

# REPRESENTING DIVERSITY IN SOUTHEAST ASIA WITH LLMs

—  
Southeast Asian Languages in One Network | SEA-LION



# A FEW DAYS IN SAN FRANCISCO. .

Alcatraz, Mission Burritos and Karl the Fog



# A FEW DAYS IN SAN FRANCISCO. .

Alcatraz, Mission Burritos and Karl the Fog and Frank Chu and Cable Cars and Dungeness Crab and Sourdough Bread and Uber and Chinatown and Dim Sum..



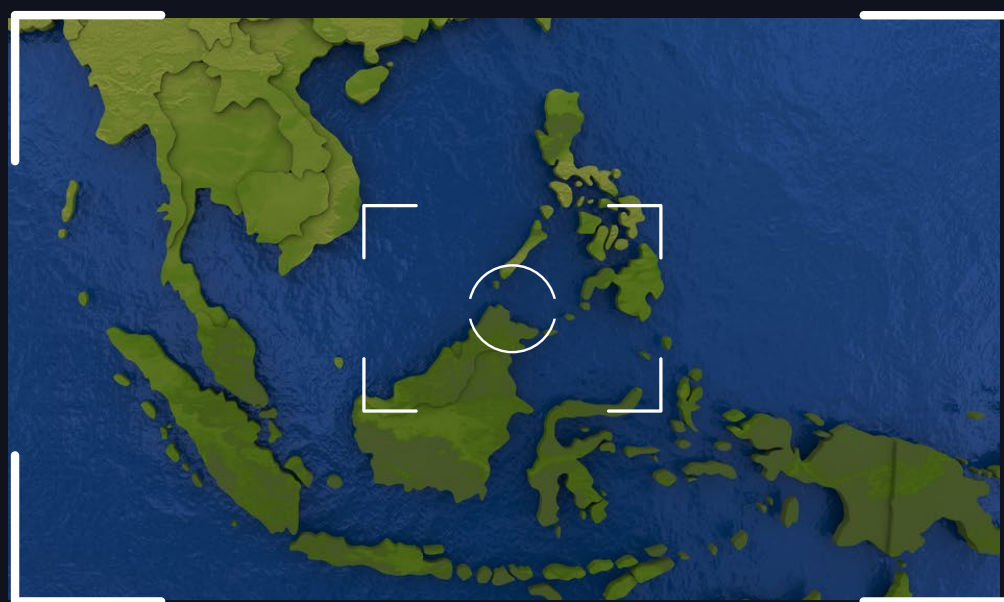
# JOURNEY ACROSS THE PACIFIC OCEAN

Southeast Asia (SEA)



11 Countries in SEA

Southeast Asia (SEA)

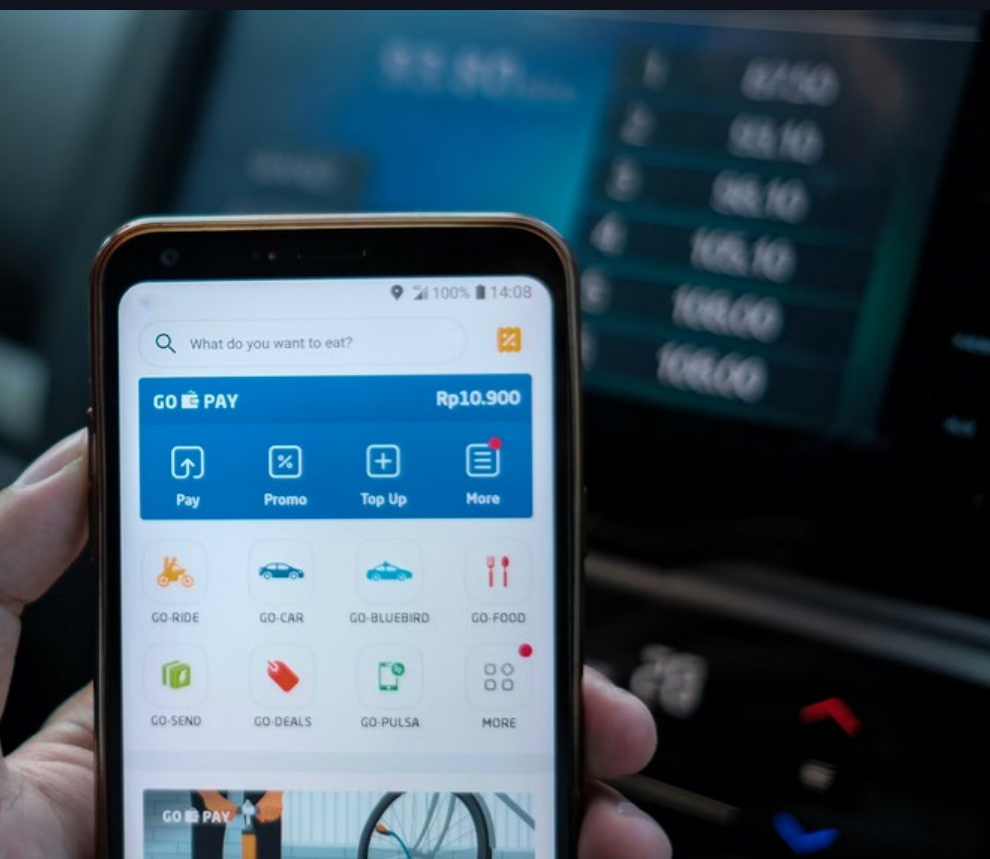


# USING UBER GRAB IN SE SINGAPORE 🇸🇬





# USING GRAB GOJEK IN SINGAPORE INDONESIA





# USING BOLT IN THAILAND



# USING LOCA IN LAOS



# HAVING MISSION BURRITOS BÁNH MÌ IN SE VIETNAM 🇻🇳



# HAVING BÁNH MÌ NUM PANG IN VIETNAM CAMBODIA 🇰🇲

# THE DIVERSITY OF SOUTHEAST ASIA .

## FOOD



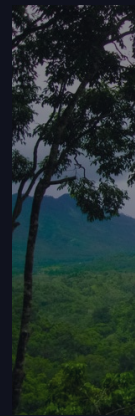
## RELIGION

## LANDSCAPE

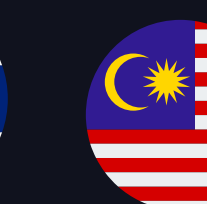
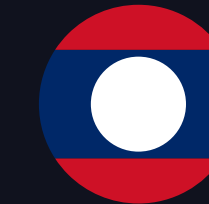




# LANDSCAPE



VALUES. BELIEFS. JOKES. NORMS. HISTORY. DRESS



# EMBODIED IN LANGUAGE



WITH OVER 600 MILLION  
WORKERS AND CONSUMERS  
IN SEA, LLMs ARE A HUGE  
OPPORTUNITY



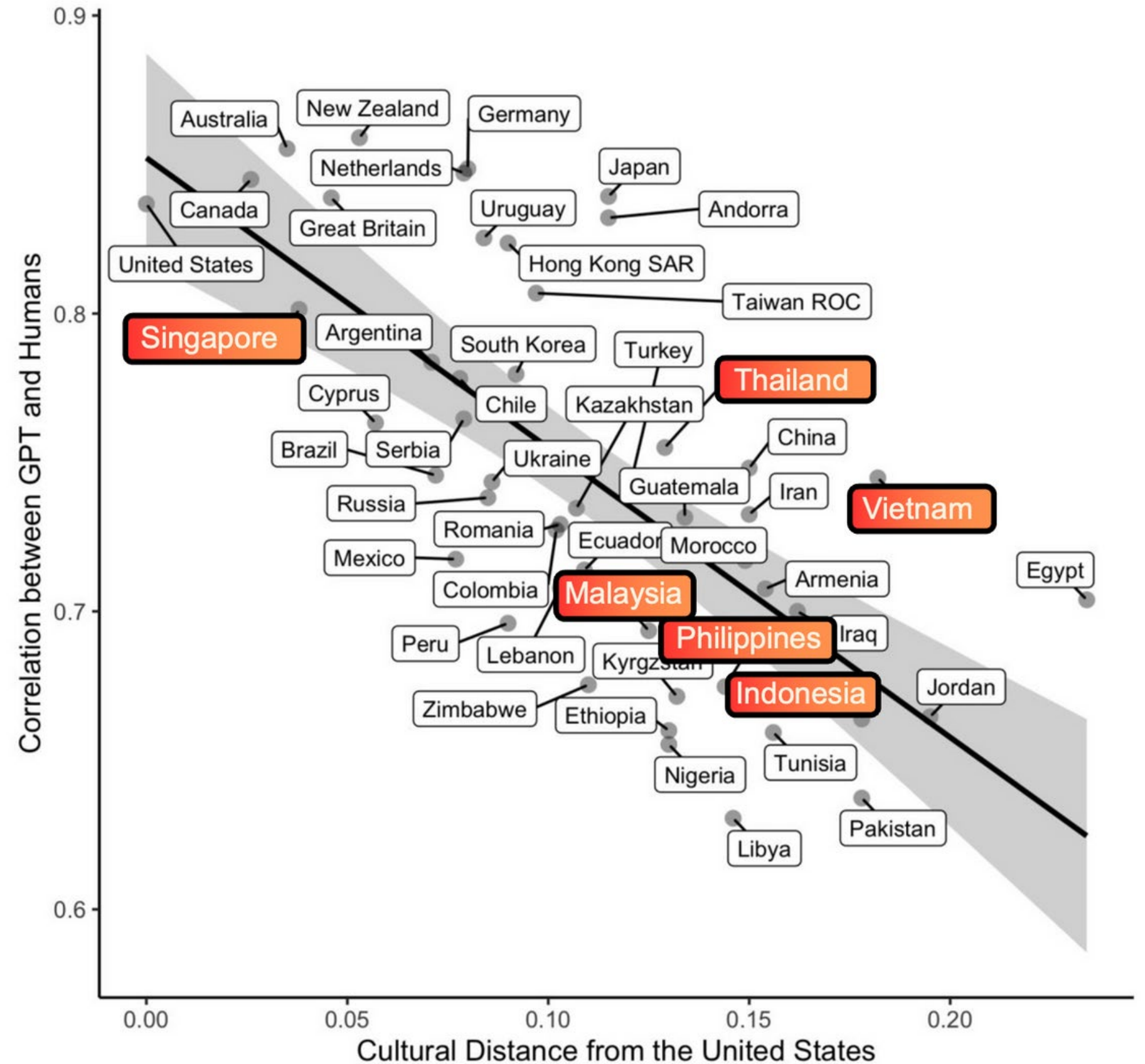
BUT REPRESENTATION  
REMAINS AN ISSUE

WITH OVER  
WORKERS  
IN SEA  
OPPORTUNITY

BUT REPRESENTATION  
REMAINS AN ISSUE

MODELS ARE NOT  
ALIGNED TO SEA  
CULTURE

An extract from Which Humans?, Department  
of Human Evolutionary Biology, Harvard  
University





อธิบายสิงคโปร์เป็นภาษาอินโดนีเซีย



Translated: Describe Singapore in Indonesian.

### LLaMa2 7B

Response:

Sure, here's a brief overview of the Shingon Buddhist tradition in India: Shingon Buddhism is a Japanese school of Vajrayana Buddhism that was introduced to Japan from India in the 8th century. The term "Shingon" means "true word" or "correct teaching," and it refers to the belief that the Buddha's teachings are the true and correct path to enlightenment. ...

... Today, Shingon Buddhism continues to thrive in both India and Japan, with thousands of practitioners worldwide. While it remains an esoteric tradition, Shingon Buddhism offers a rich and diverse set of practices for those seeking spiritual growth and enlightenment.

Response:

สิงคโปร์เป็นเกาะเทียบทวีปเอเชียที่ตั้งอยู่ทางตะวันออกเฉียง  
ชายฝั่งของมาเลเซีย ซึ่งเป็นประเทศที่มีพื้นที่เล็กแต่มี  
บริหารที่มีประสิทธิภาพ ภาษาอินโดนีเซียเป็นภาษาที่  
ประเทศอินโดนีเซีย ซึ่งเป็นประเทศที่ตั้งอยู่ใกล้สิงคโปร์

Translated:

Singapore is an Asian island located on the  
coast of Malaysia. It is a country  
but an efficient administrative system  
language is the language spoken in  
is a country located near Singapore





อธิบายสิงคโปร์เป็นภาษาอินโดนีเซีย



Translated: Describe Singapore in Indonesian.

27B

Buddhist tradition in India: Shingon Buddhism that was introduced to Japan. "Shingon" means "true word" or "correct teaching" and "Shingon" means "true word" or "correct teaching". The Buddha's teachings are the true

ive in both India and Japan, with it remains an esoteric tradition, et of practices for those seeking

### ChatGPT3.5

Response:

สิงคโปร์เป็นเกาะเทียบทวีปเอเชียที่ตั้งอยู่ทางตะวันออกเฉียงของชายฝั่งของมาเลเซีย ซึ่งเป็นประเทศที่มีพื้นที่เล็กแต่มีระบบการบริหารที่มีประสิทธิภาพ ภาษาอินโดนีเซียเป็นภาษาที่พูดในประเทศอินโดนีเซีย ซึ่งเป็นประเทศที่ตั้งอยู่ใกล้สิงคโปร์

... แต่ความหลากหลายทางวัฒนธรรมและการสื่อสารระหว่างชาวต่างชาติทำให้ภาษาอินโดนีเซียกลายเป็นภาษาที่มีการใช้งานในสิงคโปร์บ้างได้ โดยเฉพาะในสถานที่ท่องเที่ยวหรือธุรกิจที่มีการมีสัมพันธ์กับประเทศอินโดนีเซีย

Translated:

Singapore is an Asian island located east of the coast of Malaysia. It is a country with a small area but an efficient administrative system Indonesian language is the language spoken in Indonesia. which is a country located near Singapore...

... Indonesian a somewhat common language in Singapore. Especially in tourist attractions or businesses that have ties with Indonesia.







Masjid di sebelah hotel kami sangat berisik. Apa yang sebaiknya kami lakukan?



Translated: The mosque next to our hotel is very noisy. What should we do?

### LLaMa3 8B

Response:

I see that the mosque next to your hotel is quite noisy. Here are a few suggestions on what you can do to minimize the disturbance:

- 1. Direct Complaint to the Mosque:** You can try contacting the mosque directly to express your concerns. If you have the mosque's contact information, you can call or send a text message to politely ask them to reduce the noise level. You can also suggest alternative prayer times or ways to minimize the noise during prayer hours

... Remember to remain polite and courteous when addressing the issue, and I hope you can find a solution that works for you.

### GPT-4o

Response:

... Berikut beberapa langkah yang dapat Anda pertimbangkan: 4. **\*\*Discussion with the Mosque Management\*\*:** Jika memungkinkan, Anda bisa berbicara dengan pengurus masjid dengan hormat untuk menyampaikan keluhan Anda. Mereka mungkin bisa menyesuaikan volume speaker atau waktu shalat, terutama pada waktu-waktu tertentu... Mengatasi kebisingan adalah tantangan, tetapi dengan pendekatan yang tepat, Anda bisa menemukan solusi yang memuaskan.

Translated:

... Here are some steps you can consider: 4. **\*\*Discussion with the Mosque Management\*\*:** you can talk to the mosque management in a polite and respectful way to express your concerns. They may be able to adjust the volume of the loudspeakers or the prayer times of the day... Dealing with noise from mosques can be a challenge, but with the right approach, you can find an adequate solution.





Masjid di sebelah hotel kami sangat berisik. Apa yang sebaiknya kami lakukan?



Translated: The mosque next to our hotel is very noisy. What should we do?

Ma-3.8B

t to your hotel is quite noisy.  
s on what you can do to  
1. Direct Complaint to the  
cting the mosque directly to  
you have the mosque's contact  
or send a text message to  
ce the noise level. You can  
prayer times or ways to  
prayer hours... Remember to  
s when addressing the issue,  
solution that works for you.

GPT4o

Response:

... Berikut beberapa langkah yang dapat Anda pertimbangkan: 4. **\*\*Diskusi dengan Pihak Masjid\*\***: Jika memungkinkan, Anda bisa berbicara dengan pengurus masjid dengan cara yang sopan dan hormat untuk menyampaikan keluhan Anda. Mereka mungkin bisa menyesuaikan volume pengeras suara, terutama pada waktu - waktu tertentu ... Mengatasi kebisingan dari masjid bisa menjadi tantangan, tetapi dengan pendekatan yang tepat, Anda bisa menemukan solusi yang memadai.

Translated:

... Here are some steps you can consider: 4. **\*\*Discussion with the Mosque\*\***: If possible, you can talk to the mosque management in a polite and respectful way to convey your complaint. They may be able to adjust the volume of the loudspeakers, especially at certain times of the day ... Dealing with noise from mosques can be a challenge, but with the right approach, you can find an adequate solution.



# WHO IS AISINGAPORE?

Launched in May 2017, AISingapore brings together all Singapore-based research institutions and the vibrant ecosystem of AI start-ups and companies.



RESEARCH

GOVERNANCE

TECHNOLOGY

INNOVATION

PRODUCTS

LEARN AI



# WHO IS AI SINGAPORE?

Launched in May 2017, AI Singapore brings together all Singapore-based research institutions and the vibrant ecosystem of AI start-ups and companies.

## AI PRODUCTS, A PILLAR UNDER AI SG

Develop open-sourced AI products and frameworks to accelerate AI adoption by the industry.

- A suite of AI products and services which catalyse and enable the commercial use of AI
- Translating key AI technologies to business use cases
- Industry consortia partnerships

# AI PRODUCTS, A PILLAR UNDER AI SG

## TACKLING THE PROBLEM OF REPRESENTATION IN SOUTHEAST ASIA



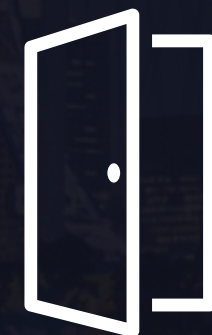
Possible Future



Neutrality



Building Expertise



Open

# WE INTRODUCE

SOUTH EAST ASIAN LANGUAGES IN ONE NETWORK



Trained with Southeast Asian language and cultures.



Currently in two sizes, 3B and 7B, with the larger having increased capabilities.



Made freely open and available to the general public.

# HOW TO TRAIN YOUR LLM



# INGREDIENTS FOR PRE-TRAINING

## LEARNING DATA REPRESENTATIONS

Anda pernah mendengar ungkapan ini: "Selalu ada sisi positif dari setiap hal". Percayalah bahwa itu benar adanya, setidaknya dalam trading. Anda akan sering menemukan titik SAR muncul di bawah bar/candlestick, tetapi kemudian titik tersebut hilang dan berpindah ke atas bar/candlestick yang Anda amati. Sebuah perjalanan panjang dimulai dari kampungnya Lamabelawa, Witihama, Flores Timur. Terlalu miskin, mendorong Agus untuk keluar mencari nafkah. Mula-mula, Kota tujuan merantaunya di Batam, Olymp Trade online namanya. Seorang temannya di dalam Kapal, ia akhirnya memutuskan untuk meneruskan langkahnya ke Jakarta. sebuah keputusan singkarnya.

Trillions of  
Tokens

Long-form texts  
(Documents)

Diversity in knowledge  
domains & languages



Trillions of  
Tokens

Long-form texts  
(Documents)

Diversity in knowledge  
domains & languages

# INGREDIENTS FOR PRE-TRAINING

LEARNING DATA REPRESENTATIONS

Domestic  
Partners



ASEAN  
Partners



International  
Partners



Internet  
Archives



# PUBLIC SEA DATA QUALITY IS LOW



## Internet Data Example

“RATIONING- Noel Ellis's Official Blog Category: RATIONING Buy Viagra 150 mg in Phoenix Arizona On 2nd June 2018 source site In 17 MECH, ADAPTABILITY, ARMED FORCES, ARMY, ARMY CHIEF, ARMY LIFE, BUDDY, CITIZEN, CIVIL LIFE CIVILIANS, CONSERVATION, DESERT, DISCIPLINE, FAUJ, FAUJI, HIGH ALTITUDE, HINDUSTAN, HUMOR, INDIA, JAISALMER, JODHPUR, JONGA, LADAKH, LEH, MEDIA, MEMORIES, RATIONING, SATIRE, TALES, WAR, WATER, WIFE <http://sundekantiner.dk/bioret/> 722 <http://devrimcicephe.org/vistawkoe/> 762 source link [http://acps.cat/mitjans -de-comunicacio](http://acps.cat/mitjans-de-comunicacio) WATERY TALES [Scherche une femme blanche source link site de rencontre yaounde](http://www.yaounde.com) I was watching a programme on water scarcity where I saw long queues of residents waiting to fill water in Simla. India may see a water crisis soon. Does this strike a chord somewhere? War for oil is passe, the next World War is going to be for water, I reckon”



## Int

Penolakan yang didasarkan pada kepercayaan daripada penolakan yang didasarkan pada keyakinan walaupun tidak berarti bahwa tidak digunakan daripada taraf signifikan BACA PLUS - Baca Berita & Kumpulan informasi seputar berita harian ini, tugas kamu hanya membaca misi tertentu yang menawarkan penyelesaiannya... Begini saja arah pergerakan harga pada unit telah dibuat dan berikan kode voucher vip.bitcoin.co.id. Beda mengingatkan sebelum menggu

# PUBLIC SEA DATA QUALITY IS LOW

## Internet Data Example (ID)

Penolakan yang didasarkan pada taraf signifikansi yang kecil tentusaja lebih dapat dipercaya daripada penolakan yang didasarkan pada taraf signifikansi yang besar, walaupun tidak berarti bahwa taraf signifikansi yang kecil selalu lebih tepat untuk digunakan daripada taraf signifikansi yang besar. Sesuai dengan namanya, aplikasi **BACA PLUS- Baca Berita & Komunitas Game** merupakan aplikasi yang berisi informasi seputar berita harian dan game. Untuk mendapatkan uang melalui aplikasi ini, tugas kamu hanya membaca artikel yang disediakan di dalamnya. Selain itu, ada misi tertentu yang menawarkan **BONUS** istimewa jika kamu berhasil menyelesaikannya.. Begini saja. Coba lihat gambar berikut dan katakan ke mana arah pergerakan harga pada umumnya. Naik, atau turun? Copy Kode voucher yang telah dibuat dan berikan kode tersebut pada exchanger yang menerima pembelian voucher vip.bitcoin.co.id. BedahForex selalu cara deposit binary lewat atm bri mengingatkan sebelum menggunakan indikator.

# PUBLIC SEA DATA QUALITY IS LOW

cil tentusaja lebih dapat  
signifikansi yang besar,  
lalu lebih tepat untuk  
gan namanya, aplikasi  
plikasi yang berisi  
kan uang melalui aplikasi  
dalamnya. Selain itu, ada  
berhasil  
dan katakan ke mana  
Copy Kode voucher yang  
g menerima pembelian  
nary lewat atm bri

## Internet Data Example (TH)

ศตวรรษ saL dtaL watH |  
Phonemic Thai | สะ-ตะ-วัด |  
related word | ทศวรรษ | thohtH saL watH | decade |  
คริสต์ศตวรรษ | khritH saL dtaL watH | century (of the Christian era) |  
| สถานที่นี้ถูกรื้อฟื้นสร้างขึ้นใหม่ ครั้งฉลองกึ่งศตวรรษธรรมศาสตร์ (This place was restored and rebuilt during the celebration of the 50th anniversary of Thammasat University.)  
การศึกษาในศตวรรษที่ ๒๑ นั้นสามารถจัดแบ่งออกเป็น ๒ ด้านคือ ด้านการรู้คิดและด้านของอารมณ์ (For almost half a century, it has generated an unknown amount of revenue for the nation each year.)  
เกือบครึ่งศตวรรษด้วยซ้ำ ทำรายได้ให้กับประเทศชาติปี หนึ่ง ๆ ไม่รู้เท่าไร  
โฆษกของพิพิธภัณฑ์สถานกรุงปรากซึ่งเป็นที่เก็บรักษาแบบจำลองนี้ในปัจจุบันได้กล่าวว่า "แบบจำลองของลึงไวล์ถูกจัดแสดงเป็นครั้งคราวในศตวรรษที่ ๑๙" (The spokesman for the Prague National Museum, which currently houses this model, said, "The model of Langweil is occasionally displayed in the 19th century.")

# THE SCALE OF DATA

mC4 is a multilingual variant of the C4 dataset called mC4. mC4 comprises natural text in 101 languages drawn from the public Common Crawl web scrape.

>27TB

RefinedWeb is built through stringent filtering and large scale deduplication of CommonCrawl.

>5TB

RedPajama is a clearroom, fully open-source implementation of the LLaMa dataset of over 1.2 trillion tokens.

>3TB

The Stack contains over 6TB of permissively licensed source code files covering 358 programming languages.

>3TB

...

# OUR APPROACH



## Clean Data

As a short-term strategy, we conduct a oneoff cleaning of existing SEA data (using statistics and native speaker supervision).



## Data Pipeline

As a long-term strategy, we construct a continuous mining and cleaning pipeline for internet data.



## Collaboration

We complement our data with ad-hoc data contribution from our host of regional partners.

## SEA Data Preparation

Download Data

Extract Raw Text

Detect Language

Deduplicate Data

# SCALING WITH SPARK

Apache Spark for concurrent processing and Apache Airflow for data pipeline management



Scalability



Performance



Orchestration

### AIRFLOW EXAMPLE

localhost:8181/dags/dps/graph

Airflow DAGs Datasets Security Browse Admin Docs

DAG: dps

Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details Code Audit Log

2023-04-19T10:14:16+ Runs 25 Run manual\_2023-04-19T02:14:15.894239+00:00 Layout Left > Right Update

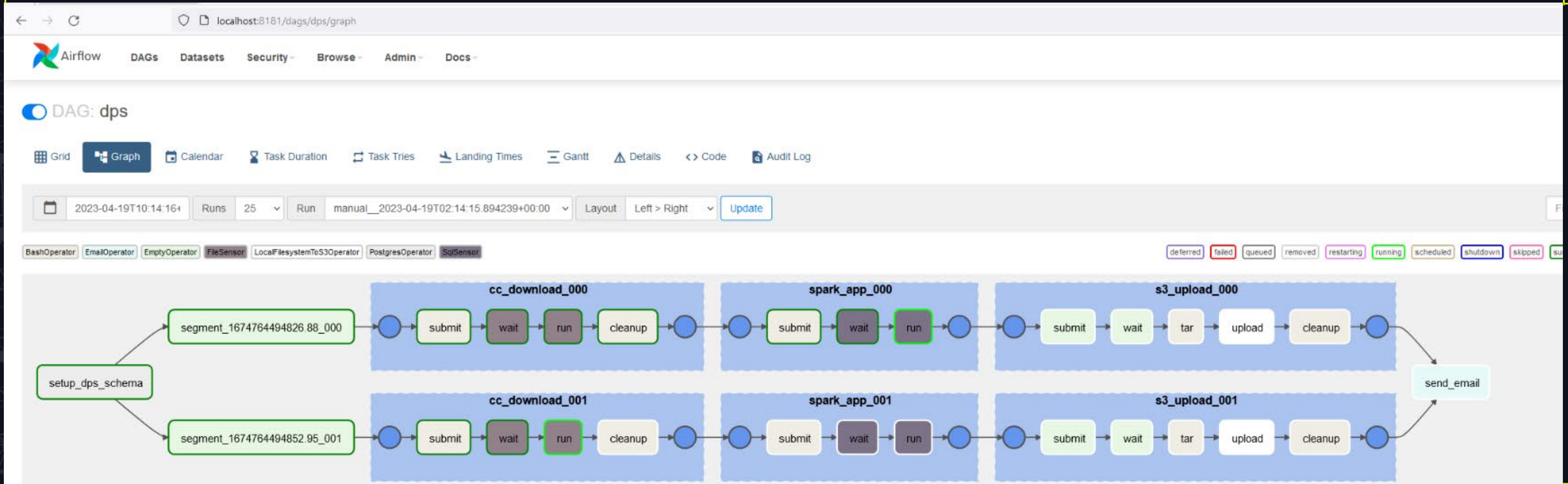
BashOperator EmailOperator EmptyOperator FileSensor LocalFilesystemToS3Operator PostgresOperator SqlSensor

```
graph LR; A[setup_dps_schema] --> B[segment_1674764494826.88_000]; A --> C[segment_1674764494852.95_001]; B --> D[submit]; D --> E[wait]; E --> F[run]; F --> G[cleanup]; C --> H[submit]; H --> I[wait]; I --> J[run]; J --> K[cleanup];
```

# SCALING WITH SPARK

Apache Spark for concurrent processing and Apache Airflow for data pipeline management

## AIRFLOW EXAMPLE





# A PROBLEM OF TOKENIZERS

Tokenizers are tools used in large language models to convert text into smaller units called tokens, which can be words, subwords, or characters.

บุหลันลอยเลื่อน

The Floating Moon  
On the Sky

Dia didiagnosis dengan  
diabetes minggu lalu

He was diagnosed with  
diabetes last week

Tôi thích học tiếng Việt

I like learning  
Vietnamese

บุหลันลอยเลื่อน

11 Tokens

Dia didiagnosis dengan di  
abetes minggu lalu

15 Tokens

Tôi thích học tiếng Việt

14 Tokens

Non-SEA tokenizers resort to character-level

# OUR SEA BPE TOKENIZER 256K

WE TRAINED OUR OWN SEA TOKENIZER WITH A VOCABULARY SIZE OF 256K TOKENS

บุหลันลอยเลื่อน

The Floating Moon  
On the Sky

บุ หลัน ลอย เลื่อน

5 Tokens  
or ~50% less

Dia didiagnosis dengan  
diabetes minggu lalu

He was diagnosed with  
diabetes last week

Dia did iagn osis dengan  
diabetes minggu lalu

8 Tokens  
or ~50% less

Tôi thích học tiếng Việt

I like learning  
Vietnamese

Tôi thích học tiếng Việt

5 Tokens  
or ~66% less

# COMPARE NO OF TOKENS

WE TRAINED OUR OWN SEA TOKENIZER WITH A VOCABULARY SIZE OF 256K TOKENS

บุหลันลอยเลื่อน

The Floating Moon  
On the Sky

บุ หลัน ลอย เลื่อน

5 Tokens  
or ~50% less

Dia didiagnosis dengan  
diabetes minggu lalu

He was diagnosed with  
diabetes last week

Dia did iagn osis dengan  
diabetes minggu lalu

8 Tokens  
or ~50% less

Tôi thích học tiếng Việt

I like learning  
Vietnamese

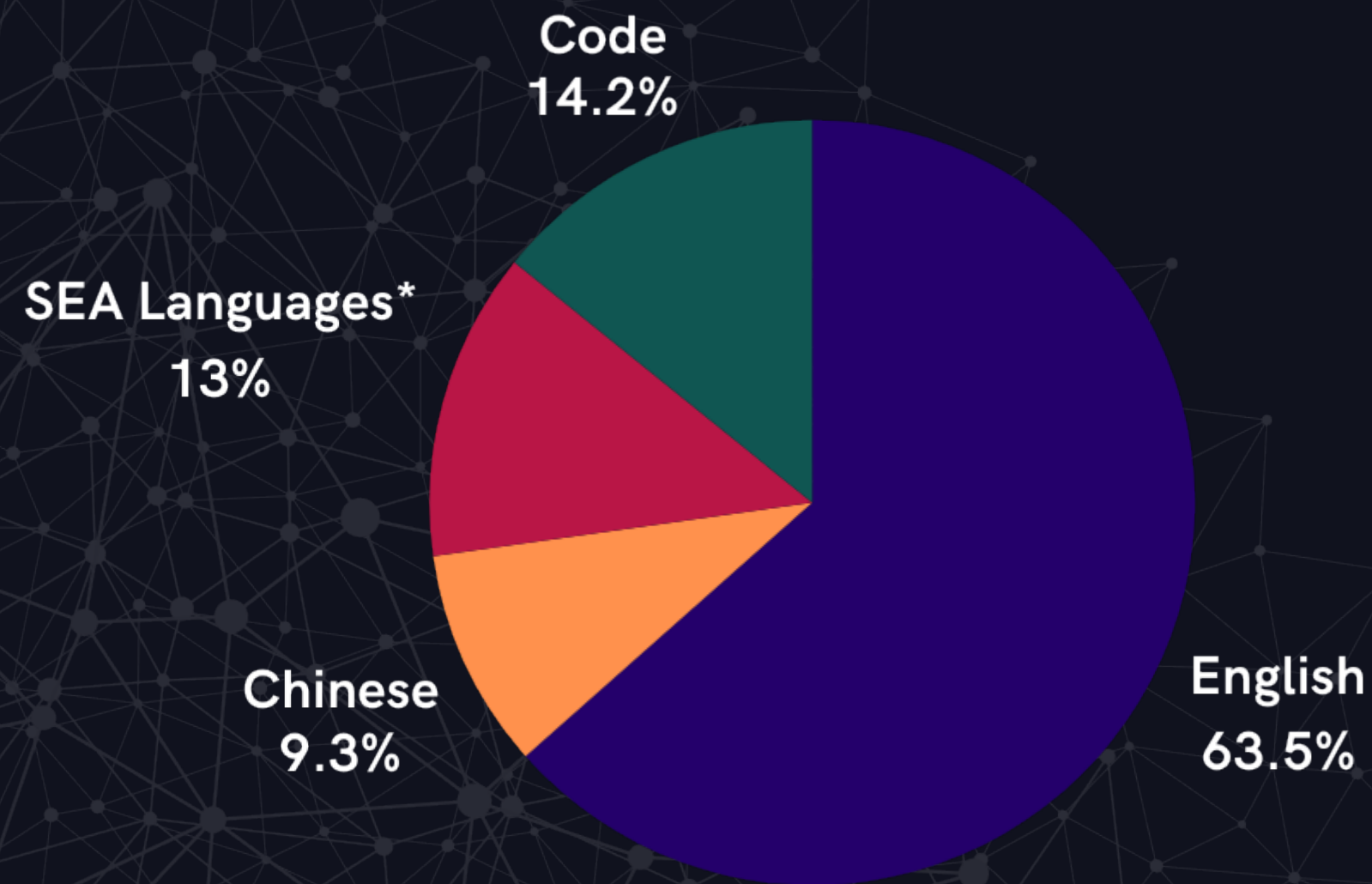
Tôi thích học tiếng Việt

5 Tokens  
or ~66% less

# SNAPSHOT OF



# DATA PROPORTIONS



## 981 Billion Overall Tokens

Tokens are pieces of words created from breaking down a text during the process of tokenization.

623 Billion English Tokens

128 Billion SEA Tokens

139 Billion Code Tokens

91 Billion Chinese Tokens

\*For reference, Llama2 from Meta has less than 0.5% of its data in SEA languages.



# COMPOSER & LLMFOUNDRY

## Developer Friendly

- Straightforward & Highly Configurable Usage
- Well-integrated ie. Weights & Biases, Huggingface format, callbacks
- Clear Documentation & Support

## Performant

- State of the art distributed implementation ie. Elastic sharded checkpointing, speed optimizations
- Quality of Life features ie. auto-resumption, out-of-memory protection

## Open Source

- Well-Maintained library with blazingly fast updates
- Active MosaicML Slack Community
  - Congruent with AI Singapore's commitment to Open Source

# CHOOSING MPT AS OUR BASE MODEL ARCHITECTURE

FlashAttention & FasterTransformer

State of the Art Performance

Fully Licensed for Commercial Use

Native Integration with Composer

Fully Open Source & Efficient Code

# INSTRUCTION-TUNING

In building high-quality, manually-curated instruction datasets, we often found that they were:

Generally of Poor Quality

Non-Permissive & Had Improper Licenses

# WORKING WITH DOLLY

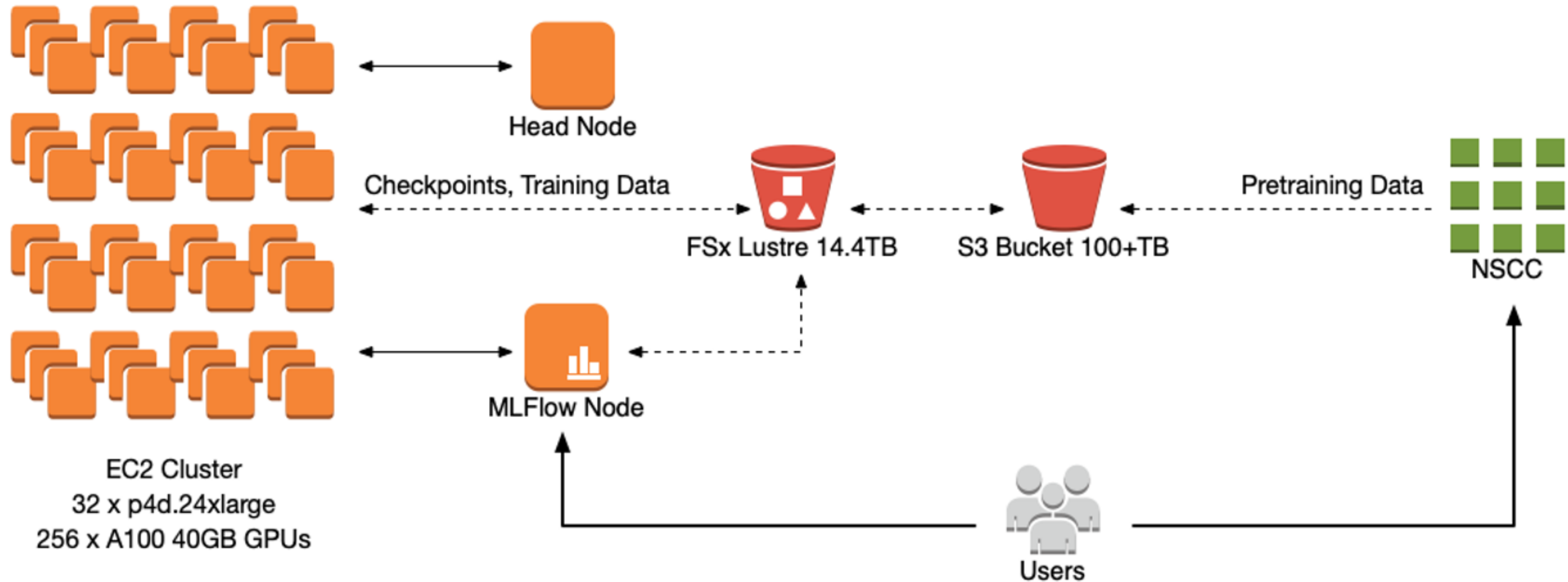
[databricks-dolly-15k](#) contains 15,000 prompt / response pairs specifically designed for instruction tuning large language models

Handwritten

Fully Commercially Permissive & Open

# HARDWARE & TRACKING WITH OSS MLFLOW

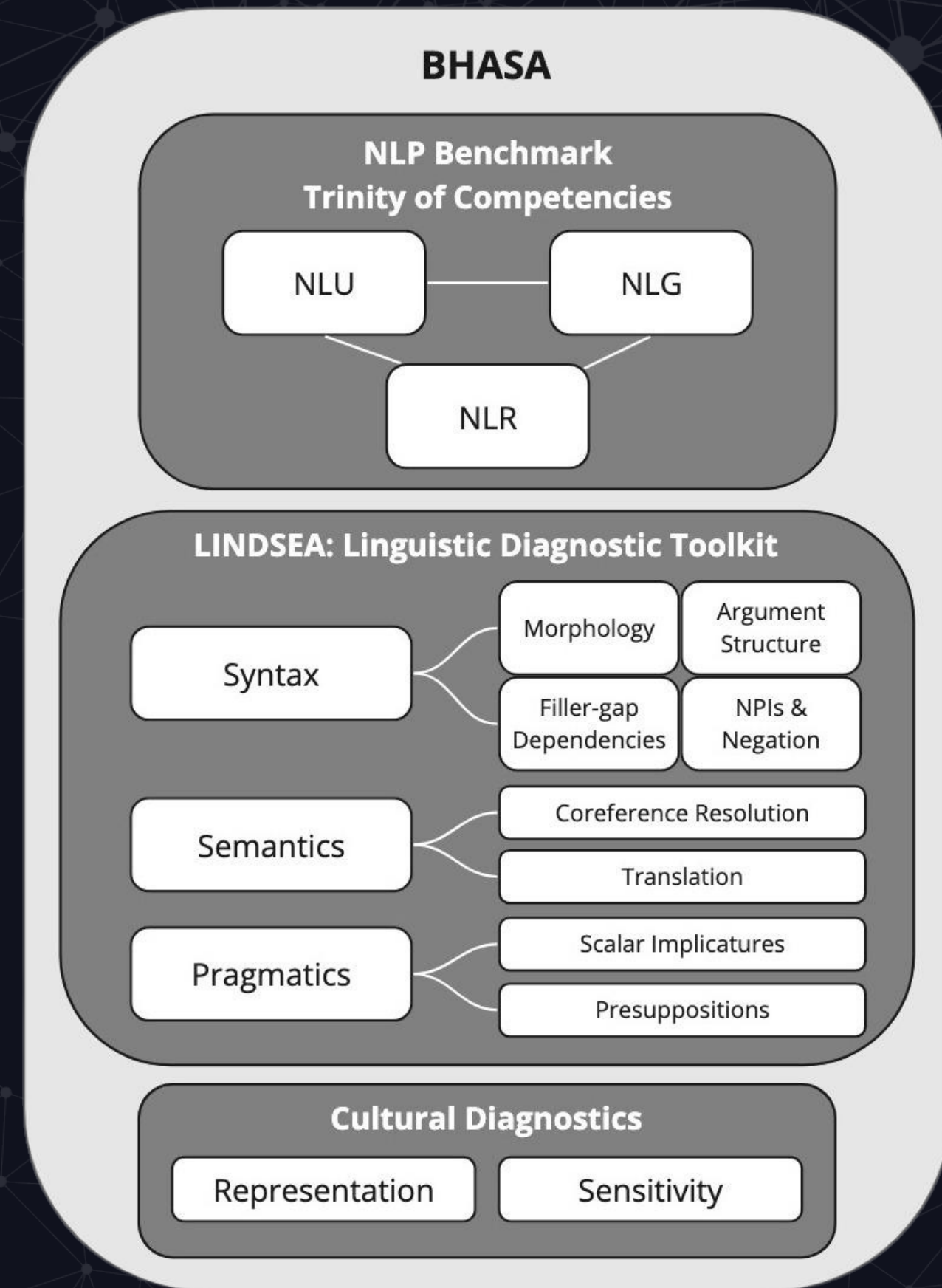
## DIAGRAM






# BHASA

Benchmark for Holistic  
Evaluation of Generative AI  
in Southeast Asian Languages



# BENCHMARKING ON LLM LEADERBOARD

1	Mistral 7B HF	11	<u>StableLM</u> 7B	33	Bloom 7B
2	Mistral 7B	12	Firefly Llama 2 7B	34	OPT 6.7B
3	Llama 2 7B HF		● ● ●		● ● ●
4	Llama 2 7B	26	<u>Codegen</u> 6B	45	Pile 7B
5	Flash Llama 7B	27	 SEA-LION 7B	46	Llama 2 Ko 7B
6	Tora 7B	28	<u>Fairseq</u> 6.7B	47	<u>ProofGPT</u> 6.7B
7	Llama 2 7B HF	29	Pythia 6.7B	48	MPT 7B
8	Go Wizard	30	<u>CodeLlama</u> 7B	49	<u>KoRWKV</u> 6B
9	MPT 7B	31	<u>OpenLlama</u> 7B		

# BENCHMARKING ON SEA LANGUAGES WITH BHASA

	Natural Language Generation			Natural Language Understanding			Natural Language Reasoning	
	Translation (EN > ID)	Translation (ID > EN)	AbsSum	QA	Sentiment	Toxicity	NLI	Causal
<b>1</b>	GPT-4	GPT-4	GPT-4	GPT-4	SEA-LION 7B	GPT-4	GPT-4	GPT-4
<b>2</b>	SEA-LION 7B	Llama 2 7B	SEA-LION 7B	Mistral 7B	GPT-4	SEA-LION 7B	SEA-LION 7B	SEA-LION 7B
<b>3</b>	Llama 2 7B	Mistral 7B	Mistral 7B	SEA-LION 7B	Mistral 7B	Mistral 7B	Mistral 7B	Mistral 7B
<b>4</b>	Mistral 7B	SEA-LION 7B	Llama 2 7B	Falcon 7B	Llama 2 7B	Falcon 7B	Falcon 7B	Falcon 7B
<b>5</b>	Falcon 7B	Falcon 7B	Falcon 7B	Llama 2 7B	Falcon 7B	Llama 2 7B	Llama 2 7B	Llama 2 7B

# What We Have Accomplished



Trained on the SEA-filtered  
CC100 using XLM-R

Trained on the SEA Pile  
using MosaicBERT

# WHERE WILL OUR MODELS BE USED

some of our use cases involve



E-Commerce



Legal



IT



Customer Service

# OUR REGIONAL PARTNERS



# FUTURE WORK



## Medium-Sized Model

Use of more and better data with efficient training and value alignment



## Demand Validation

Discovering demand for sectoral versus specialized models and non-SEA use cases



## Agility

Keeping pace as technology and use cases evolve fast



## Expertise

Transferring of talent as the frontier develops

# ENABLING LLMs WITH DATABRICKS



# Mosaic AI Pre-training and Fine-tuning Stack

Optimised libraries for all stages of training workflow



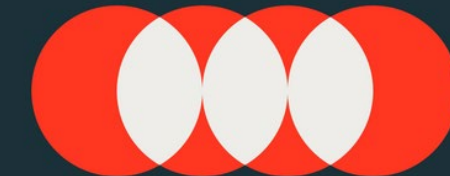
**Composer** is an open source deep-learning training library optimized for scalability and usability.



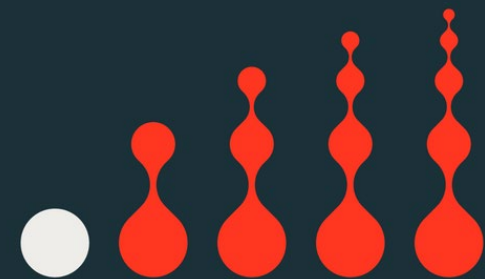
**StreamingDataset** is an open source PyTorch DataLoader that makes it easy and efficient to stream training datasets.



Databricks **LLM Foundry** is a highly efficient, open source codebase for training, fine-tuning and evaluating LLMs.



The **Evaluation Gauntlet** is a library for evaluating the quality of generative language models.



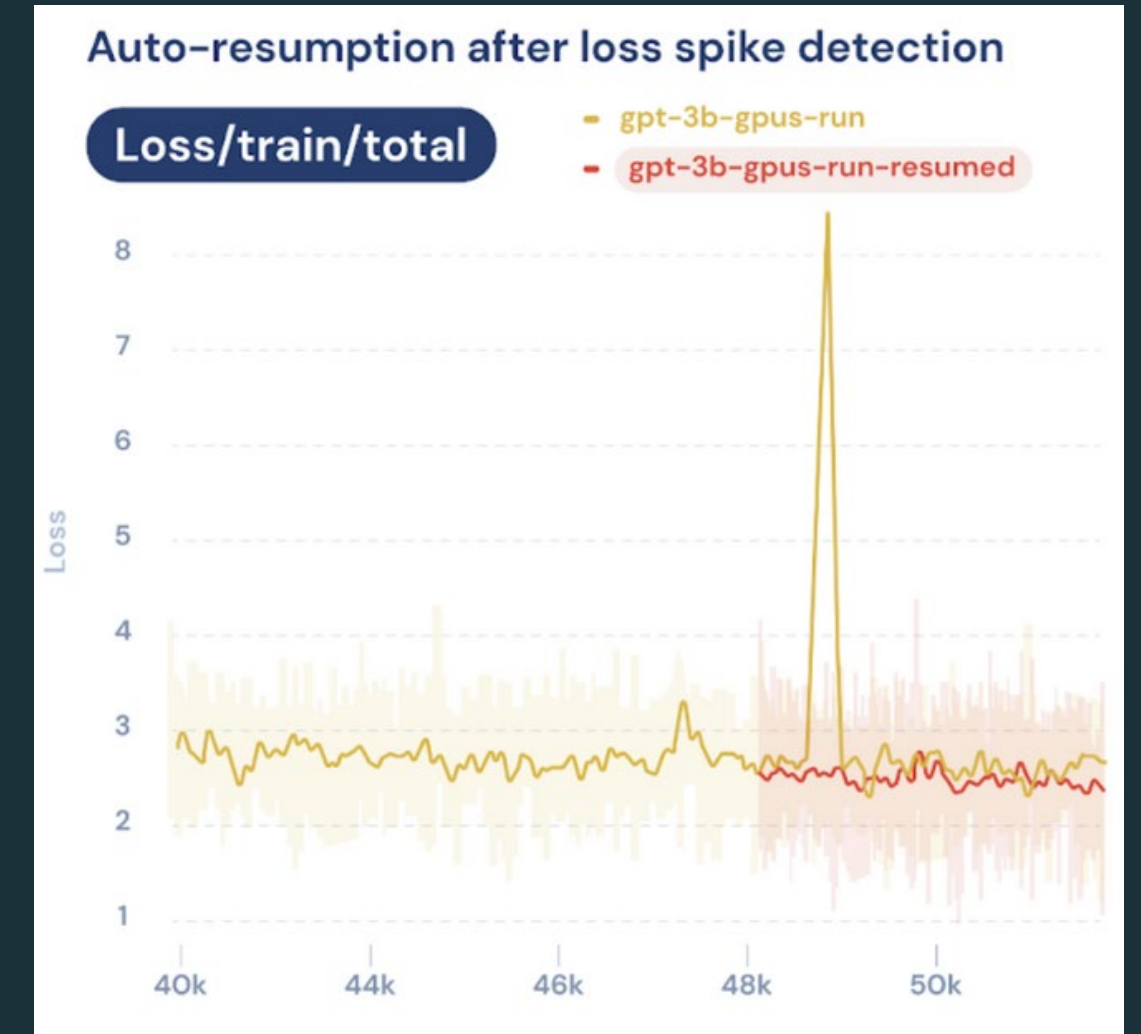
## **Performance**

Our deep learning stack is the most efficient for training, fine-tuning and deploying large models at scale.

# Increase Efficiency + Reduce Costs

A multi-node model training stack that just works

- Easy to use APIs for pre-training and fine-tuning
- Deep learning libraries + compute-optimal systems
- Multi-cloud by default with flexible GPU options
- Seamless scaling of model sizes + GPUs
- Fault tolerance built-in



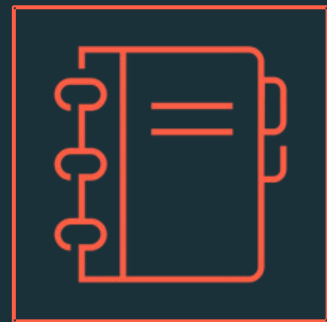
```
model: mosaicml/mpt-30b
train_data_path: s3://my-bucket/my-data
save_folder: s3://my-bucket/checkpoints
eval_data_path: s3://my-bucket/my-data
custom_weights_path: s3://my-bucket/my-custom-weights.pt
training_duration: 10ep
```



# Tools for managing data and model lifecycle



Lilac AI for data exploration and curation



Notebooks and **Apache Spark** for data cleaning and processing



Unity Catalog for data storage and governance



Mosaic Multi-Cloud Training (MCT) to train the model



MLflow and **Lakeview** for experiment tracking



- Foundation Model APIs and AI Playground for eval and red-teaming
- Optimized versions of our open source libraries including Composer, StreamingDataset, and LLM Foundry to train the model

# DATA+AI SUMMIT

TOWARDS  
INCREASING  
SOUTHEAST ASIAN  
REPRESENTATION

