達摩院
DAMO ACADEMY

Bridging the Digital Divide with SeaLLMs: Strategies for Inclusive Digital Transformation

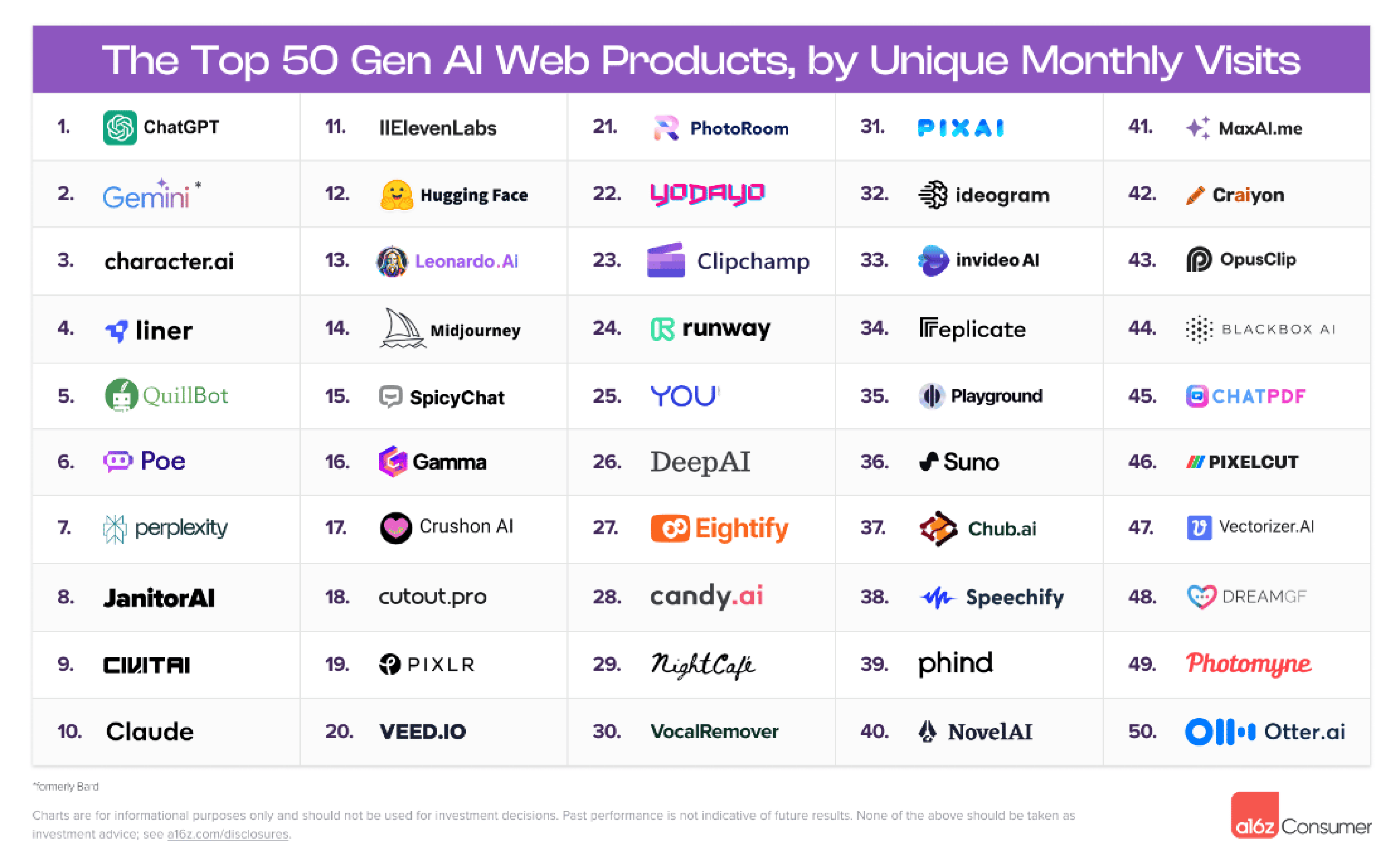
July 31st, 2024. Mekong Forum

Lidong Bing

Director of the Language Technology Lab

DAMO Academy of Alibaba Group

Which have you used?



The Top 50 Gen AI Web Products, by Unique Monthly Visits

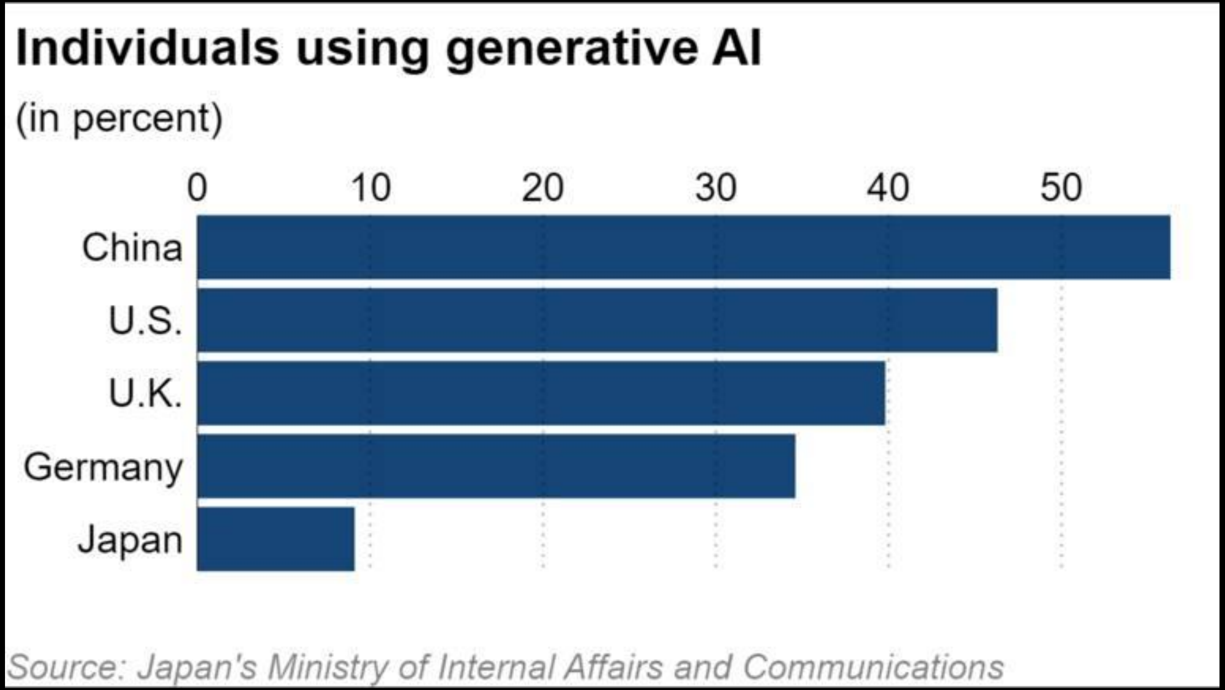
1.	ChatGPT	11.	IIElevenLabs	21.	PhotoRoom	31.	PIXAI	41.	MaxAI.me
2.	Gemini*	12.	Hugging Face	22.	yodayo	32.	ideogram	42.	Craiyon
3.	character.ai	13.	Leonardo.Ai	23.	Clipchamp	33.	invideo AI	43.	OpusClip
4.	liner	14.	Midjourney	24.	runway	34.	replicate	44.	BLACKBOX AI
5.	QuillBot	15.	SpicyChat	25.	YOU	35.	Playground	45.	CHATPDF
6.	Poe	16.	Gamma	26.	DeepAI	36.	Suno	46.	PIXELCUT
7.	perplexity	17.	Crushon AI	27.	Eightify	37.	Chub.ai	47.	Vectorizer.AI
8.	JanitorAI	18.	cutout.pro	28.	candy.ai	38.	Speechify	48.	DREAMGF
9.	CIVITAI	19.	PIXLR	29.	NightCafe	39.	phind	49.	Photomyne
10.	Claude	20.	VEED.IO	30.	VocalRemover	40.	NovelAI	50.	Otter.ai

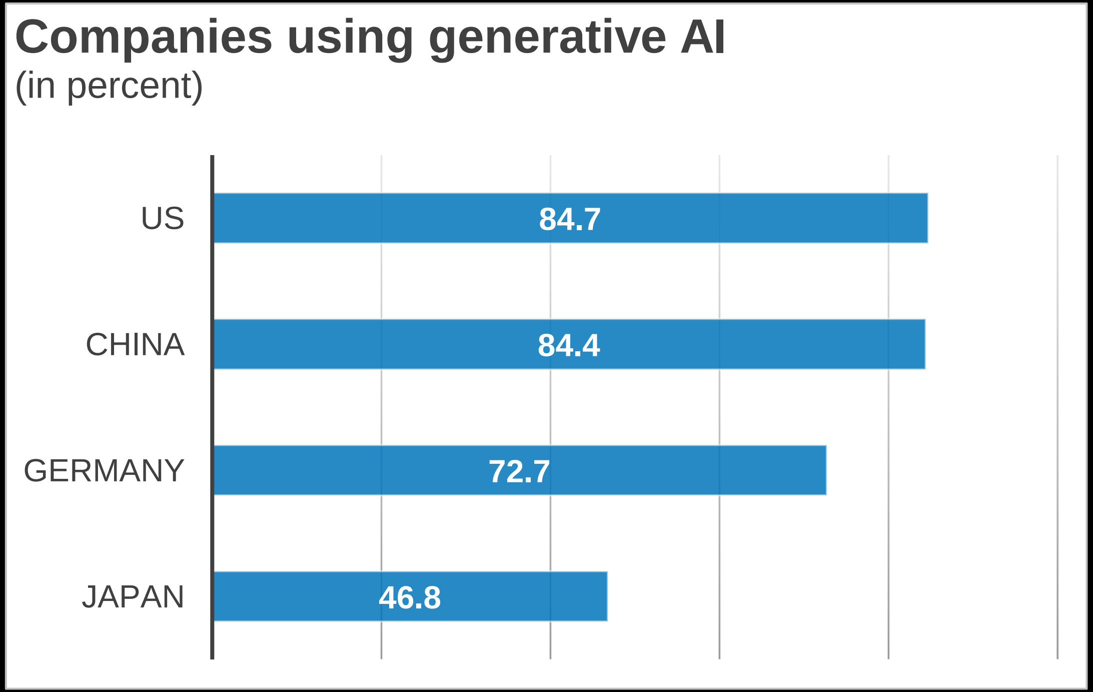
*formerly Bard

Charts are for informational purposes only and should not be used for investment decisions. Past performance is not indicative of future results. None of the above should be taken as investment advice; see a16z.com/disclosures.

a16z Consumer

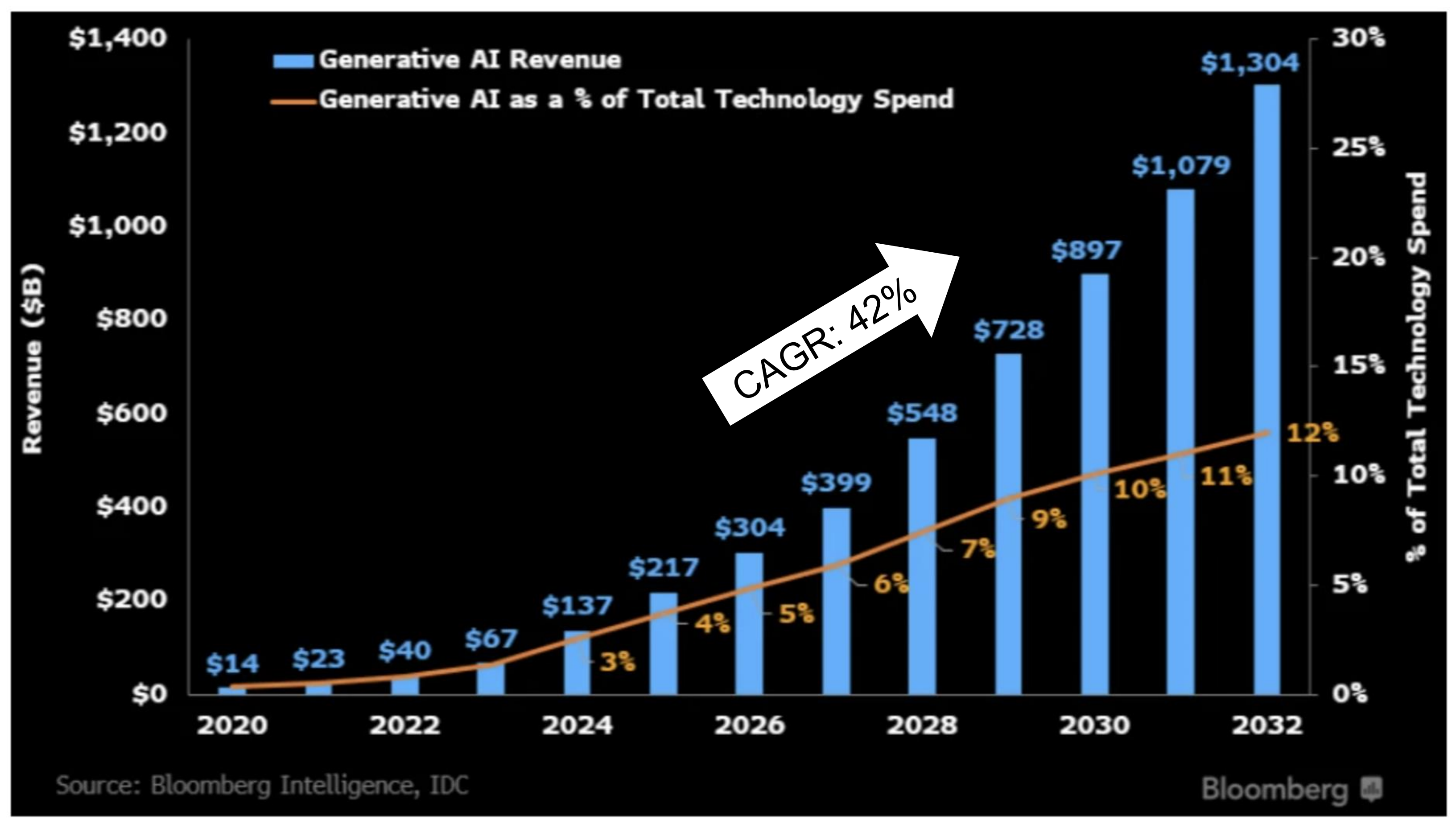
How wildly GenAI is adopted





https://www.hcamag.com/asia/specialisation/hr-technology/japan-lags-behind-china-us-in-use-of-gen-ai/496241

Generative AI will be every where on our planet 🌍

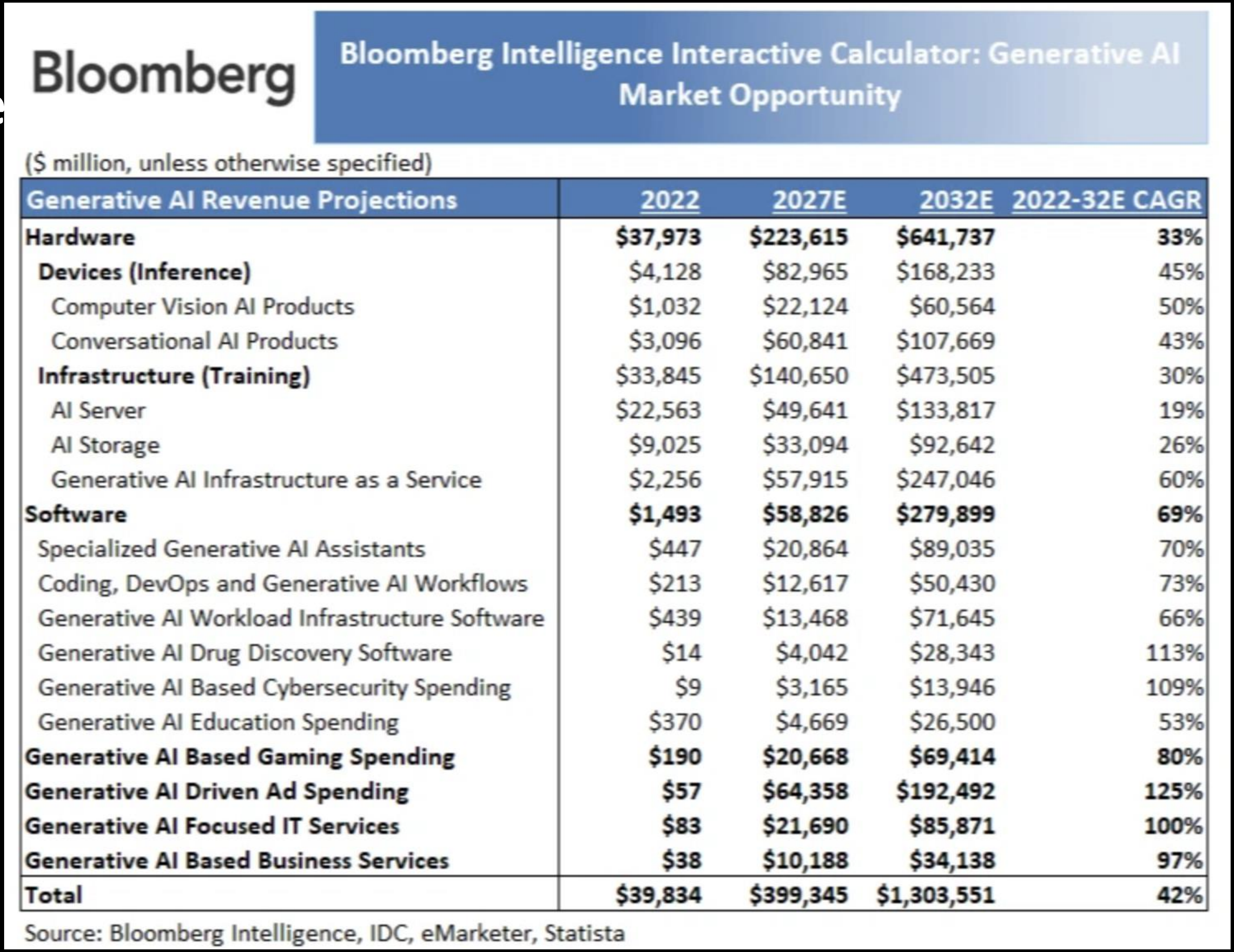


CAGR: 42%

Source: Bloomberg Intelligence, IDC

https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/

Ge



Bloomberg

Bloomberg Intelligence Interactive Calculator: Generative AI Market Opportunity

($ million, unless otherwise specified)

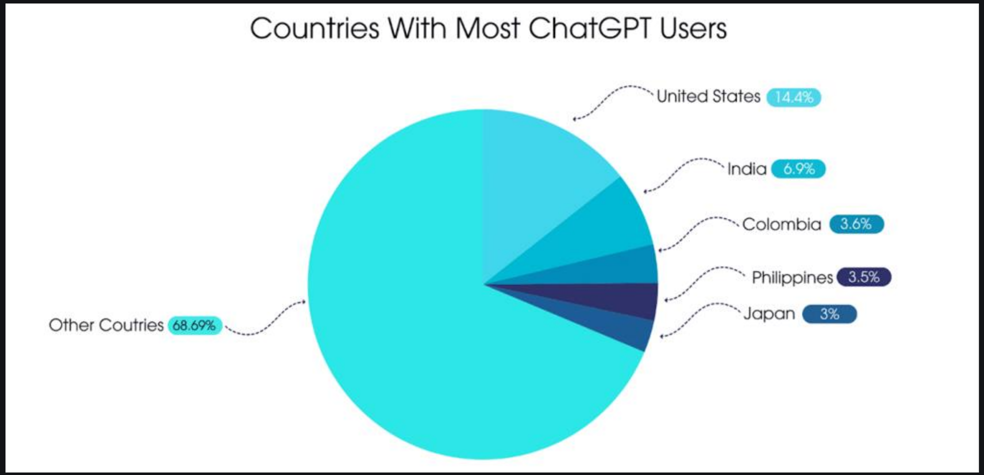
Generative AI Revenue Projections	2022	2027E	2032E	2022-32E CAGR
Hardware	**$37,973**	**$223,615**	**$641,737**	**33%**
Devices (Inference)	$4,128	$82,965	$168,233	45%
Computer Vision AI Products	$1,032	$22,124	$60,564	50%
Conversational AI Products	$3,096	$60,841	$107,669	43%
Infrastructure (Training)	$33,845	$140,650	$473,505	30%
AI Server	$22,563	$49,641	$133,817	19%
AI Storage	$9,025	$33,094	$92,642	26%
Generative AI Infrastructure as a Service	$2,256	$57,915	$247,046	60%
Software	**$1,493**	**$58,826**	**$279,899**	**69%**
Specialized Generative AI Assistants	$447	$20,864	$89,035	70%
Coding, DevOps and Generative AI Workflows	$213	$12,617	$50,430	73%
Generative AI Workload Infrastructure Software	$439	$13,468	$71,645	66%
Generative AI Drug Discovery Software	$14	$4,042	$28,343	113%
Generative AI Based Cybersecurity Spending	$9	$3,165	$13,946	109%
Generative AI Education Spending	$370	$4,669	$26,500	53%
Generative AI Based Gaming Spending	**$190**	**$20,668**	**$69,414**	**80%**
Generative AI Driven Ad Spending	**$57**	**$64,358**	**$192,492**	**125%**
Generative AI Focused IT Services	**$83**	**$21,690**	**$85,871**	**100%**
Generative AI Based Business Services	**$38**	**$10,188**	**$34,138**	**97%**
Total	**$39,834**	**$399,345**	**$1,303,551**	**42%**

Source: Bloomberg Intelligence, IDC, eMarketer, Statista

How inclusive LLMs are?

LLMs (ChatGPT, Claude, LLaMA, Mistral, etc.) are widely used globally, how multilingual they are?

- ❑ Most (famous) models generally exhibit strong performance in English
- ❑ High-resource languages (e.g., Chinese) also receive relatively good support
- ❑ What about other languages?

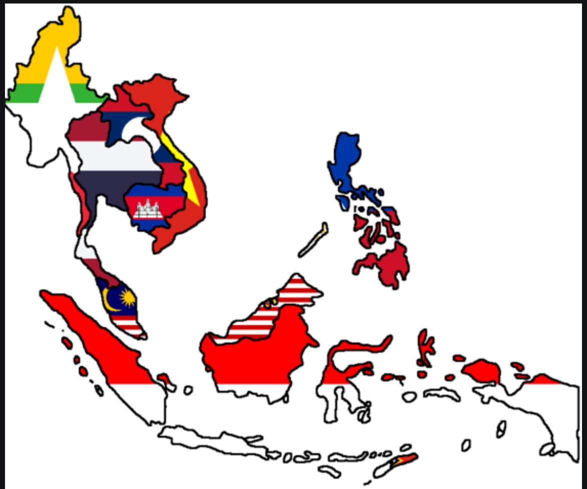


Stats source: https://www.demandsage.com/chatgpt-statistics/

State of LLMs for Southeast Asia

Not all languages are created equal!

- ❑ Linguistic studies have revealed that there are more than 6,500 human languages in the world.
- ❑ Southeast Asia is a linguistically diverse region of the world, e.g. 300 dialects in ID.
- ❑ Global models lack SEA-lang support.
- ❑ Latin vs non-Latin performance contrast
- ❑ Some SEA languages lack data severely
- ❑ Lack multilingual instruction data

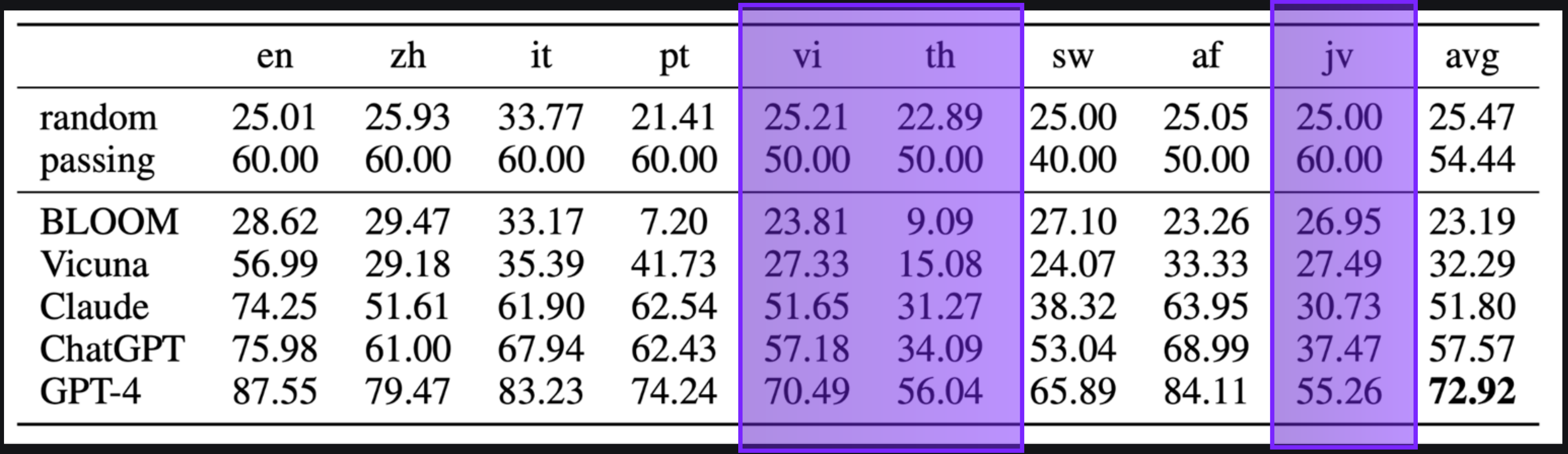


https://commons.wikimedia.org/wiki/File:Flag_map_of_South_East_Asia.png

State of LLMs for Southeast Asia

Not all languages are created equal - big performance gap between:

- ❑ latin-script v.s. non-latin script
- ❑ high-resource v.s. low-resource



	en	zh	it	pt	vi	th	sw	af	jv	avg
random	25.01	25.93	33.77	21.41	25.21	22.89	25.00	25.05	25.00	25.47
passing	60.00	60.00	60.00	60.00	50.00	50.00	40.00	50.00	60.00	54.44
BLOOM	28.62	29.47	33.17	7.20	23.81	9.09	27.10	23.26	26.95	23.19
Vicuna	56.99	29.18	35.39	41.73	27.33	15.08	24.07	33.33	27.49	32.29
Claude	74.25	51.61	61.90	62.54	51.65	31.27	38.32	63.95	30.73	51.80
ChatGPT	75.98	61.00	67.94	62.43	57.18	34.09	53.04	68.99	37.47	57.57
GPT-4	87.55	79.47	83.23	74.24	70.49	56.04	65.89	84.11	55.26	**72.92**

M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models. NIPS Dataset and Benchmarks 2023.

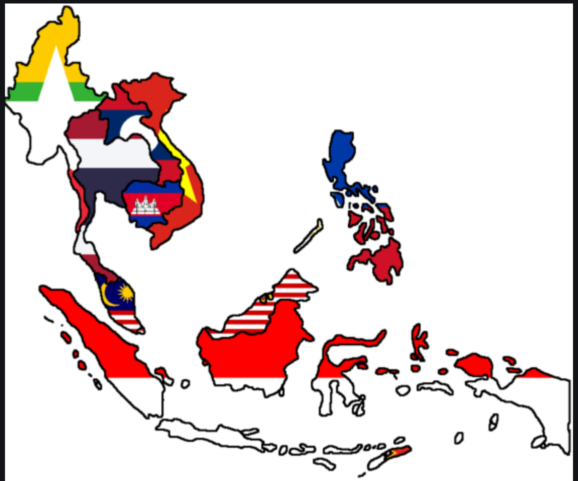
SeaLLMs for Southeast Asian Languages

Goal of SeaLLMs

❑ Built to serve Southeast Asia with support for English, Chinese, Indonesian, Vietnamese, Malay, Thai, Lao, Khmer, Burmese & Tagalog.

❑ Aim to achieve greater reception from research communities and industries in Southeast Asian countries.

❑ Adapted to local culture and regulations



SeaLLMs for Southeast Asian Languages

- ❑ Current status (https://huggingface.co/collections/SeaLLMs/)

  Nov 2023, SeaLLMs-7b - released

  Feb 2024, SeaLLMs-7b-v2 - released

  Apr 2024, SeaLLMs-7b-v2.5 - released

  Jul 2024, SeaLLMs-7b-v3 - released
- ❑ Won "Best Innovate for Impact Award" by ITU of United Nations
- ❑ Explore more at: https://damo-nlp-sg.github.io/SeaLLMs/

DEMO

SeaLLMs – How It’s Built

❑ Starts with English-centered open-source base model

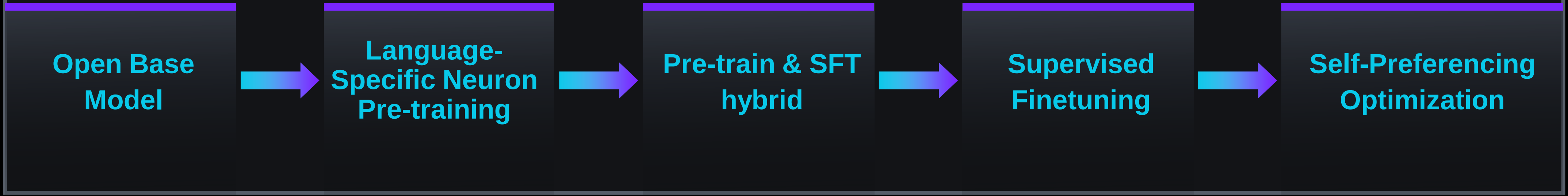
❑ Language-specific neuron pretraining

❑ Pretrain & SFT hybrid

❑ Supervised finetuning (SFT)

❑ Self-preferencing optimization

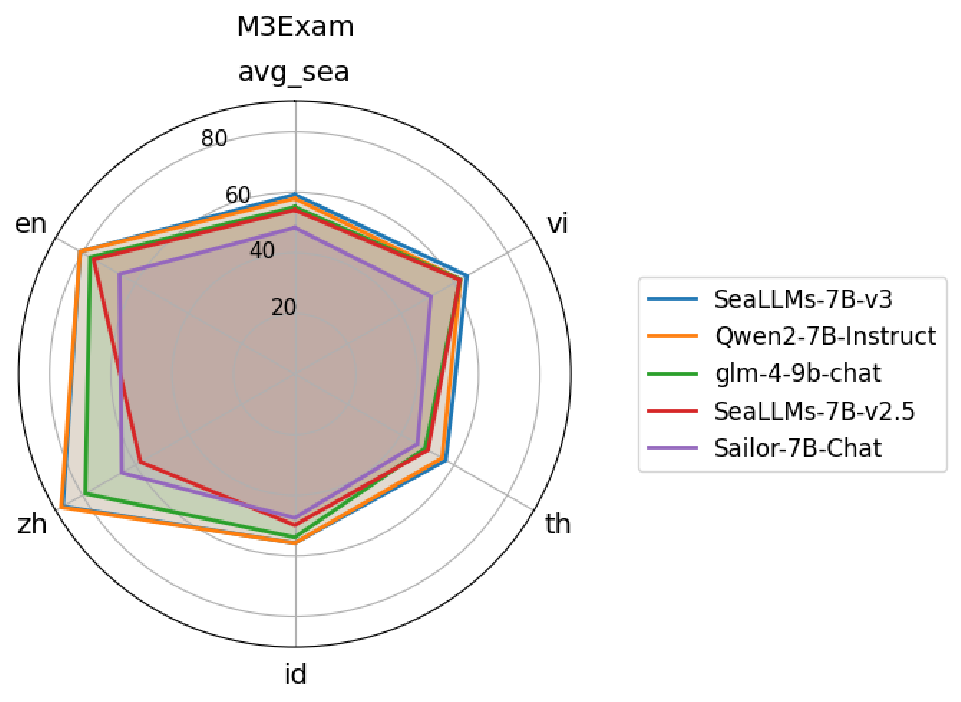
Technical Report

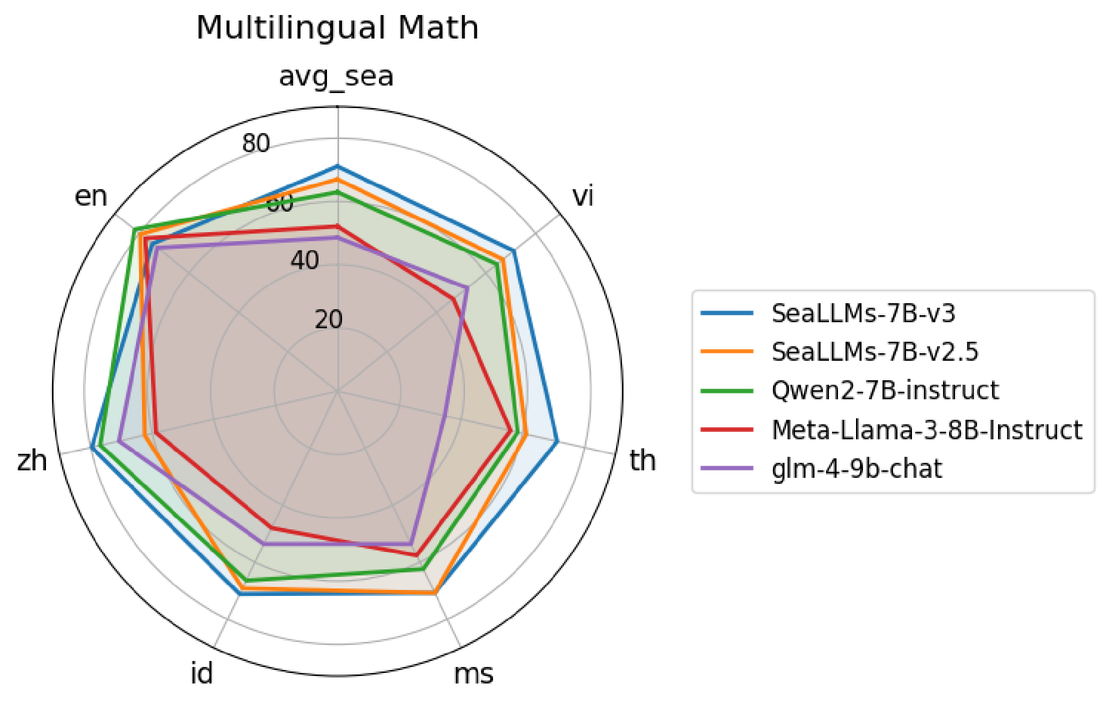


Open Base Model → Language-Specific Neuron Pre-training → Pre-train & SFT hybrid → Supervised Finetuning → Self-Preferencing Optimization

SeaLLMs – How It Performs

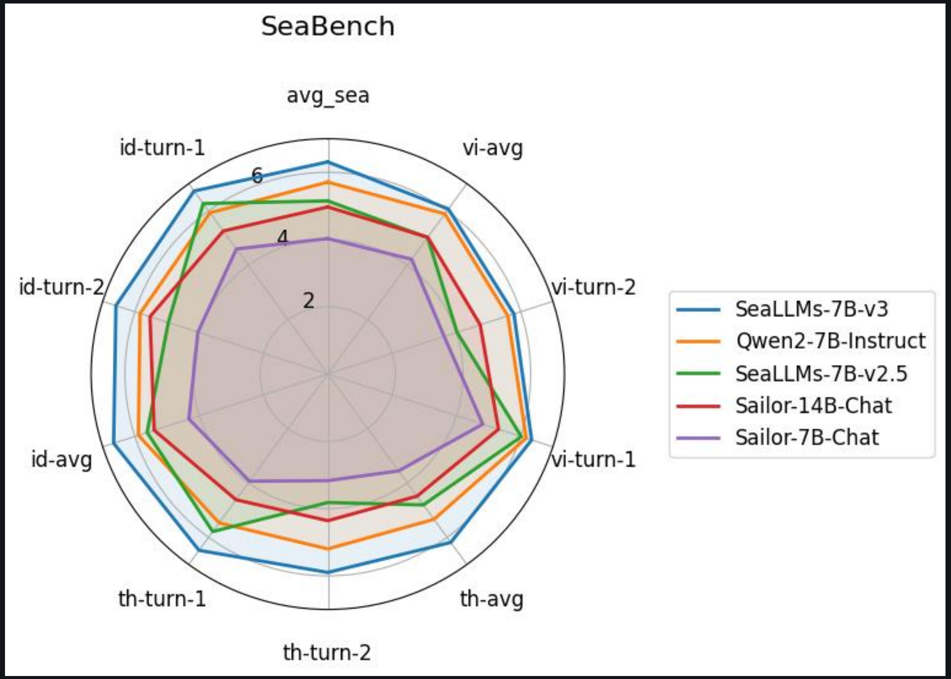
World Knowledge & Math



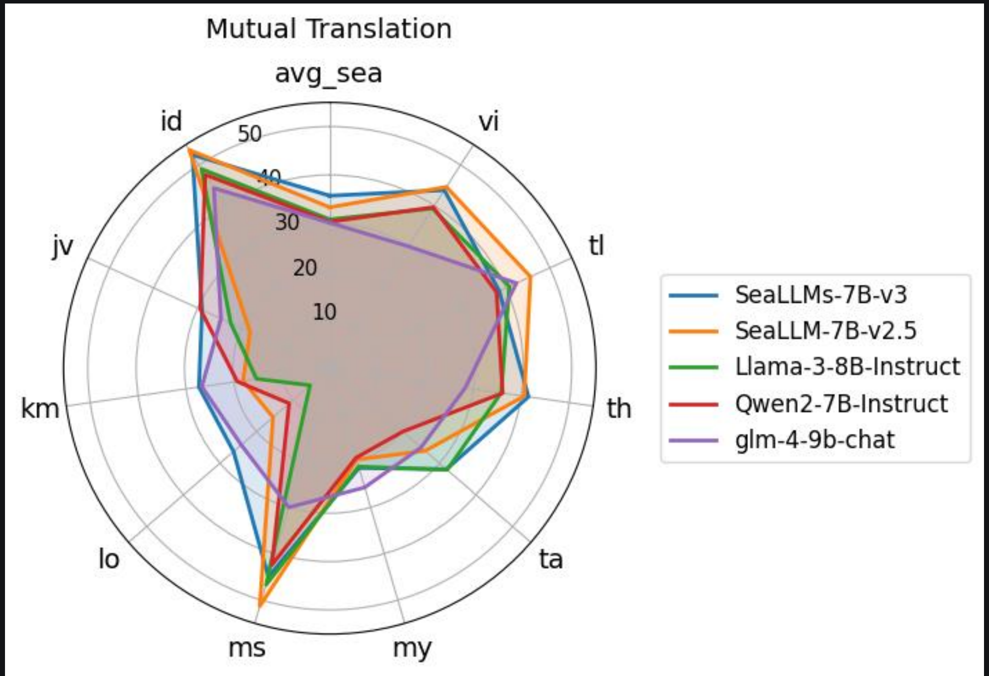


SeaLLMs – How It Performs

Multi-turn & Translation



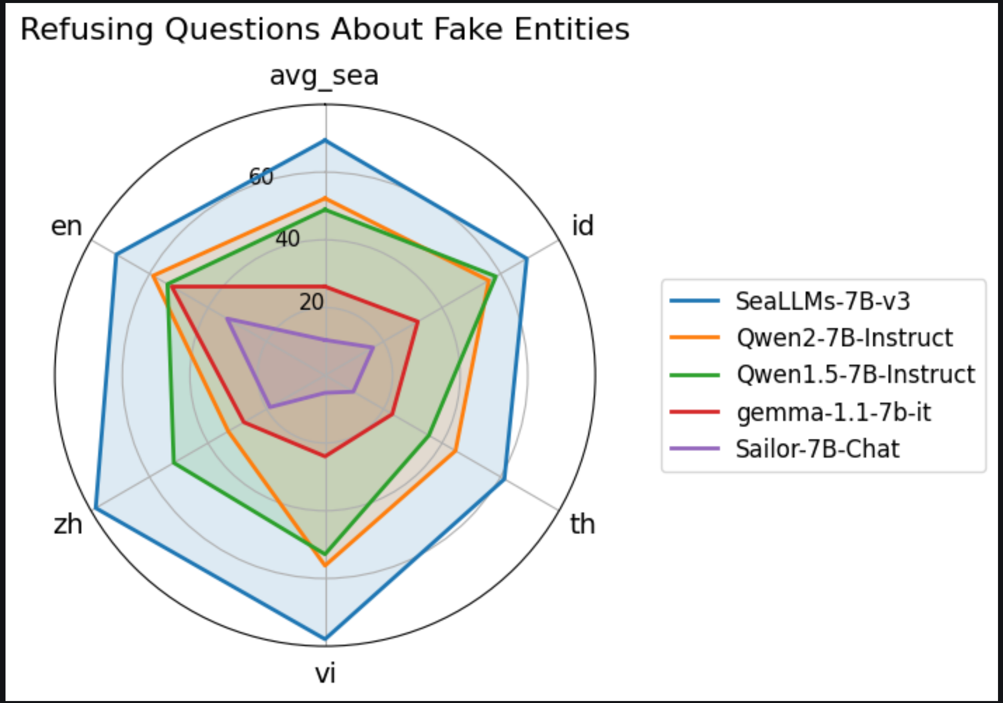
SeaBench



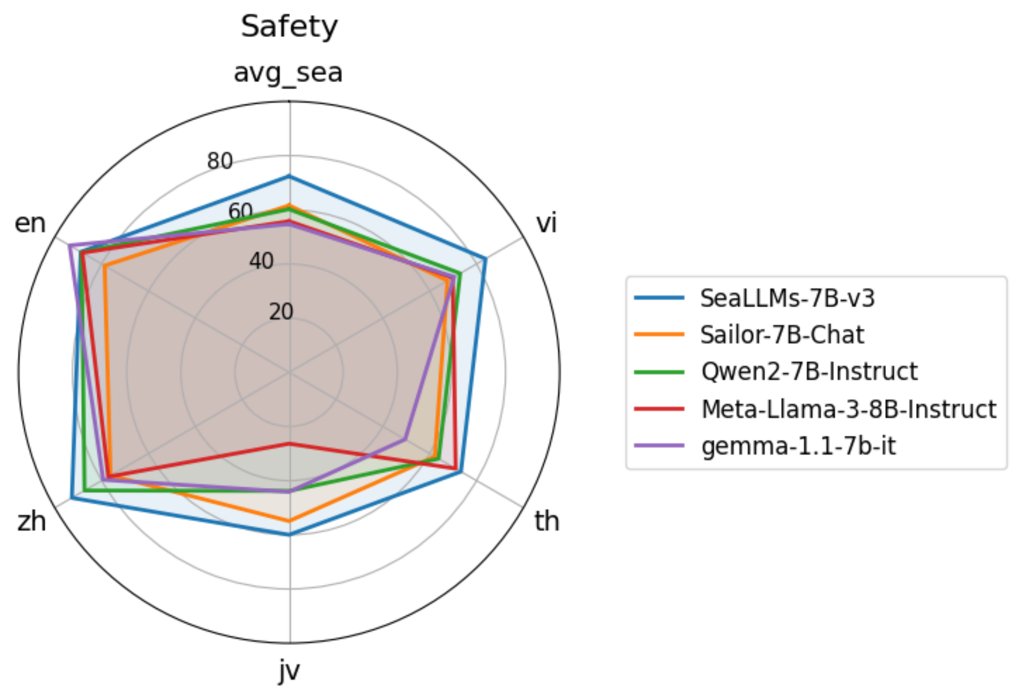
Mutual Translation

SeaLLMs – How It Performs

Hallucination & Safety



Refusing Questions About Fake Entities

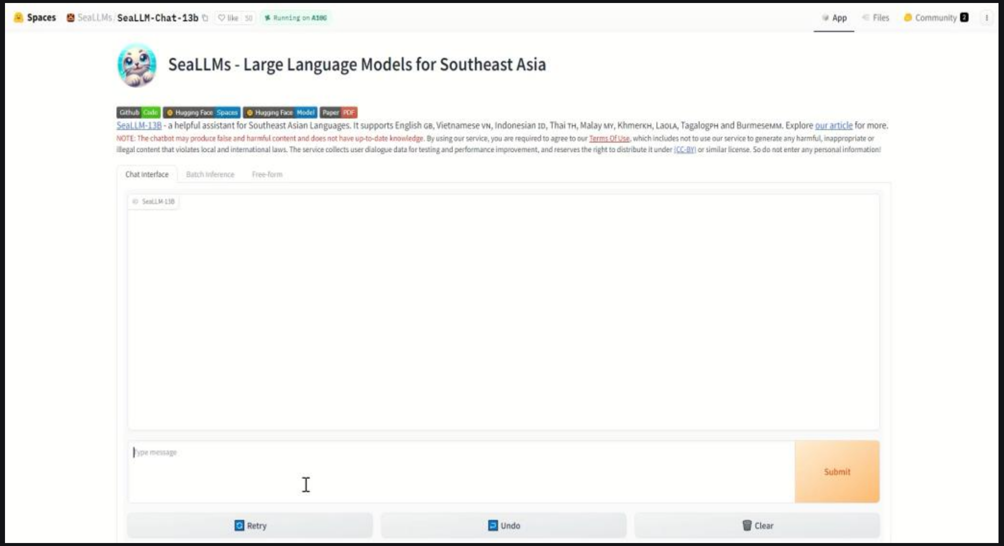


Safety

達摩院
DAMO ACADEMY

Give it a try!

Downloads at HuggingFace: 120K+ (Jul, 2024)



SeaLLMs - Large Language Models for Southeast Asia

SeaLLM-13B - a helpful assistant for Southeast Asian Languages. It supports English GB, Vietnamese VN, Indonesian ID, Thai TH, Malay MY, KhmerKH, LaoLA, TagalogPH and BurmeseMM. Explore our article for more.

NOTE: The chatbot may produce false and harmful content and does not have up-to-date knowledge. By using our service, you are required to agree to our Terms Of Use, which includes not to use our service to generate any harmful, inappropriate or illegal content that violates local and international laws. The service collects user dialogue data for testing and performance improvement, and reserves the right to distribute it under (CC-BY) or similar license. So do not enter any personal information!

DEMO

Technical Report

THANK YOU