update the model parameters.

Thank You