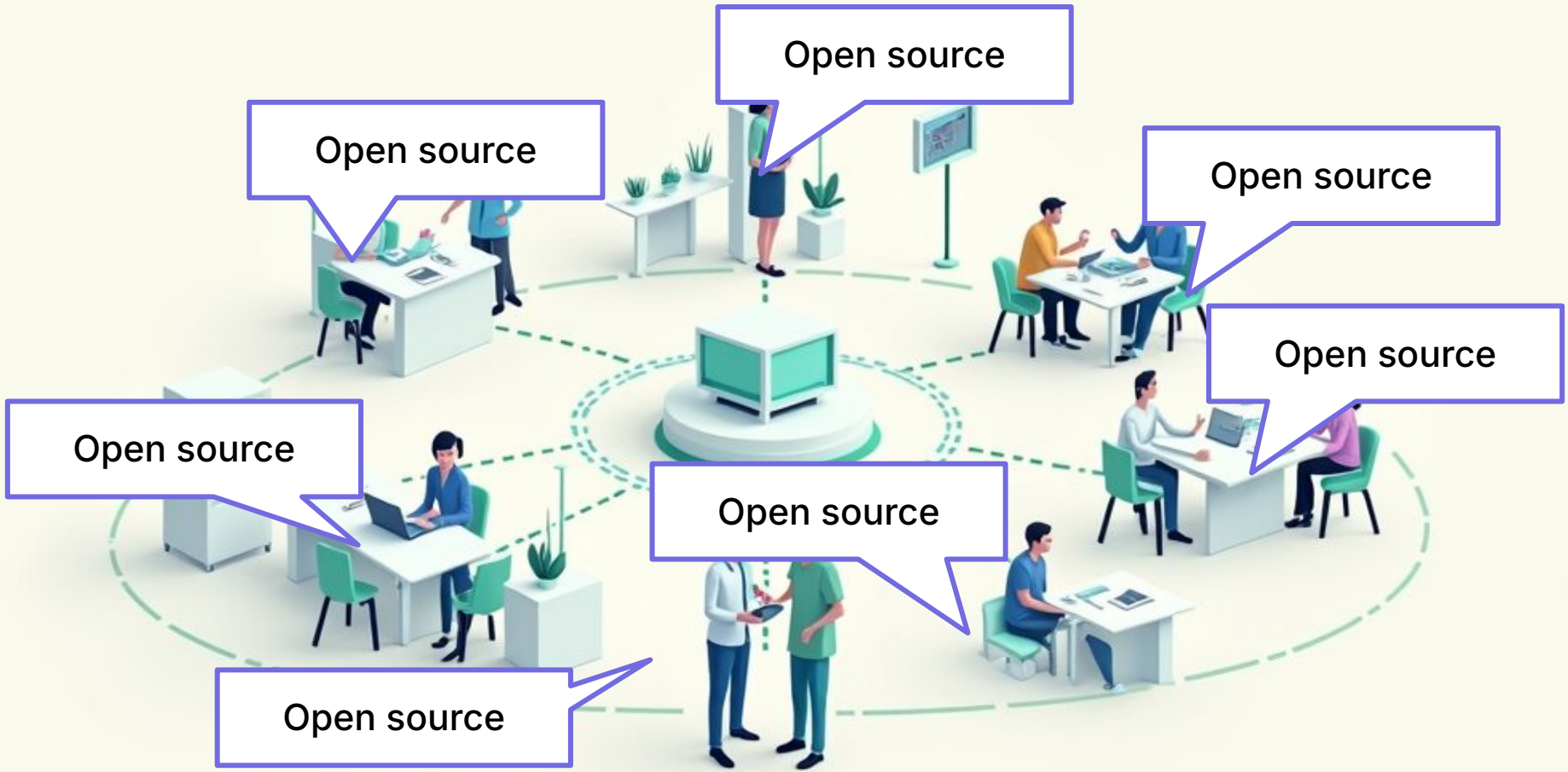
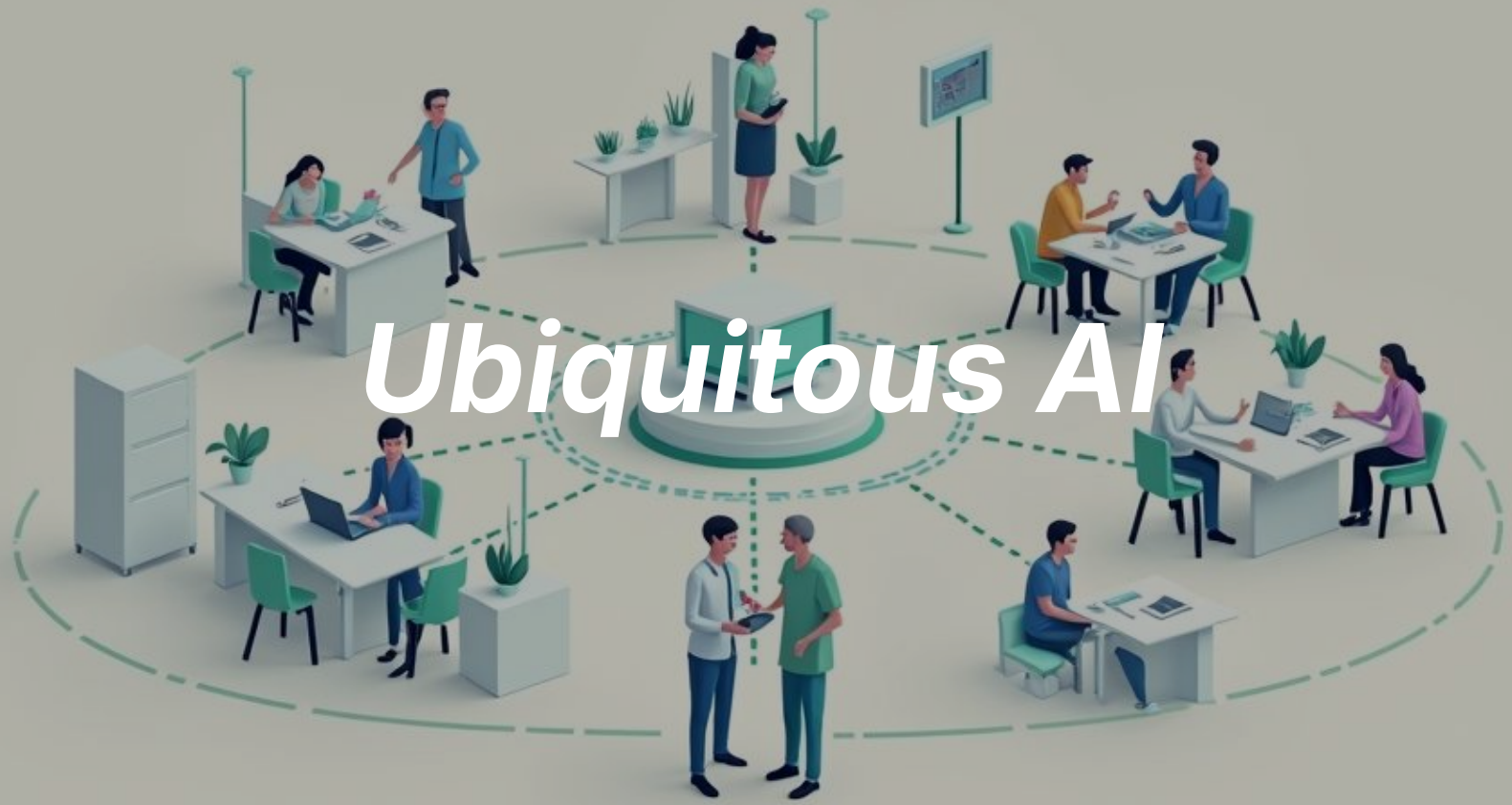
TYPHOON

Open-Source Thai Language Technologies

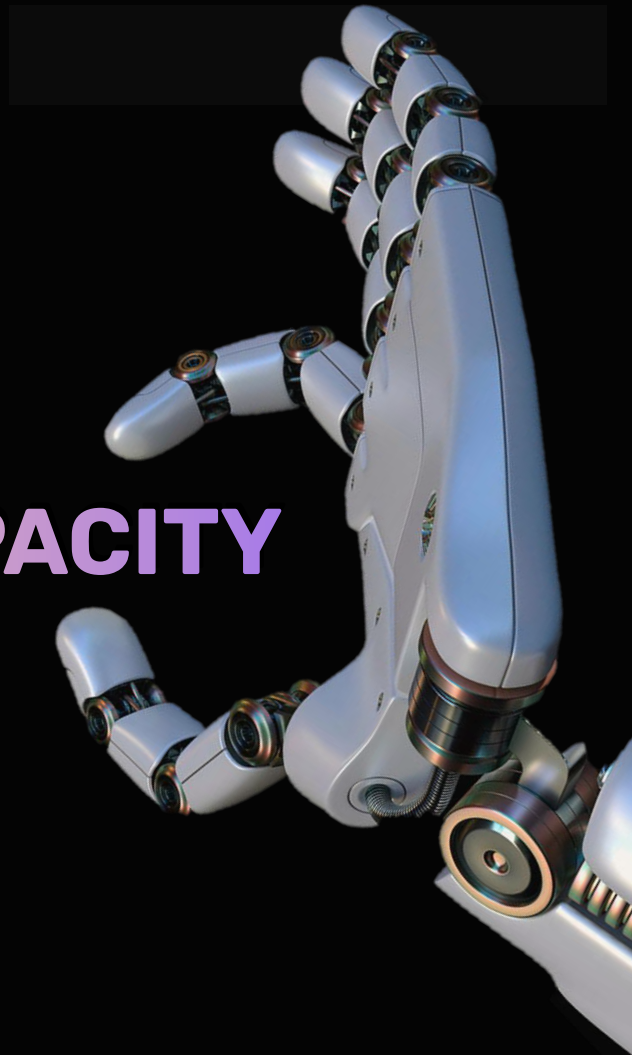
Pittawat (Pete) Taveekitworachai
Research Scientist, SCB 10X

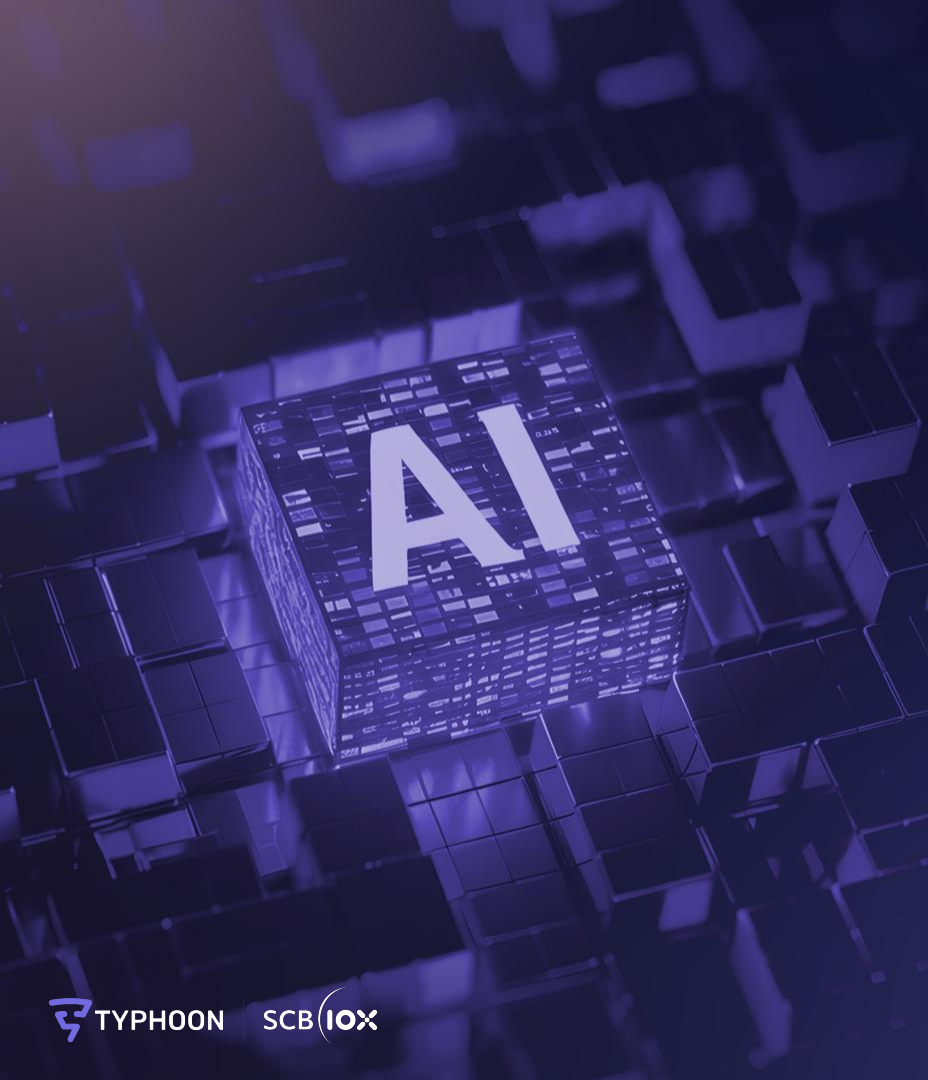
FOSSASIA Summit 2025
15 March 2025


Open source
Open source
Open source
Open source
Open source
Open source
Open source


Ubiquitous AI

LIMITED TECHNICAL CAPACITY





Sovereign AI

Thailand needs its own AI models to preserve our linguistic and cultural identity, build our own capabilities, and shape our technological destiny. By investing in open-source LLMs, we can drive our own innovation and engage with global players as peers in the long term.

TYPHOON

What Is Typhoon?

Typhoon is an **advanced research initiative** focused on developing **open-source large language technologies for the Thai language.** We provide **models, datasets, tools, and research** to **advance Thai language AI and multimodal capabilities**

Efficient Speed & Cost

Improved Thai Knowledge and Instruction-Following Performance

Open Source

Open access to resources fosters collaboration and drives AI innovation

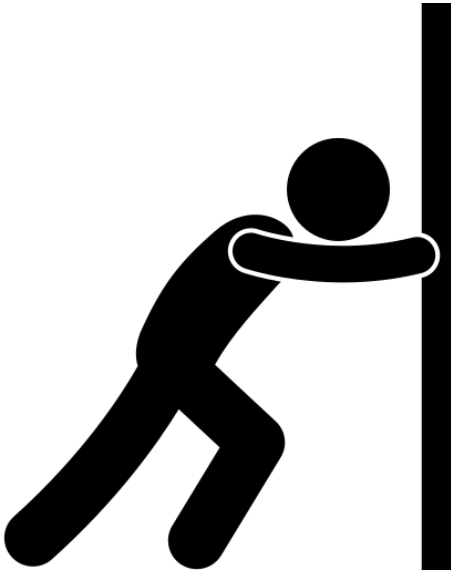
Open Source *Accelerates* Humanity's Progress

📀 *Open Data*

🤖 *Open Model Weights*

💻 *Open Source Code*

📗 *Open Knowledge*



Collaboration Over Competition

Recent Releases

Reasoning Models

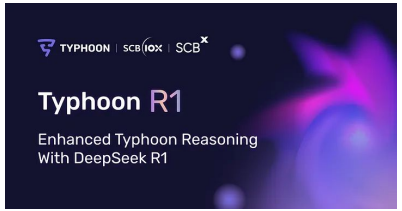
Cutting Edge Research

Typhoon 2

Our latest release, building on Typhoon 1.5 and 1.5X. It includes models ranging from compact, edge-capable options (1B and 3B) to 70 billion parameters, specifically optimized for Thai applications.

Typhoon T1

Southeast Asia's first open reasoning model. Typhoon T1 3B, the debut model in our "Typhoon T" series, is setting a new benchmark for structured, thoughtful AI reasoning—excelling in math, coding, and other complex tasks.



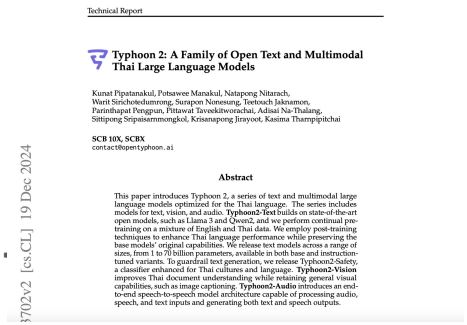
Typhoon R1

Built on the solid foundations of Typhoon 2 and Deepseek R1, Typhoon R1 enhances Typhoon 2 with Deepseek R1's reasoning capabilities while maintaining Typhoon's Thai capabilities via model merging.

Recent Releases

Reasoning Models

Cutting Edge Research


Technical Report

Typhoon 2: A Family of Open Text and Multimodal Thai Large Language Models

Kunat Pipatanakul, Potsawee Manakul, Natapong Nitarach, Warit Sirichotedumrong, Surapon Nonesung, Teetouch Jaknamon, Parinthapat Pengpun, Pittawat Taveekitworachai, Adisai Na-Thalang, Sittipong Sripaisarnmongkol, Krisanapong Jirayoot, Kasima Tharnpipitchai

SCB 10X, SCBX
contact@opentyphoon.ai

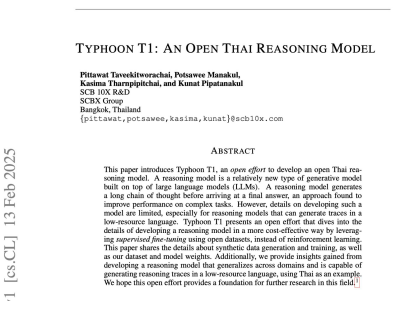
Abstract

This paper introduces Typhoon 2, a series of text and multimodal large language models optimized for the Thai language. The series includes models for text, vision, and audio. **Typhoon2-Text** builds on state-of-the-art open models, such as Llama 3 and Qwen2, and we perform continual pre-training on a mixture of English and Thai data. We employ post-training techniques to enhance Thai language performance while preserving the base models' original capabilities. We release text models across a range of sizes, from 1 to 70 billion parameters, available in both base and instruction-tuned variants. To guardrail text generation, we release Typhoon2-Safety, a classifier enhanced for Thai cultures and language. **Typhoon2-Vision** improves Thai document understanding while retaining general visual capabilities, such as image captioning. **Typhoon2-Audio** introduces an end-to-end speech-to-speech model architecture capable of processing audio, speech, and text inputs and generating both text and speech outputs.

9702v2 [cs.CL] 19 Dec 2024

Typhoon 2: A Family of Open Text and Multimodal Thai Large Language Models

https://arxiv.org/abs/2412.13702


TYPHOON T1: AN OPEN THAI REASONING MODEL

Pittawat Taveekitworachai, Potsawee Manakul, Kasima Tharnpipitchai, and Kunat Pipatanakul
SCB 10X R&D
SCBX Group
Bangkok, Thailand
{pittawat,potsawee,kasima,kunat}@scb10x.com

ABSTRACT

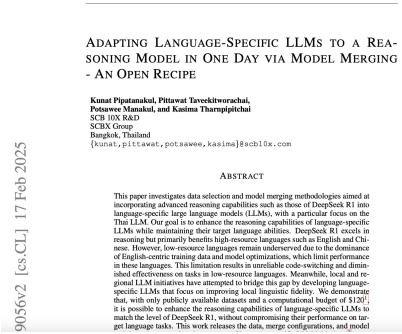
This paper introduces Typhoon T1, an *open effort* to develop an open Thai reasoning model. A reasoning model is a relatively new type of generative model built on top of large language models (LLMs). A reasoning model generates a long chain of thought before arriving at a final answer, an approach found to improve performance on complex tasks. However, details on developing such a model are limited, especially for reasoning models that can generate traces in a low-resource language. Typhoon T1 presents an open effort that dives into the details of developing a reasoning model in a more cost-effective way by leveraging *supervised fine-tuning* using open datasets, instead of reinforcement learning. This paper shares the details about synthetic data generation and training, as well as our dataset and model weights. Additionally, we provide insights gained from developing a reasoning model that generalizes across domains and is capable of generating reasoning traces in a low-resource language, using Thai as an example. We hope this open effort provides a foundation for further research in this field.

1 [cs.CL] 13 Feb 2025

Typhoon T1: An Open Thai Reasoning Model

Accepted at ICLR 2025 SCI-FM Workshop

https://arxiv.org/abs/2502.09042


ADAPTING LANGUAGE-SPECIFIC LLMS TO A REASONING MODEL IN ONE DAY VIA MODEL MERGING - AN OPEN RECIPE

Kunat Pipatanakul, Pittawat Taveekitworachai, Potsawee Manakul, and Kasima Tharnpipitchai
SCB 10X R&D
SCBX Group
Bangkok, Thailand
{kunat,pittawat,potsawee,kasima}@scb10x.com

ABSTRACT

This paper investigates data selection and model merging methodologies aimed at incorporating advanced reasoning capabilities such as those of DeepSeek R1 into language-specific large language models (LLMs), with a particular focus on the Thai LLM. Our goal is to enhance the reasoning capabilities of language-specific LLMs while maintaining their target language abilities. DeepSeek R1 excels in reasoning but primarily benefits high-resource languages such as English and Chinese. However, low-resource languages remain underserved due to the dominance of English-centric training data and model optimizations, which limit performance in these languages. This limitation results in unreliable code-switching and diminished effectiveness on tasks in low-resource languages. Meanwhile, local and regional LLM initiatives have attempted to bridge this gap by developing language-specific LLMs that focus on improving local linguistic fidelity. We demonstrate that, with only publicly available datasets and a computational budget of $120, it is possible to enhance the reasoning capabilities of language-specific LLMs to match the level of DeepSeek R1, without compromising their performance on target language tasks. This work releases the data, merge configurations, and model

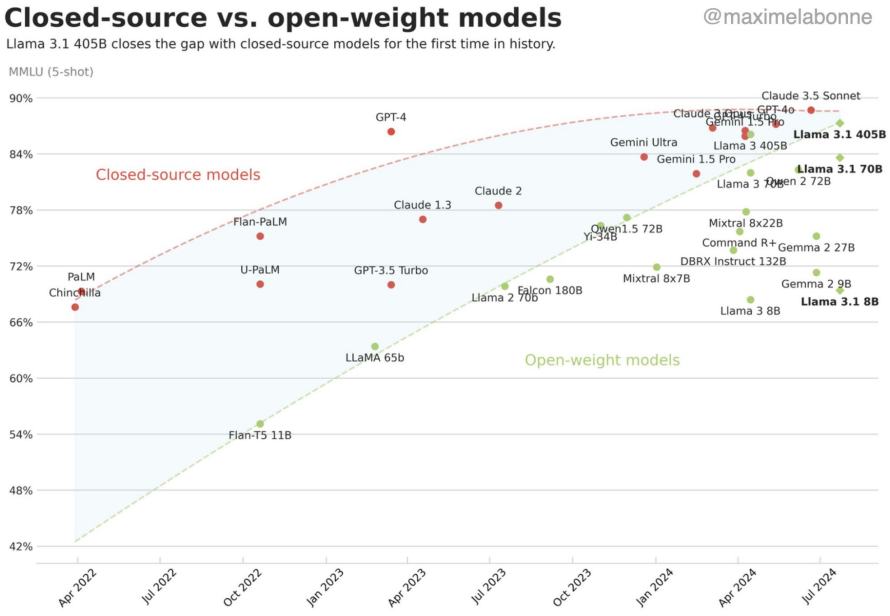
9056v2 [cs.CL] 17 Feb 2025

Adapting Language-Specific LLMs to a Reasoning Model in One Day via Model Merging—An Open Recipe

Accepted at ICLR 2025 SCI-FM Workshop

https://arxiv.org/abs/2502.09056

Open Model VS Proprietary Model
***Performance Gap*?**



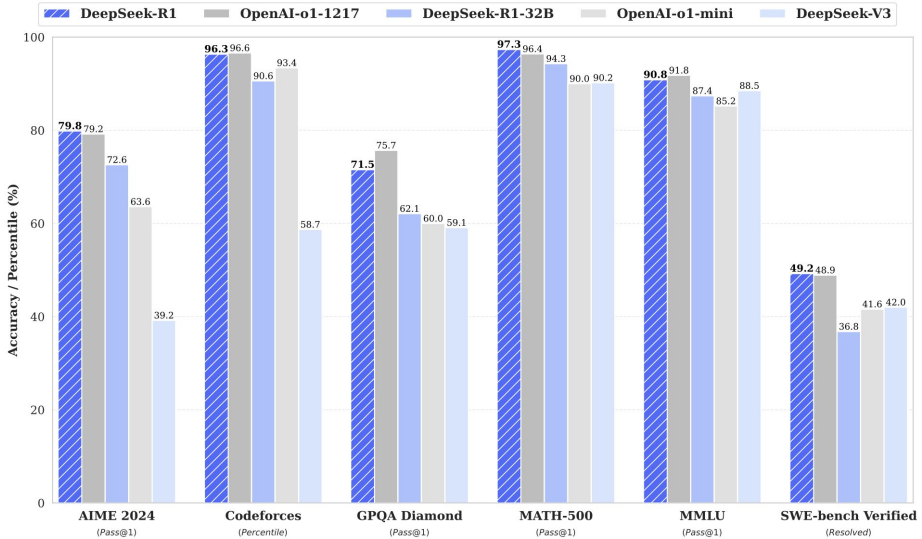


Figure 1 | Benchmark performance of DeepSeek-R1.

arxiv.org/abs/2501.12948

Open Collaborations

Let's work together!

TYPHOON | SCB 10X

Be Part of the Open-Source LLM Revolution!

Connect & Collaborate

- Join our Discord community
- Chat with us today in person!

Build & Experiment

- Access our models in Hugging Face
- Create your own LLM application

Join Our Team

Now hiring: • Full-Stack Engineers • Research Scientists

Start Here: opentyphoon.ai