LLMs for Low Resource Languages in Multilingual, Multimodal and Dialectal Settings

https://llm-low-resource-lang.github.io

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

EACL 2024, 21th March, 2024

Speakers

Firoj Alam
Scientist

Shammur Chowdhury
Scientist

Sabri Boughorbel
Scientist

Maram Hasanain
Post Doctoral Researcher

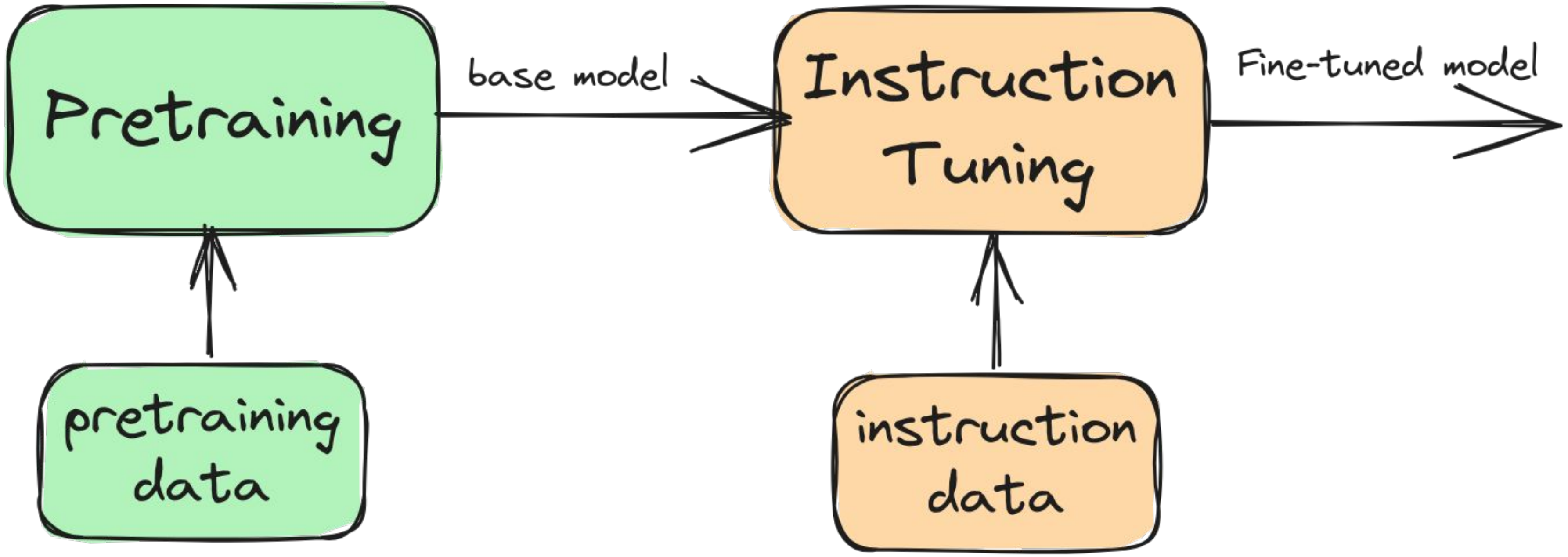
Qatar Computing Research Institute

Content

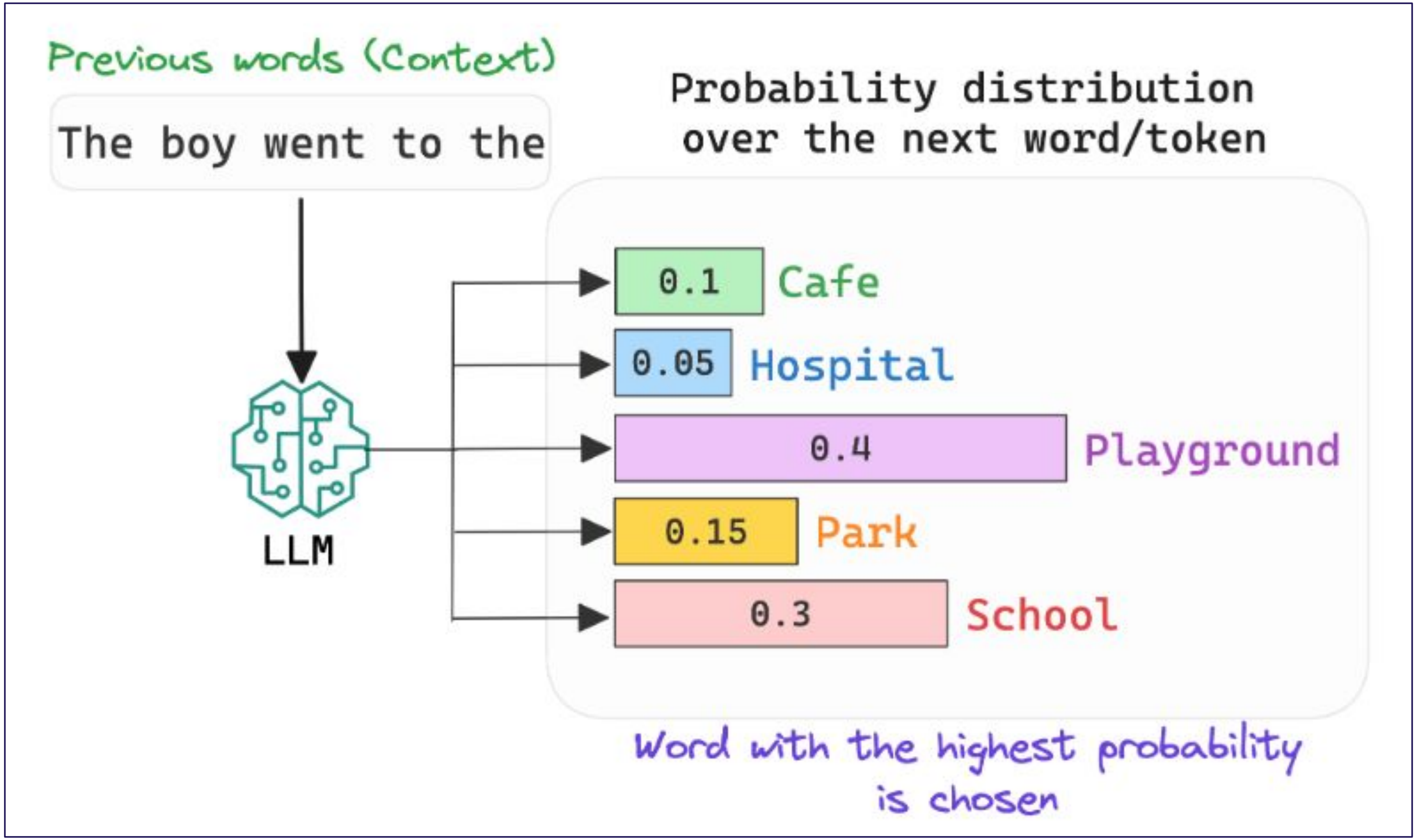
- Introduction **[20 mins]**
- Models and their capabilities for low-resource languages **[70 mins]**
  - NLP models [40 mins]
  - Multimodality [25 mins]
    - Overview
      - Multimodality
      - Speech
  - QA [5 mins]
- Coffee break **[30 mins]**
- Prompting + Benchmarking Tool **[60 mins]**
  - Prompt Engineering [40 mins]
    - Prompting techniques
    - Cross-/multi-lingual prompting
  - Prompt and Benchmarking tools [15 mins]
  - QA: [5 mins]
- Other Related Aspects [**20 mins**]

Models and their Capabilities for Low-Resource Languages

LLMs for Text Input

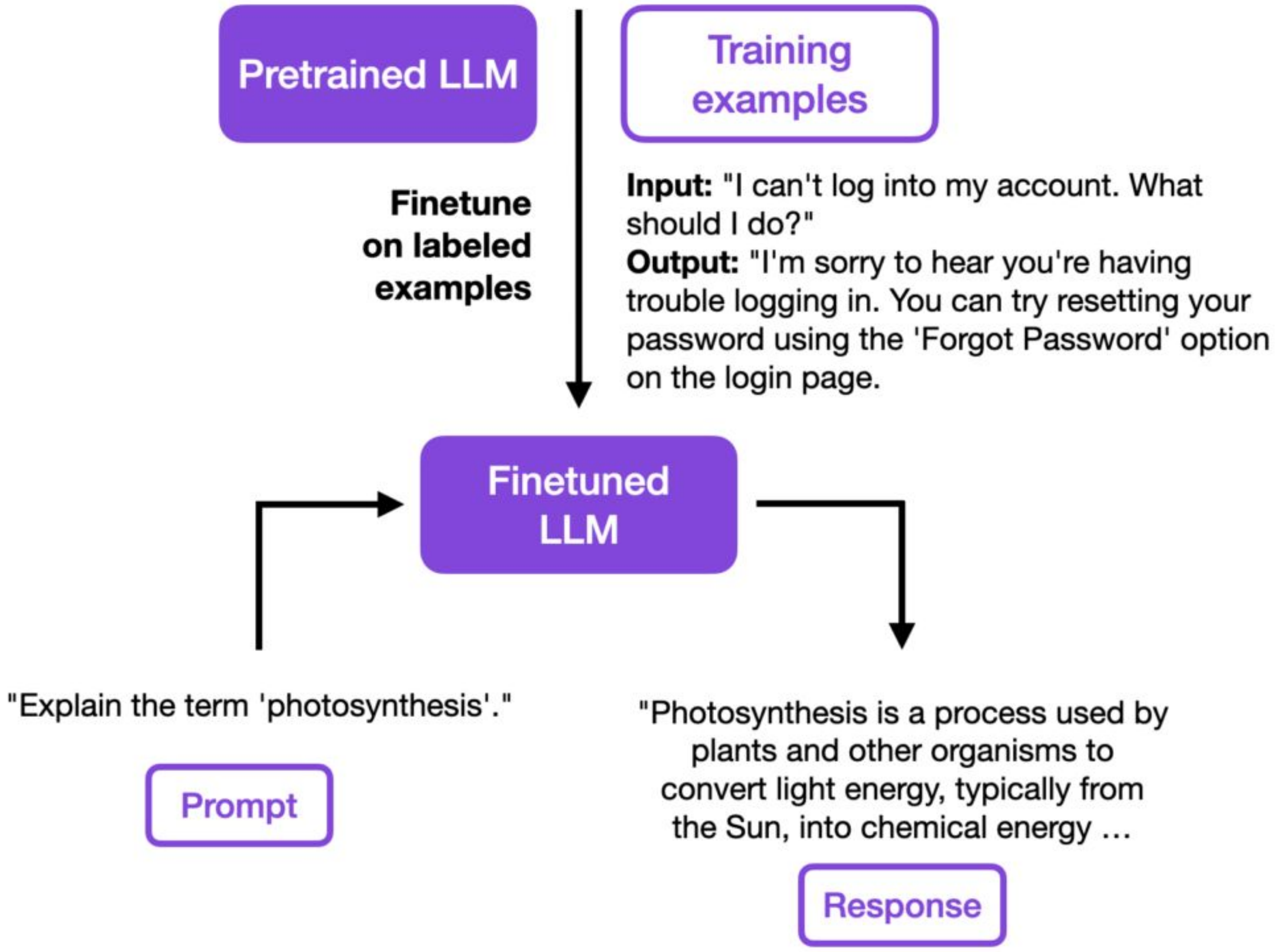


Pretraining



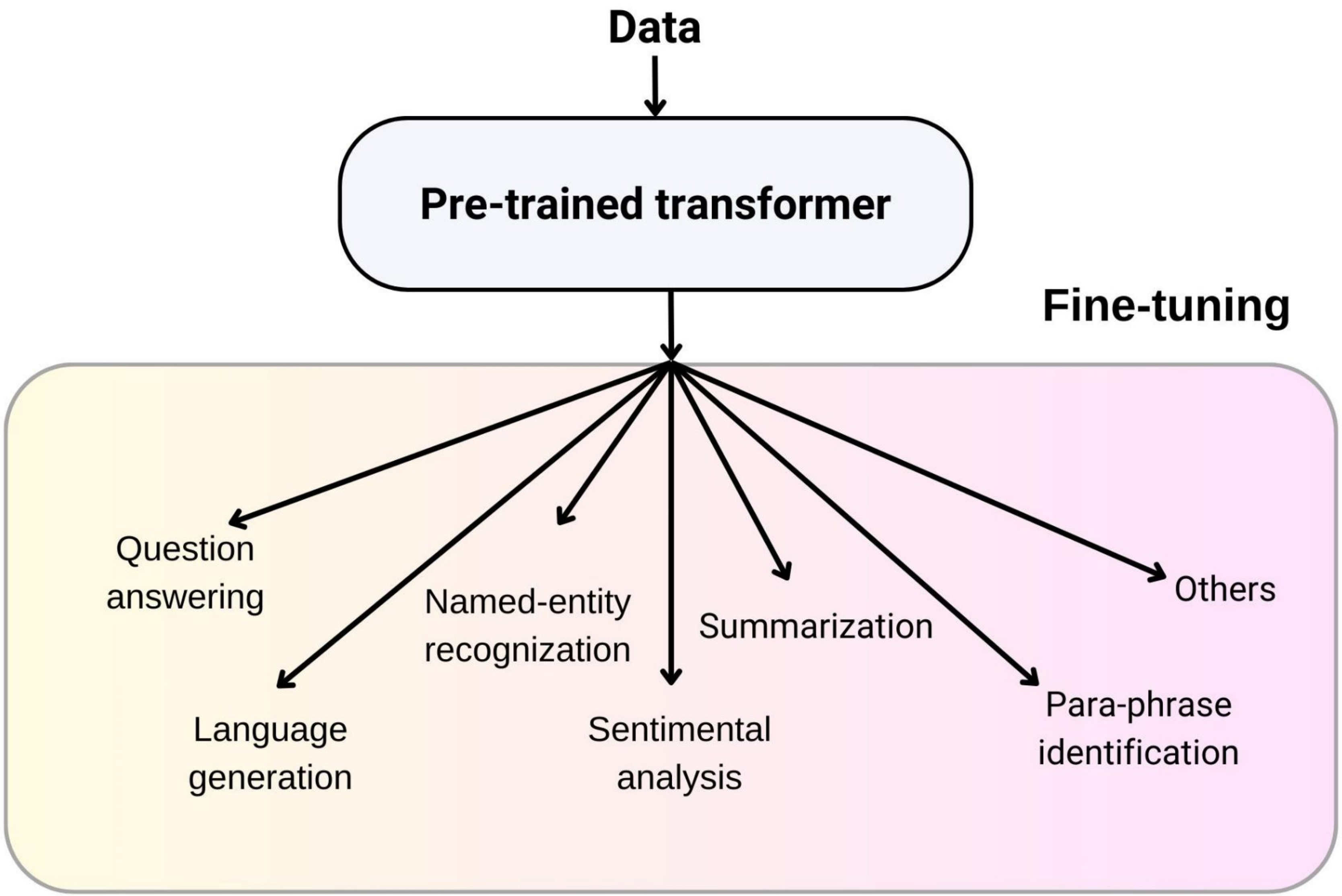
Source: https://undefined.photos/photo-gallery/a-sentence-that-has-the-word-predict-in-it

Instruction Tuning



Source: https://lightning.ai/pages/community/finetuning-falcon-efficiently/

Instruction Tuning



Source: https://datasciencedojo.com/blog/fine-tuning-llms

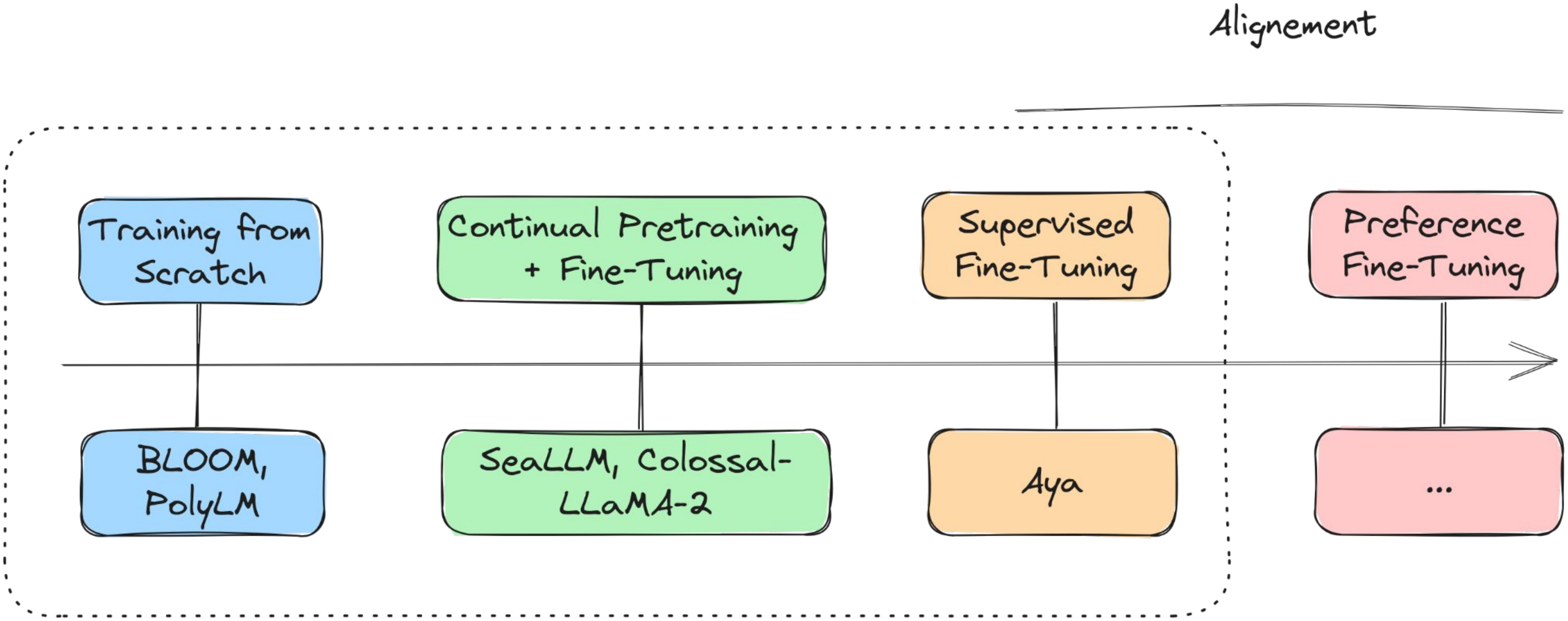
Different Scenarios

Scenarios	Data requirement	Compute requirement
Training from scratch + fine-tuning	++++	++++
Further pretraining + fine-tuning	+++	++
Fine-tuning existing LLM	+	+

Multilingual LLMs

https://medium.com/grabngoinfo/how-to-access-llama-2-free-generative-ai-llm-alternative-to-chatgpt-api-359569b27c3a

LLM Training Pipeline



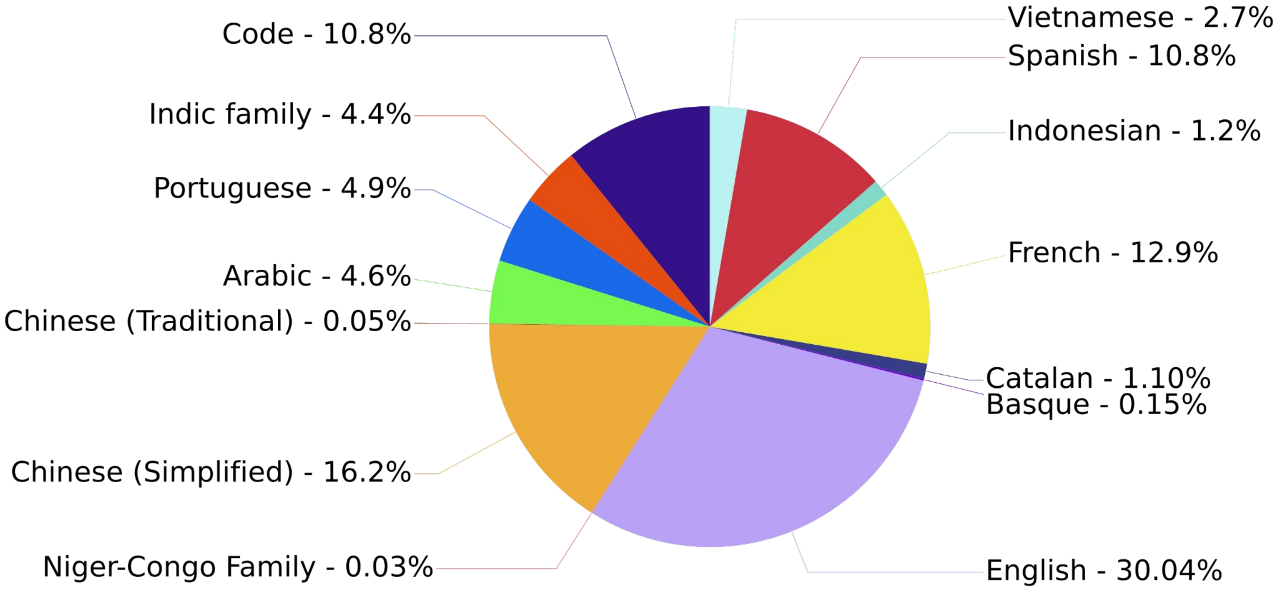
BLOOM

a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

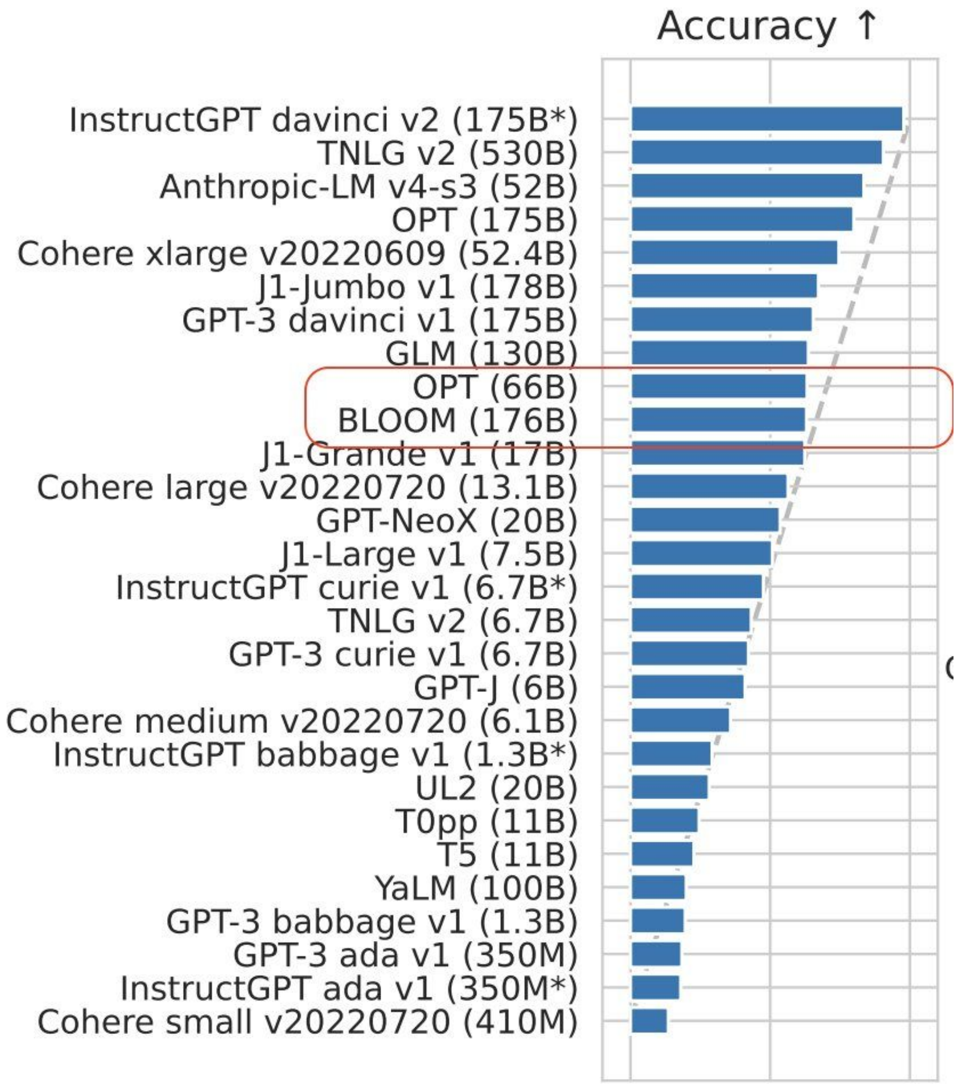
- From BigScience consortium
- Model family: 560m, 1.7B, 3B, 7B, 176B
- Instruction-tuned: BLOOMZ using xP3
- Training data (ROOTS corpus)
  - 498 Hugging Face datasets
  - 46 languages
  - 13 programming languages
  - 350B tokens
  - 250K vocabulary size tokenizer



https://huggingface.co/bigscience/bloom-7b1

BLOOM

- BLOOM-176B performance in English is not at expectation but smaller version could
- It can be useful for low resource language
  - 60% of its data in non-English
  - example of fine-tuned bloom-7b: *phoenix-chat-7b*



BLOOM

bloom

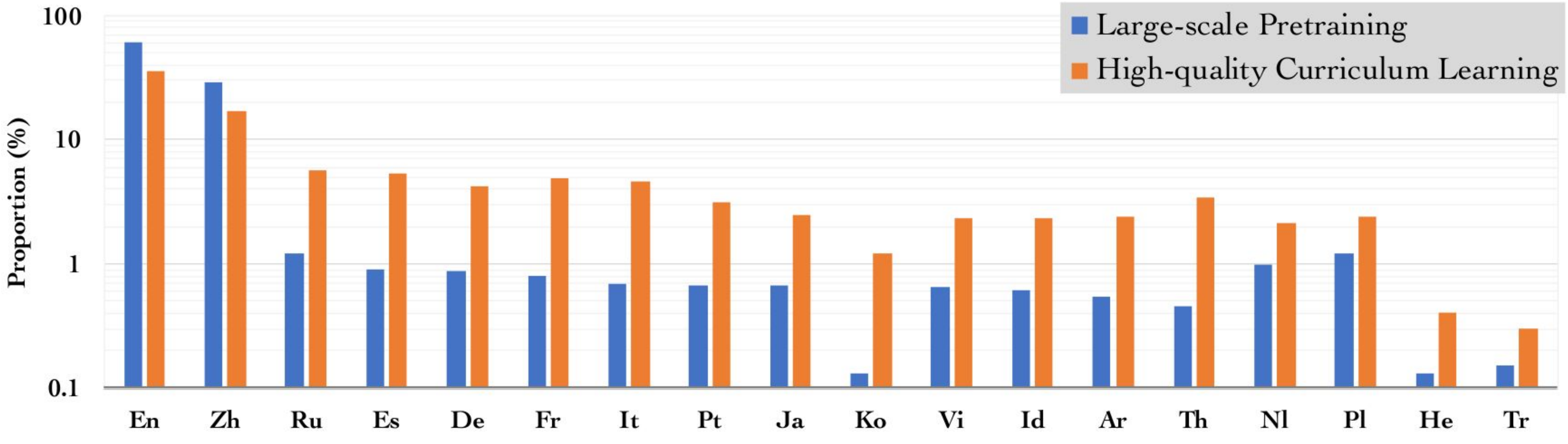
Model	Language	Code	Average	ARC (25-shot)	HellaSwag (0-shot)	MMLU (25-shot)
bloom-7b1	French	fr	41	36.7	56.6	29.9
bloom-7b1	Spanish	es	41	38.1	56.7	28.9
bloom-7b1	Portuguese	pt	40.7	40	55.1	28.8
bloom-7b1	Chinese	zh	39.1	37.3	51.2	29.1
bloom-7b1	Catalan	ca	38.7	34.7	51.2	28.8
bloom-7b1	Vietnamese	vi	38.7	33.7	48.3	28.1
bloom-7b1	Indonesian	id	38.5	36	49.5	28.1
bloom-7b1	Arabic	ar	36.2	31.4	43.3	27.5
bloom-7b1	Italian	it	35.3	29	40.8	27.6
bloom-7b1	Hindi	hi	34.4	29.2	36.4	27.5

https://huggingface.co/spaces/uonlp/open_multilingual_llm_leaderboard

PolyLM

- Trained on 638B tokens in two sizes 1.7B and 13B
- Tokenizer: vocabulary size is 256K
  - Reduced bias towards high resource language by increasing vocab size of LRL

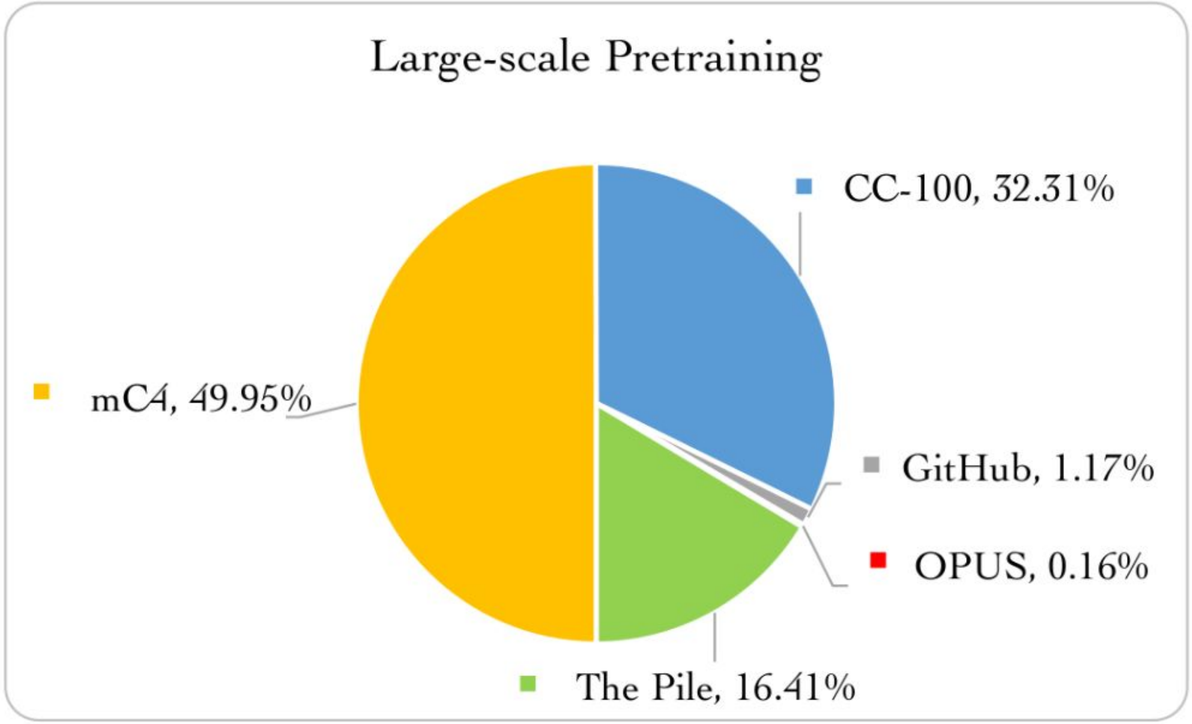
Source	Fraction	Tokens	Type
mC4	49.95%	321.7B	Web-text (Multilingual)
CC-100	32.31%	208.1B	Web-text (Multilingual)
The Pile	16.41%	105.7B	Web-text & books (English)
GitHub	1.17%	7.5B	Code
OPUS	0.16%	1.0B	Parallel Multilingual Data
Sum	-	638B	

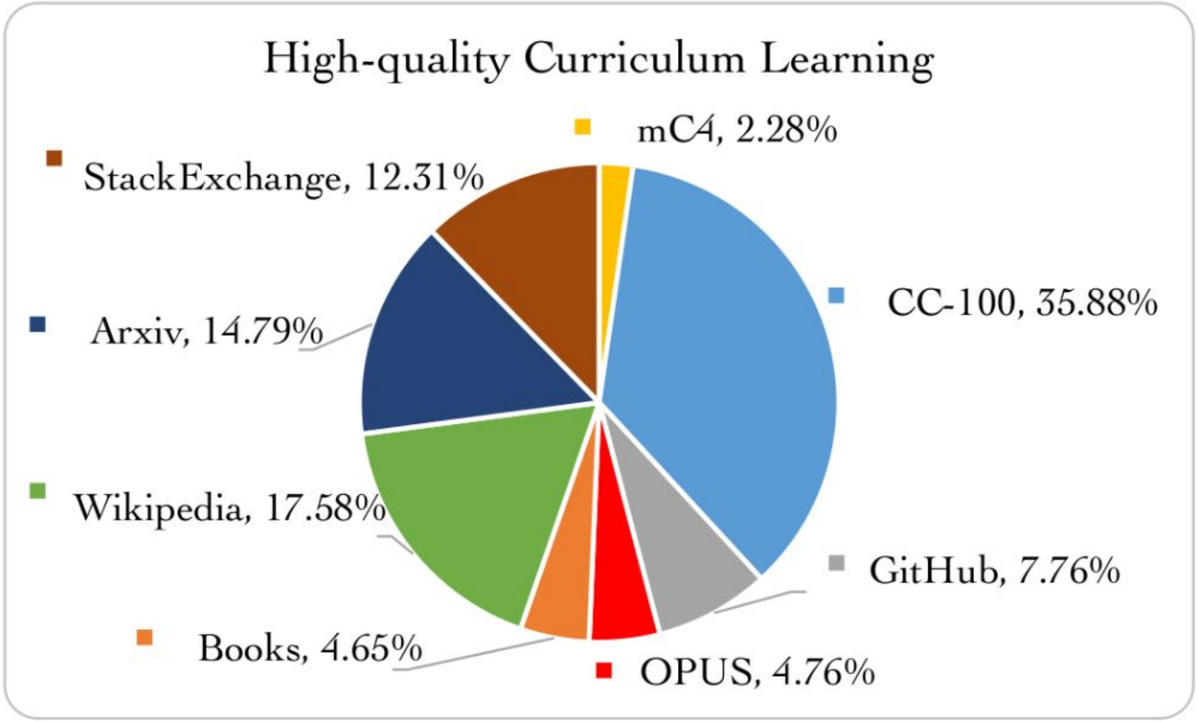


Polylm: An open source polyglot large language model (Wei, et al., 2024)

PolyLM

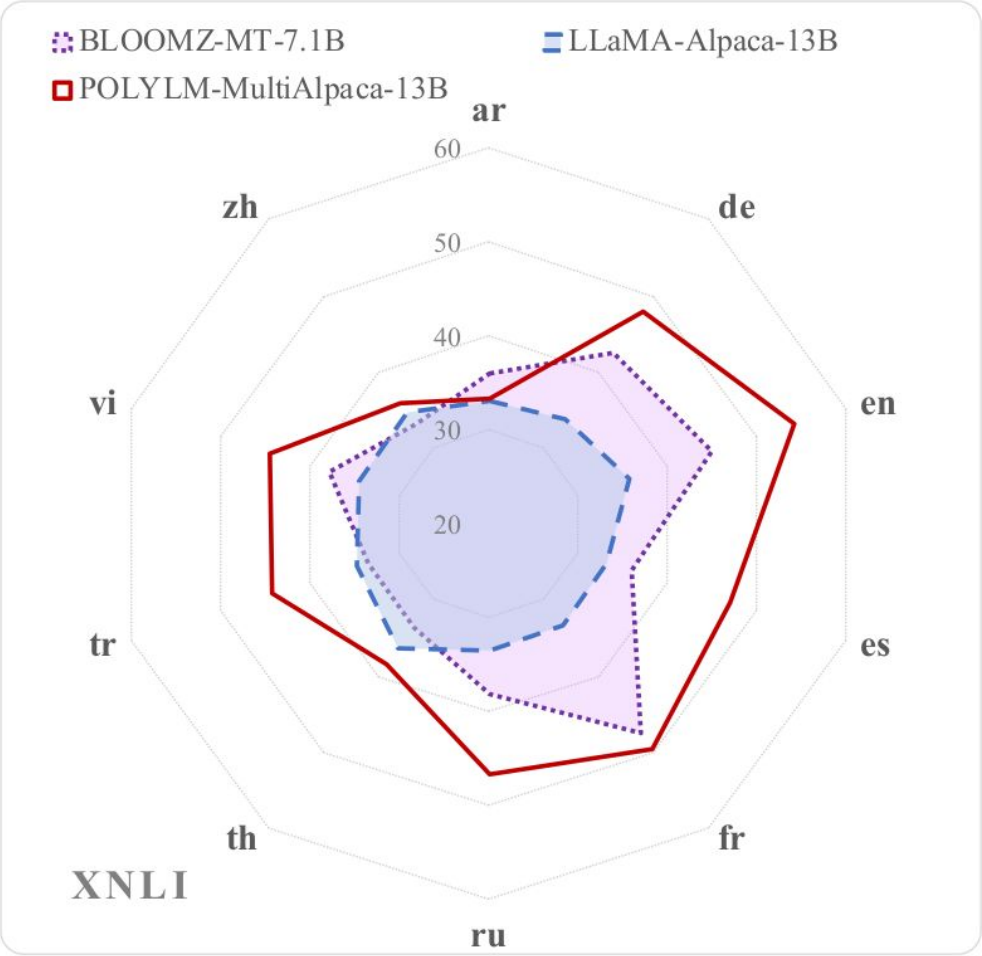
- Curriculum Learning:
  - Increased non-English data 30% to 60%
- Bilingual data into training data;

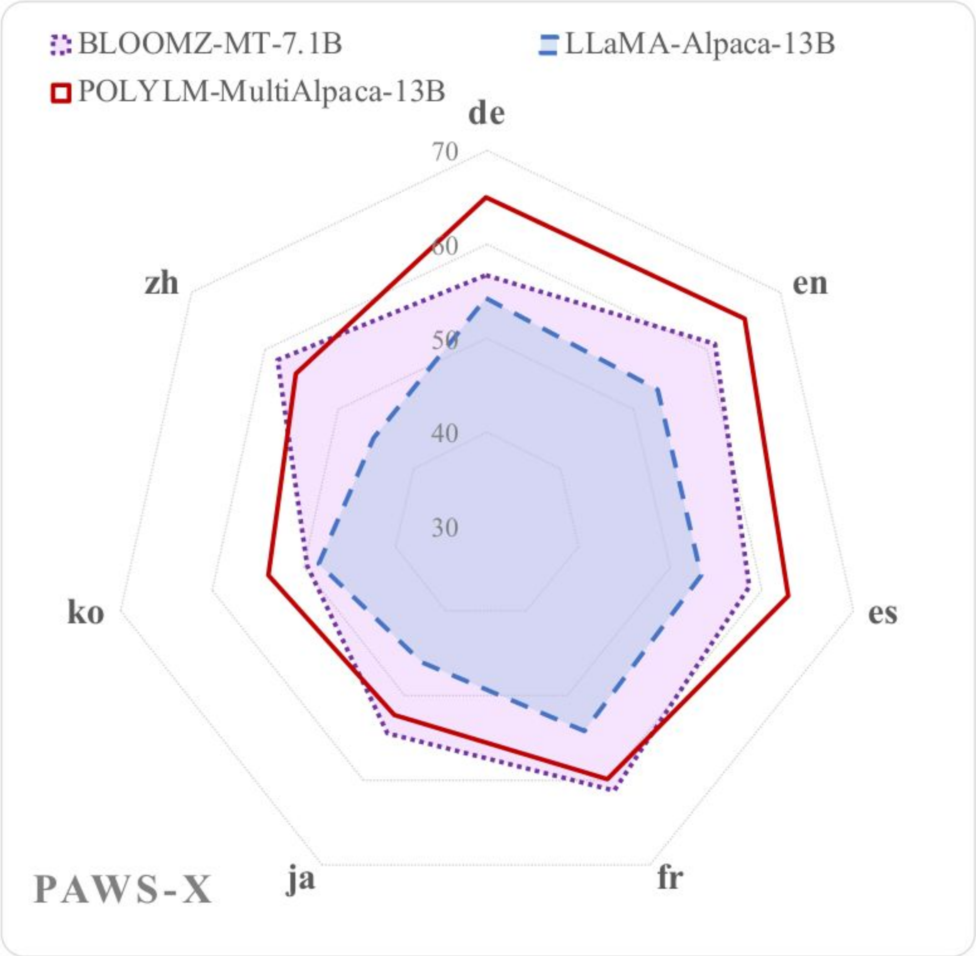




Polylm: An open source polyglot large language model (Wei, et al., 2024)

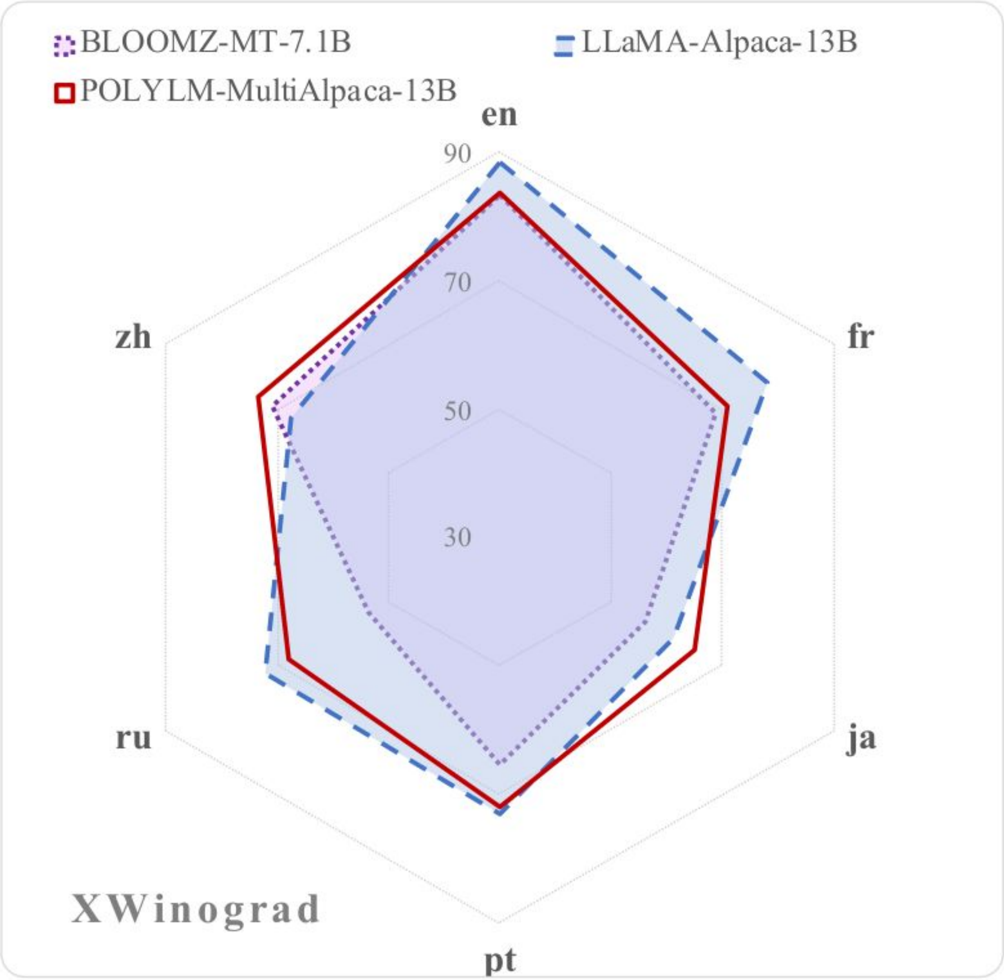
PolyLM

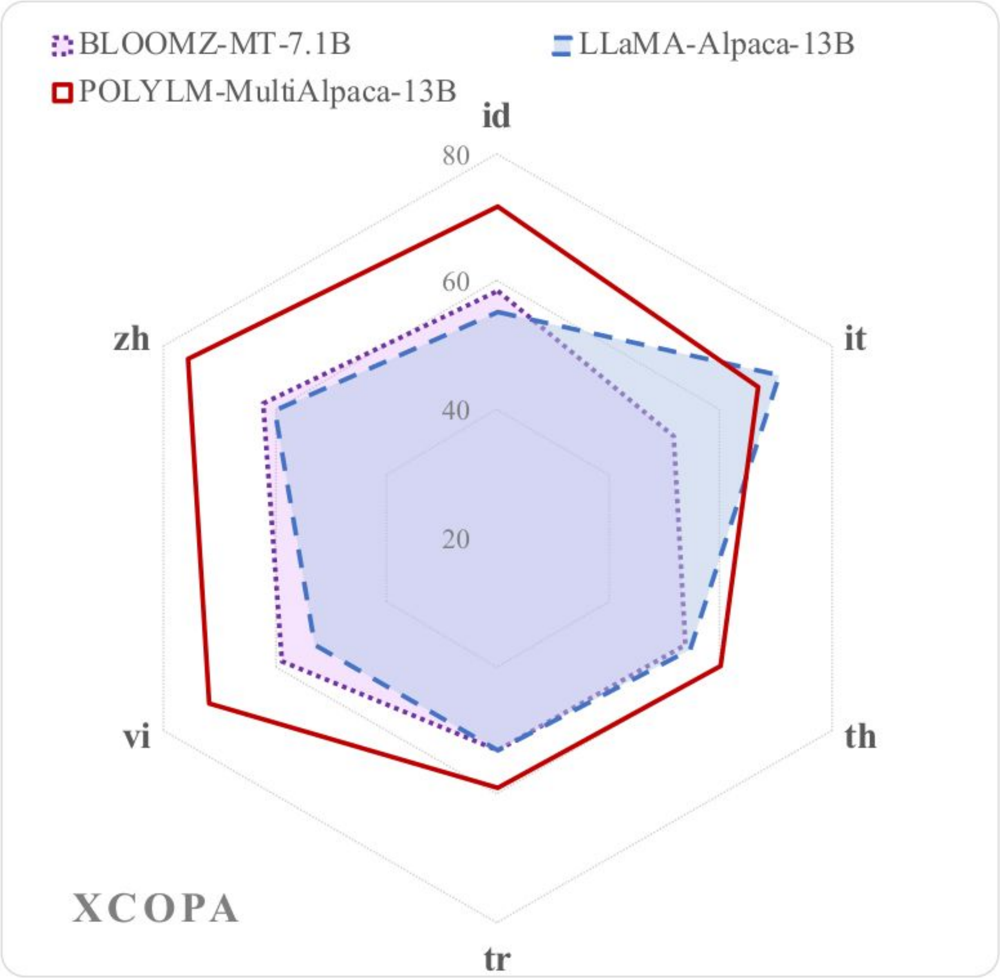




Polylm: An open source polyglot large language model (Wei, et al., 2024)

PolyLM



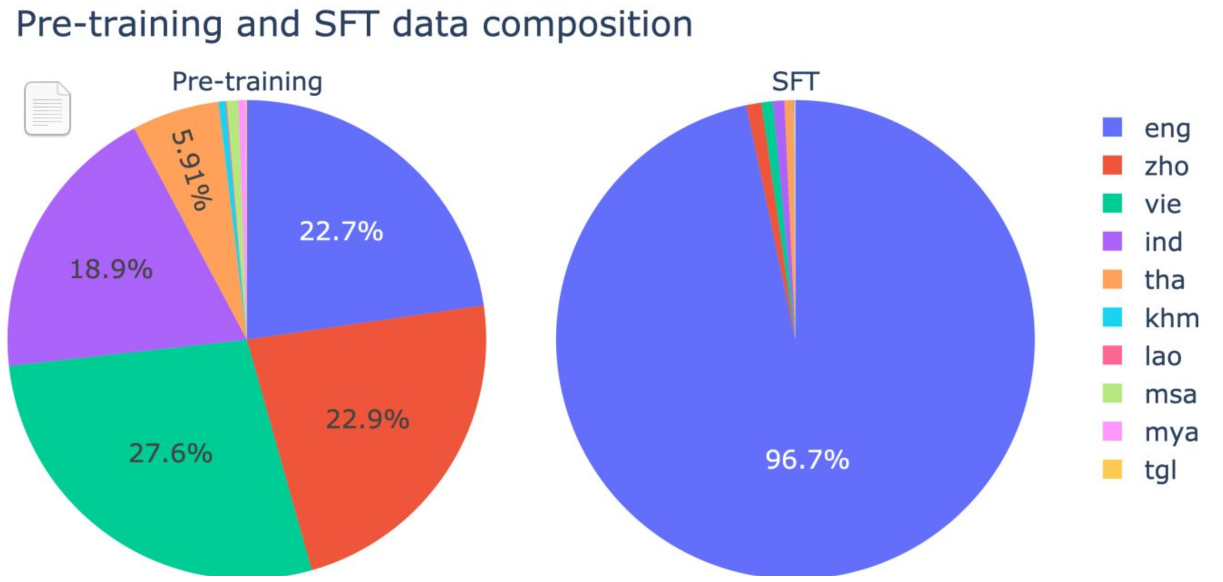


Polylm: An open source polyglot large language model (Wei, et al., 2024)

SeaLLM

https://github.com/DAMO-NLP-SG/SeaLLMs

- SeaLLMs - Large Language Models for Southeast Asia:
  - Thai, Vietnamese, Indonesian, Chinese, Khmer, Lao, Malay, Burmese, and Tagalog
- Base model: Llama-2-13B
- Extended vocabulary: 16K



SeaLLMs--Large Language Models for Southeast Asia. (Nguyen, Xuan-Phi, et al. 2023).

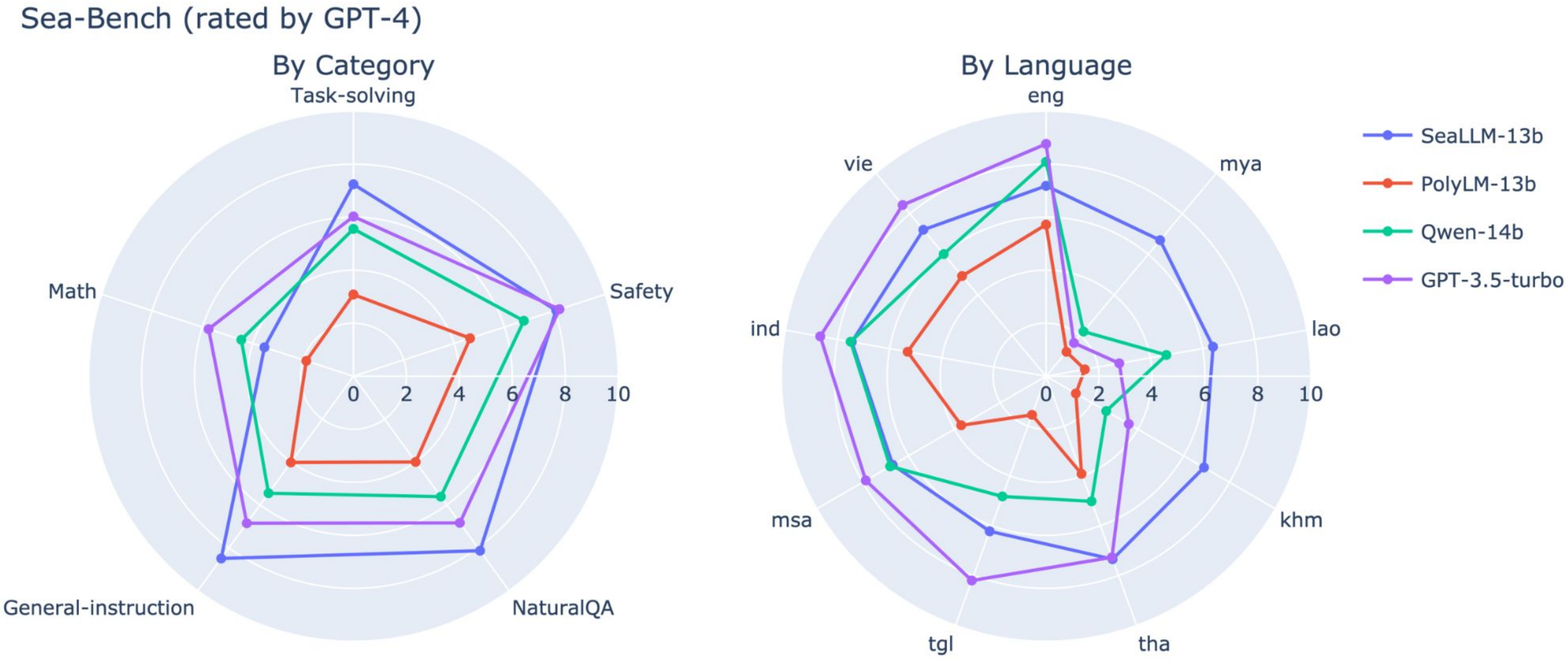
SeaLLM

- **Vocabulary expansion**
  - Exhaustive Merge
  - Pruning low frequency
- **Pretraining**
  - Different languages into a single training sequence
  - high-quality documents for each language -> lower quality -> high-quality
- **Pre-training and SFT Hybrid**
  - pre-training corpus, labeled data from traditional NLP tasks, and significant quantities of open-source instruction-following data
- **SFT**
  - native-language data, selective translation, self-instruction

SeaLLM

Model	M3Exam					MMLU
	Eng	Zho	Vie	Ind	Tha	Eng
ChatGPT-3.5	75.46	60.20	58.64	49.27	37.41	70.00
Llama-2-7b	49.58	37.58	29.82	28.93	19.89	45.62
Llama-2-13b	61.17	43.29	39.97	35.50	23.74	53.50
Polylm-13b	32.23	29.26	29.01	25.36	18.08	22.94
SeaLLM-7b	54.89	39.30	38.74	32.95	25.09	47.16
SeaLLM-13b-5L	**63.20**	**45.13**	**49.13**	**40.04**	**36.85**	**55.23**
SeaLLM-13b-10L	62.69	44.50	46.45	39.28	36.39	52.68

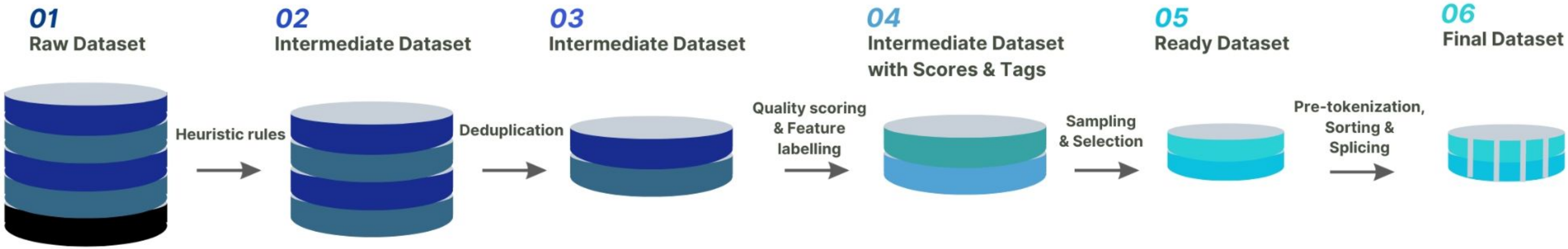
SeaLLM



Sea-Bench (rated by GPT-4)
By Category
Task-solving
Safety
NaturalQA
General-instruction
Math
0
2
4
6
8
10
By Language
eng
mya
lao
khm
tha
tgl
msa
ind
vie
0
2
4
6
8
10
SeaLLM-13b
PolyLM-13b
Qwen-14b
GPT-3.5-turbo

Colossal-LLaMA-2-7B

- Continual pre-training of 8.5 billion tokens over a duration of 15 hours with 64 A800 GPUs (<$1,000)
- Vocabulary size: 32,000 to 69,104
- High quality data



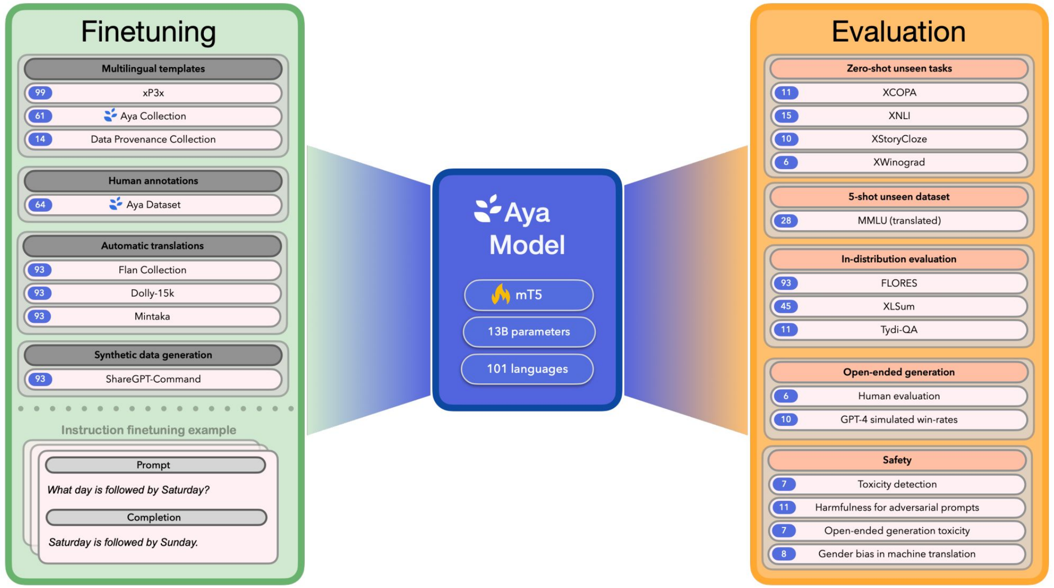
Colossal-LLaMA-2-7B

https://github.com/hpcaitech/ColossalAI

Model	Backbone	Tokens Consumed	MMLU (5-shot)	CMMLU (5-shot)	AGIEval (5-shot)	GAOKAO (0-shot)	CEval (5-shot)
Baichuan-7B	-	1.2T	42.32	44.53	38.72	36.74	42.8
ChatGLM2-6B	-	1.4T	44.74	49.40 (-)	46.36	45.49	51.7
Qwen-7B	-	2.2T	54.29	56.03	52.47	56.42	59.6
Llama-2-7B	**-**	**2.0T**	**44.47**	**32.97 (-)**	**32.6**	**25.46**	**-**
Linly-AI/Chinese-LLaMA-2-7B-hf	Llama-2-7B	1.0T	37.43	29.92	32	27.57	-
FlagAlpha/Atom-7B	Llama-2-7B	0.1T	49.96	41.1	39.83	33	-
IDEA-CCNL/Ziya-LLaMA-13B-v1.1	Llama-13B	0.11T	50.25	40.99	40.04	30.54	-
Colossal-LLaMA-2-7b-base	Llama-2-7B	**0.0085T**	53.06	49.89	51.48	58.82	50.2
Colossal-LLaMA-2-13b-base	Llama-2-13B	**0.025T**	56.42	61.8	54.69	69.53	60.3

Aya

- Instruction-tuned mT5 (13B)
- 101 languages of which over 50% are considered as lower-resourced
- 250k vocabulary size
- Evaluation suites for 99 languages
- Instruction datasets are open sourced



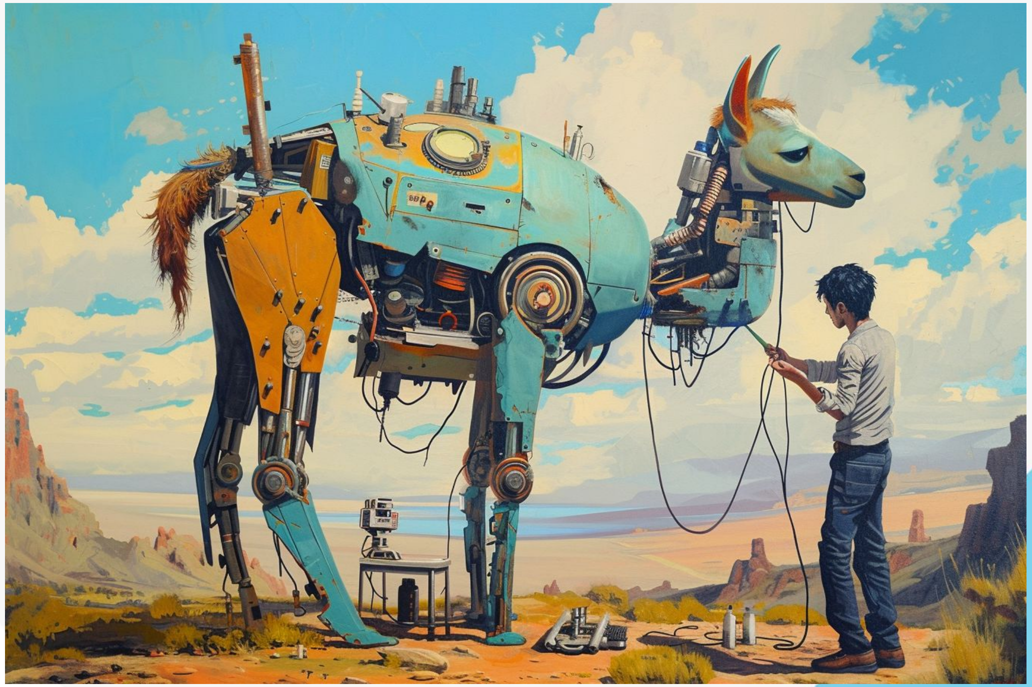
Group	Category	Languages	Examples
Higher-Resourced	5	7	Arabic, Chinese, English, French, Spanish
	4	17	Hindi, Italian, Portuguese, Russian, Turkish
Mid-Resourced	3	24	Afrikaans, Indonesian, Kazakh, Latin, Latvian
Lower-Resourced	2	11	Hausa, Icelandic, Irish, Lao, Maltese
	1	29	Albanian, Gujarati, Igbo, Luxembourgish
	0	13	Kurdish, Kyrgyz, Nyanja, Sinhala, Yiddish

Aya

Model	Base Model	IFT Mixture	Held out tasks (Accuracy %) XCOPA	XNLI	XSC	XWG	Avg
46 Languages							
mT0	mT5 13B	xP3	75.6	55.3	87.2	73.6	72.9
BLOOMZ	BLOOM 176B	xP3	64.3	52.0	82.6	63.3	65.5
52 Languages							
Bactrian-X 13B	Llama 13B	Bactrian-X	52.4	34.5	51.8	50.5	47.3
101 Languages							
mT0x	mT5 13B	xP3x	71.7	45.9	85.1	60.6	65.8
Aya (human-anno-heavy)	mT5 13B	All Mixture	76.5	**59.2**	89.3	70.6	73.9
Aya (template-heavy)	mT5 13B	All Mixture	**77.3**	58.3	**91.2**	**73.7**	**75.1**
★ **Aya** (translation-heavy)	mT5 13B	All Mixture	76.7	58.3	90.0	70.7	73.9

Aya

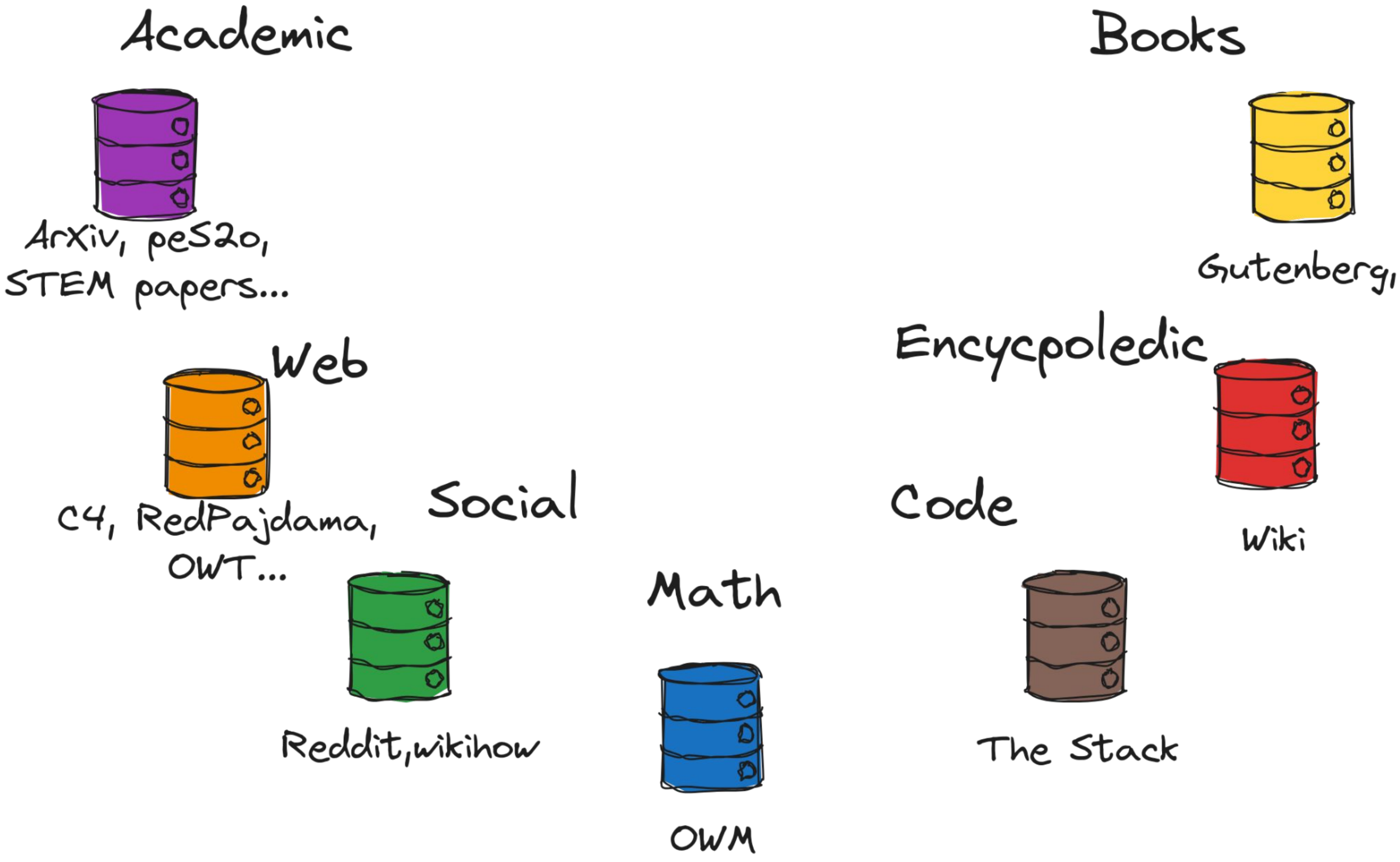
	arb	cat	deu	eus	fra	hin	hrv	hun	ita	nld	por	rud	ser	spa	swe	vie
Okapi‡	27.7	30.5	31.7	27.9	30.7	26.5	30.0	30.1	30.4	31.1	30.1	30.6	30.4	30.9	29.3	27.5
mT0	31.5	32.8	32.7	29.7	32.1	32.0	31.1	32.3	32.4	32.0	32.1	32.8	30.9	32.1	31.6	30.9
mT0x	31.6	32.6	32.5	29.2	32.7	31.6	31.1	31.7	31.3	32.1	32.0	31.7	31.4	32.2	32.8	31.1
Aya	38.2	39.6	39.7	36.0	39.7	38.7	37.5	38.8	39.0	40.1	39.0	39.2	38.1	39.7	39.7	34.8
	zho	ben	dan	ind	ron	slk	tam	ukr	guj	hye	kan	mal	mar	npi	tel	**Avg**
Okapi‡	28.2	26.8	31.8	27.5	30.9	30.2	26.0	31.6	27.4	27.5	26.8	25.8	26.1	25.2	25.9	28.8
mT0	32.5	31.6	33.0	33.3	32.4	32.3	29.4	31.5	29.5	28.4	30.9	28.6	31.6	32.4	29.0	31.5
mT0x	31.6	30.2	32.0	32.3	31.8	31.4	27.7	32.3	28.5	26.7	28.9	26.7	29.7	30.1	27.9	30.8
Aya	38.3	35.8	39.7	40.0	39.5	39.4	31.2	39.9	33.6	30.0	34.5	30.4	36.0	37.2	32.1	**37.3**



https://www.datacamp.com/tutorial/fine-tuning-llama-2

Pre-training Data

Multi-Source Corpora



Pretraining Datasets

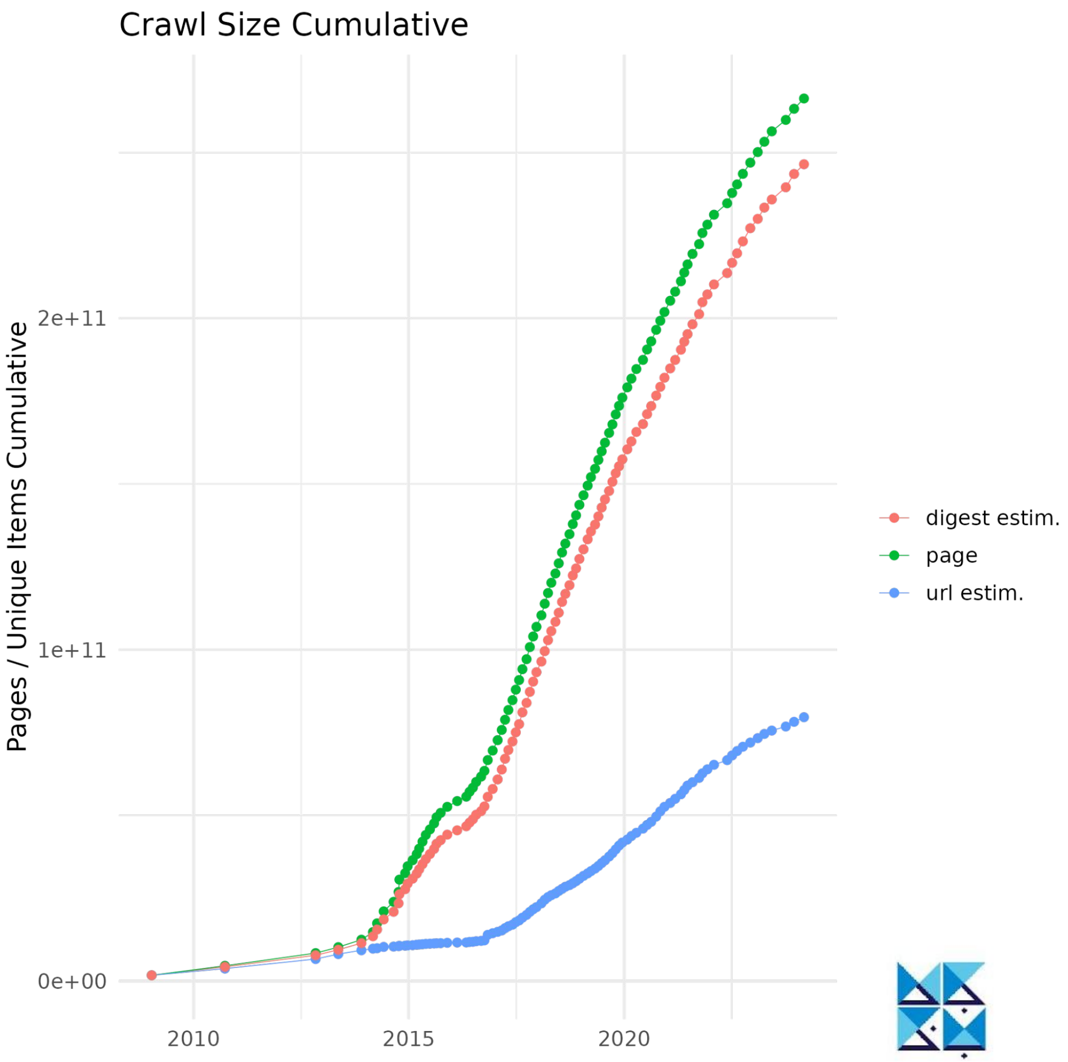
- **Multilingual datasets**
  - Common Crawl, mC4, OSCAR, CulturaX
- **Creating own dataset using data preparation pipelines**
  - RedPajama
  - Dolma
- **Machine translation for data augmentation**

Common Crawl

COMMON CRAWL

https://commoncrawl.org/

- Open repository of web crawl data
- Petabytes of data, regularly collected since 2008
  - 250 billion pages over 17 years
  - 3-5 billion new pages added each month
  - In June 2023, 3 billion web pages and ~400 TB of uncompressed data.



https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

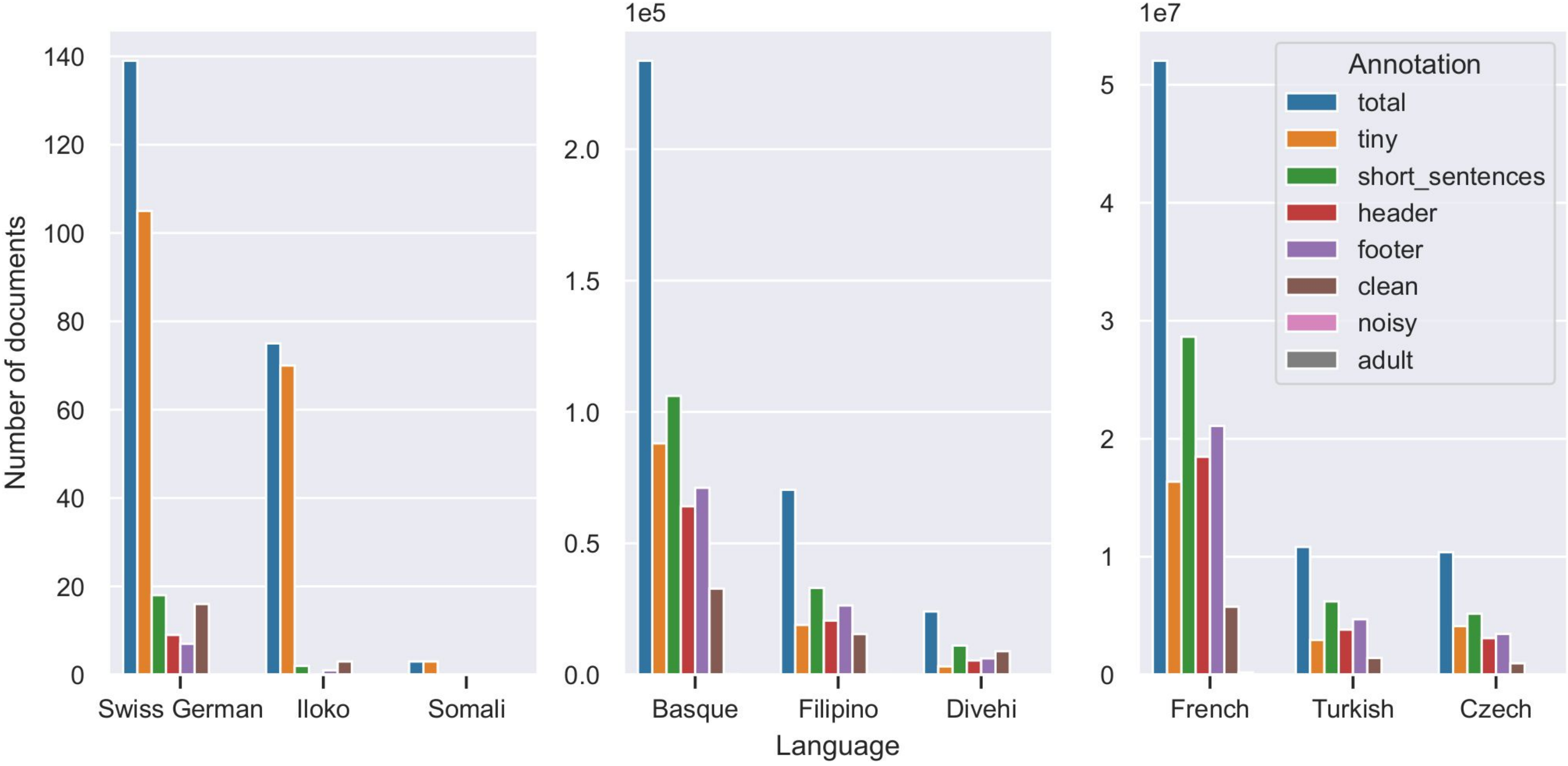
OSCAR

https://oscar-project.org/

- **O**pen **S**uper-large **C**rawled **A**ggregated co**R**pus
- 151 different languages (12GB multilingual corpus)
- It has been used to train known models, e.g., BART
- Moved from line-oriented to documented-oriented
- Added Annotations:
  - Length-based
  - Noise detection (ratio letters/non-letters, unicode categories)
  - Adult content

OSCAR

https://oscar-project.org/



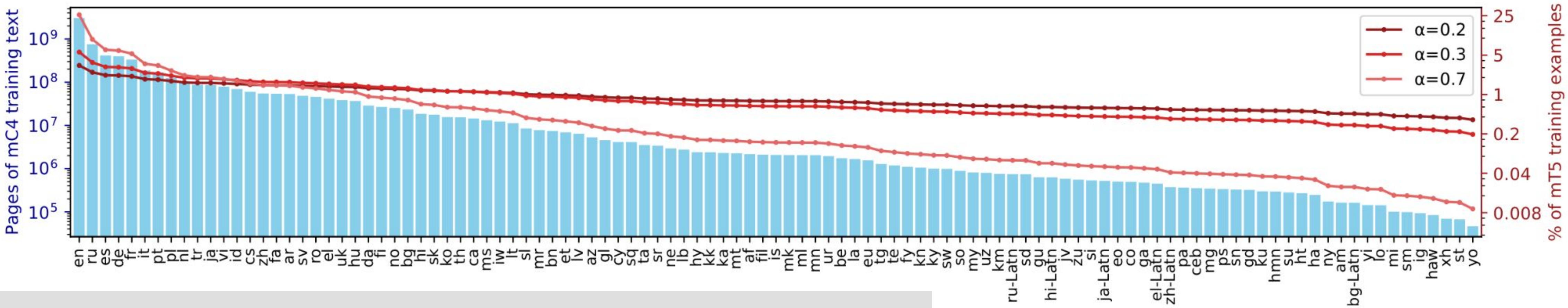
Annotation count

Towards a cleaner document-oriented multilingual crawled corpus (Julien, et al., 2022)

mC4: Multilingual C4

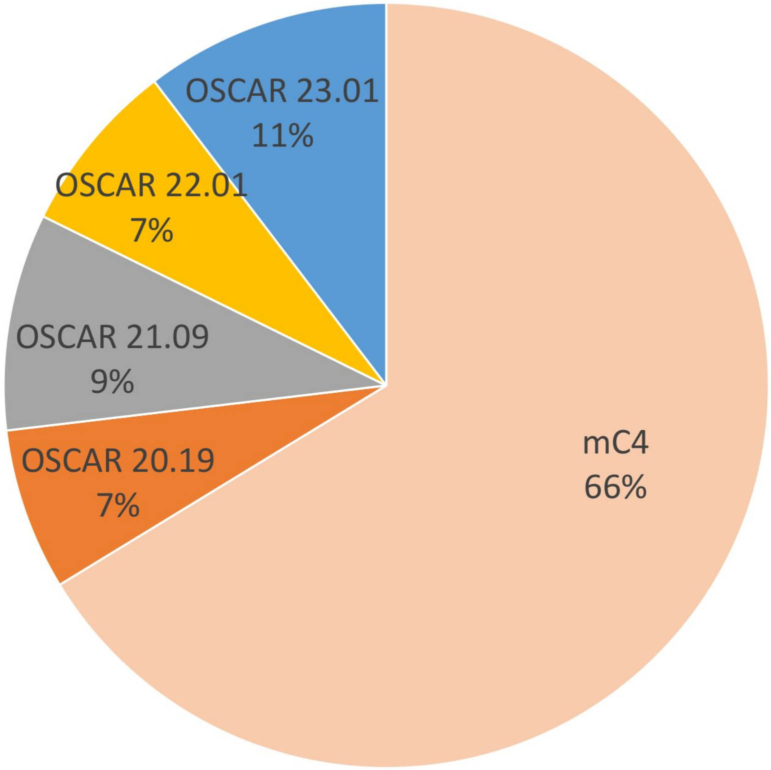
https://huggingface.co/datasets/mc4

- **M**ultilingual **C**olossal, **C**leaned version of **C**ommon **C**rawl's web crawl corpus
- mC4 has been used to train Google’s mT5 model
- 2.7T tokens English, 3.6T tokens multilingual
- Language identification using CLD3



mT5: A massively multilingual pre-trained text-to-text transformer (Xue, et al., 2020)

CulturaX



- Combines: mC4 and OSCAR
  - 6.3B tokens
  - 167 languages
- Extensive cleaning and deduplication
  - Language Identification: FastText identification on mC4
  - URL-based Filtering
  - Metric-based Cleaning:
    - MinHash &URL-based Deduplication

Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. (Nguyen et al., 2023)

RedPajama

	RedPajama	LLaMA*
CommonCrawl	878 billion	852 billion
C4	175 billion	190 billion
Github	59 billion	100 billion
Books	26 billion	25 billion
ArXiv	28 billion	33 billion
Wikipedia	24 billion	25 billion
StackExchange	20 billion	27 billion
Total	1.2 trillion	1.25 trillion

- Open source dataset with two versions
- English-centric dataset
- Llama dataset clone
  - same performance over 20 benchmarking datasets

Task/Metric	GPT-J 6B	LLaMA 7B	LLaMA 13B	OpenLLaMA 3Bv2	OpenLLaMA 7Bv2	OpenLLaMA 3B	OpenLLaMA 7B	OpenLLaMA 13B
Average	0.52	0.55	0.57	0.53	0.56	0.53	0.55	0.57

https://github.com/togethercomputer/RedPajama-Data

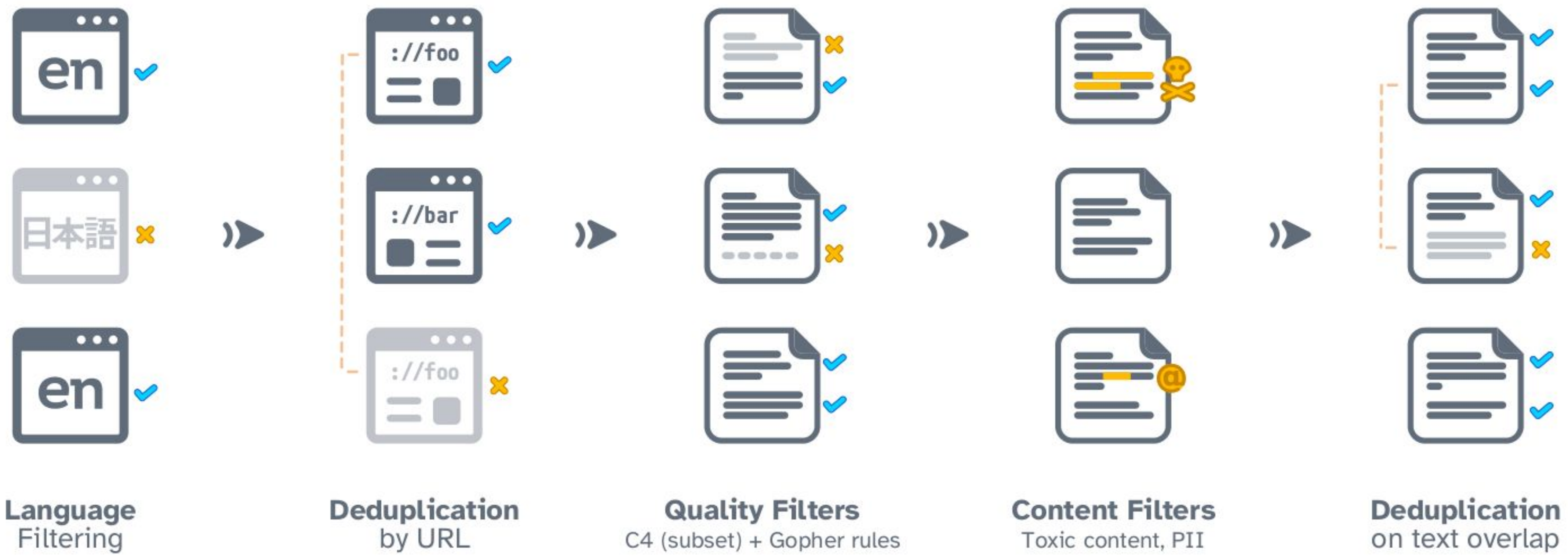
RedPajama V2

- 84 CommonCrawl snapshots
- Processed using the CCNet pipeline
- Quality Signals (>40 quality signals)
- Deduplication
- Open source pipeline
- **Interesting direction:**
  - multilingual RedPajama

	# Documents	Estimated Token count (deduped)
en	14.5B	20.5T
de	1.9B	3.0T
fr	1.6B	2.7T
es	1.8B	2.8T
it	0.9B	1.5T
Total	20.8B	30.4T

https://github.com/togethercomputer/RedPajama-Data

Dolma

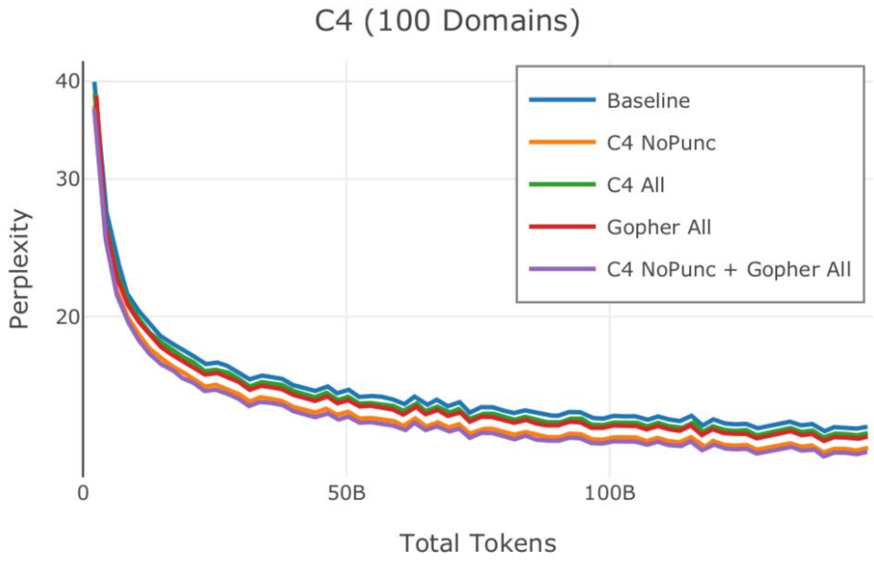

en
日本語
en
://foo
://bar
://foo
Language Filtering
Deduplication by URL
Quality Filters
C4 (subset) + Gopher rules
Content Filters
Toxic content, PII
Deduplication on text overlap

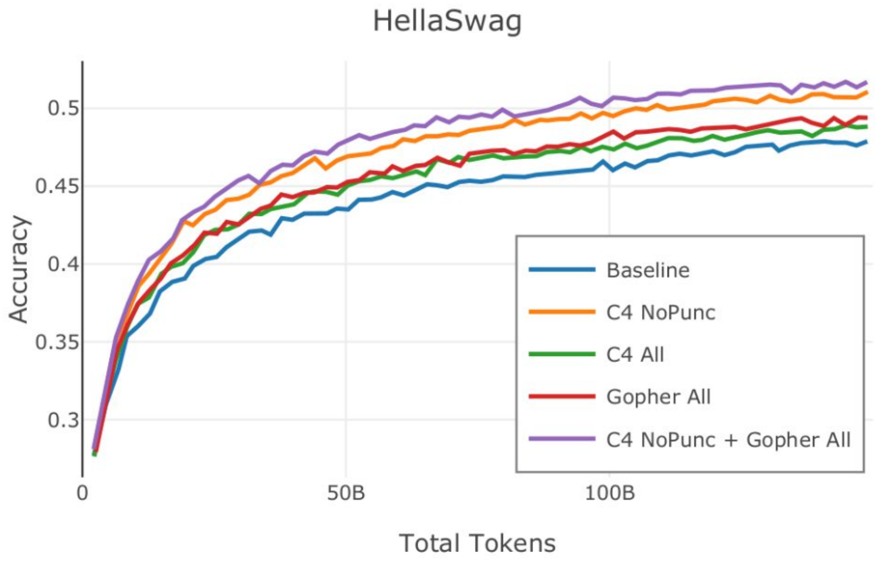
Dolma

Source	Doc Type	UTF-8 bytes (GB)	Documents (millions)	Unicode words (billions)	Llama tokens (billions)
Common Crawl	web pages	9,022	3,370	1,775	2,281
The Stack	code	1,043	210	260	411
C4	web pages	790	364	153	198
Reddit	social media	339	377	72	89
PeS2o	STEM papers	268	38.8	50	70
Project Gutenberg	books	20.4	0.056	4.0	6.0
Wikipedia, Wikibooks	encyclopedic	16.2	6.2	3.7	4.3
Total		**11,519**	**4,367**	**2,318**	**3,059**

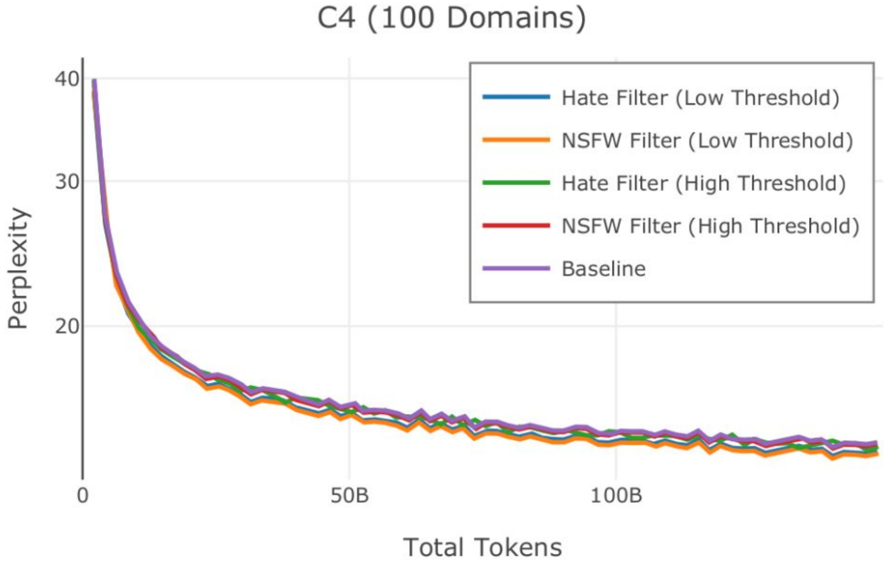
Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research, L. Soldaini et al , 2024

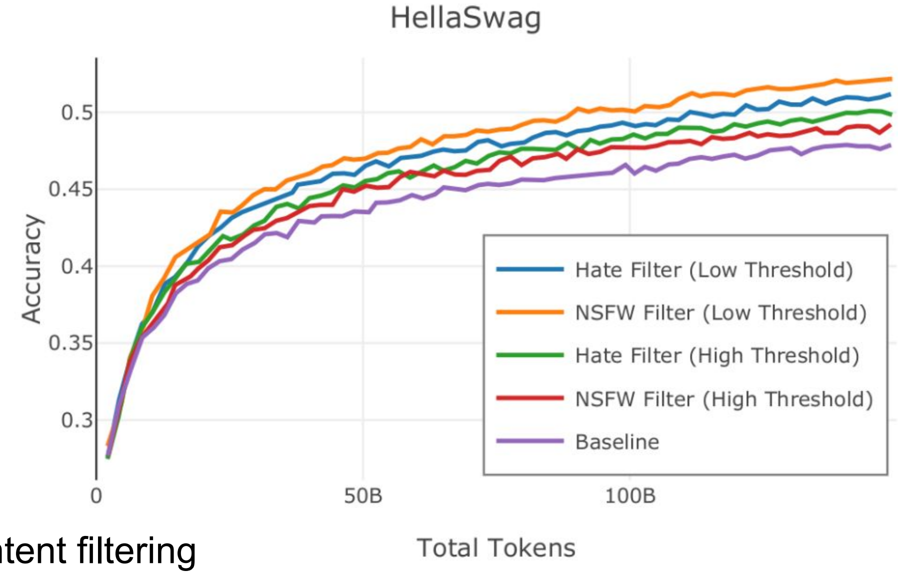
Dolma





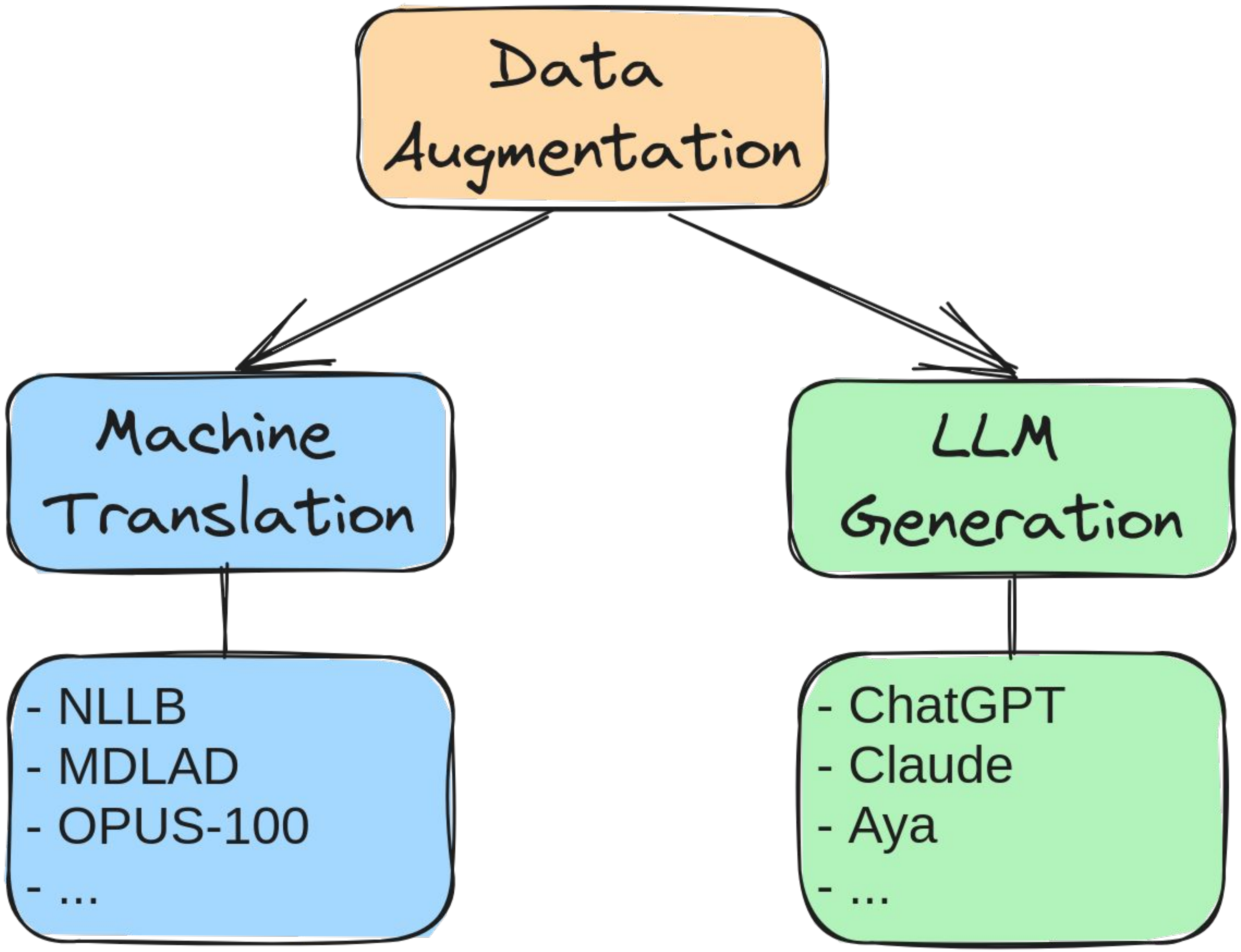
Quality filtering



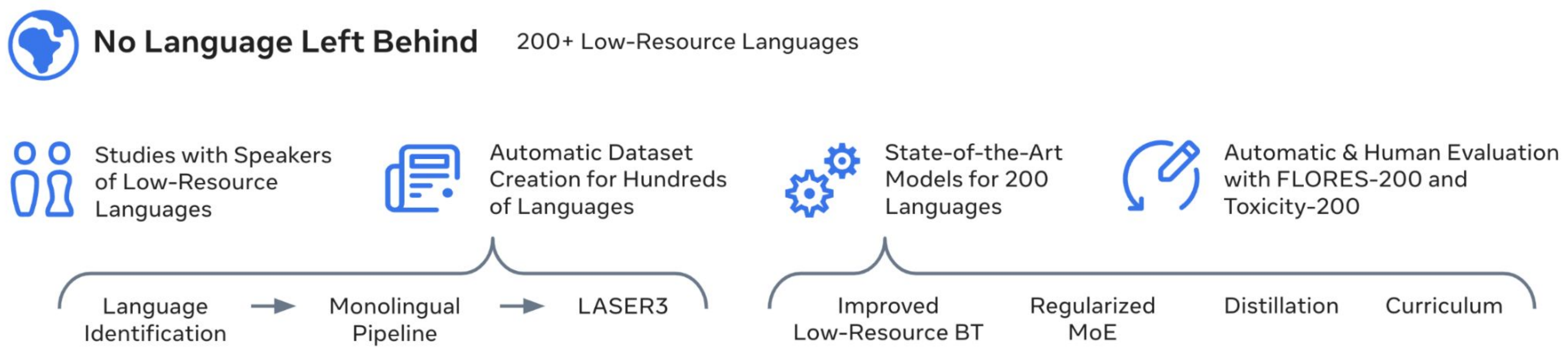


Content filtering

Data Augmentation



NLLB



- 200 languages
- Sparsely Gated Mixture of Experts
- Trained on data tailored for low-resource languages
- 44% BLEU relative to the previous state-of-the-art
- Variants: distilled-600M, 1.3B, distilled-1.3B,3.3B, moe-54B

MADLAD

- MADLAD-400 is a multilingual machine translation model based on the T5 architecture
- Trained on 250 billion tokens covering over 450 languages using publicly available data.
- MADLAD variants: 3B, 7B and 10B

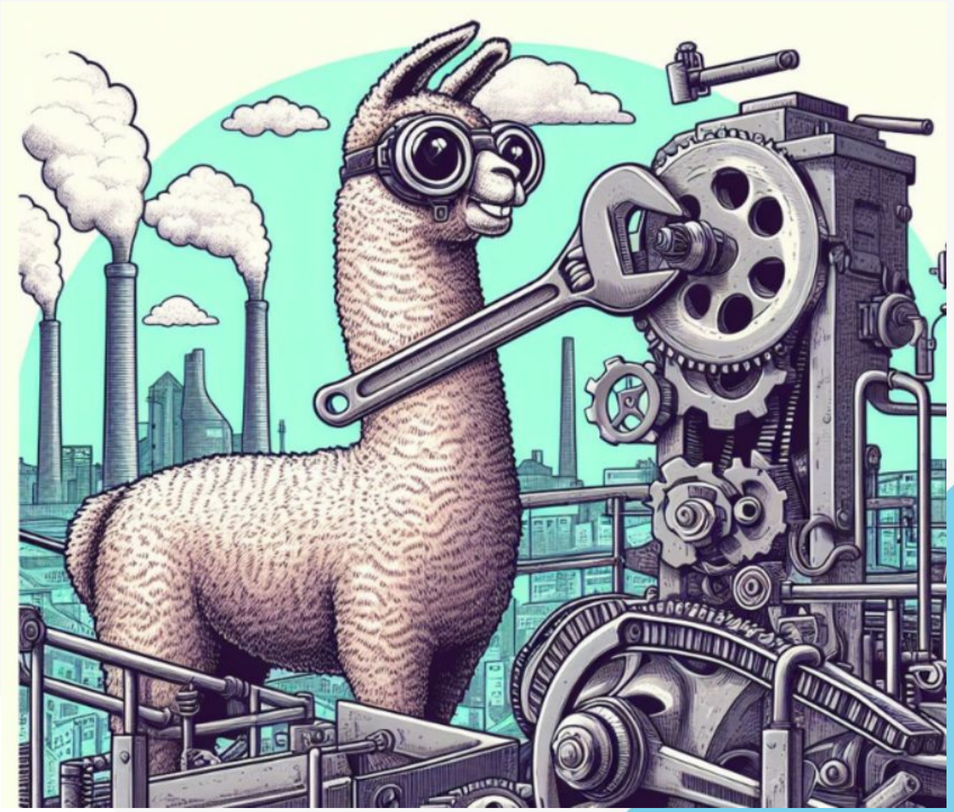
Continent	# Languages
Asia	149
Americas	66
Africa	87
Europe	89
Oceania	26
Constructed	2

Madlad-400: A multilingual and document-level large audited dataset (Kudugunta et al., 2024)

Limitations of Data Augmentation

- Accuracy of Machine Translation varies by content
- Risks of distortion of the semantic using Machine Translation
- Could carry model bias into augmented data
- Copyright restriction on LLM generated data

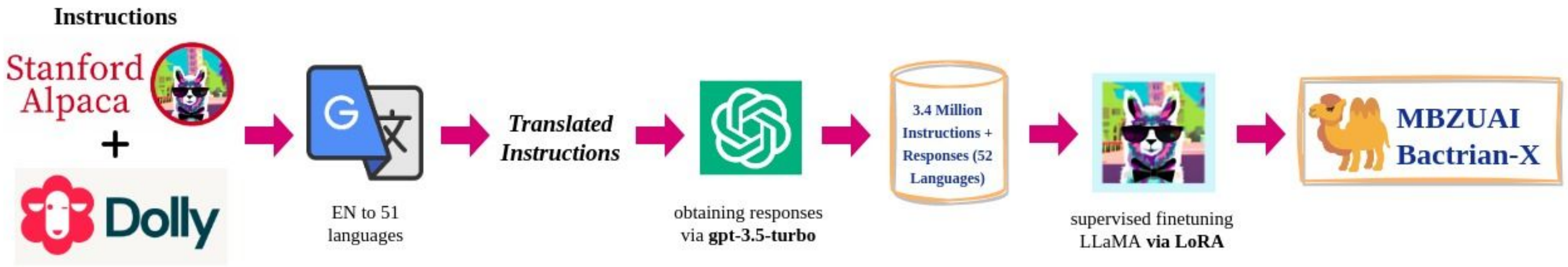
Instruction-Tuning Data



https://www.datacamp.com/tutorial/fine-tuning-llama-2

Instruction-Tuning Datasets

- **Bactrian-X:**
  - 3.4M pairs of instructions and responses in 52 languages
  - alpaca-52k, and dolly-15k translated into 52 languages using gpt-3.5-turbo



- [MBZUAI/bactrian-x-llama-7b-lora]()
- [MBZUAI/bactrian-x-llama-13b-lora]()
- [MBZUAI/bactrian-x-bloom-7b1-lora]()

Instruction Tuning Datasets

Dataset	#Instances	#Langs	% English	Generation method	Permissive license
Llama2 IFT data [Touvron et al., 2023]	NA	27	90%	Human-annotations SFT datasets	✗
Alpaca [Taori et al., 2023]	52K	1	100%	Synthetic data generation IFT datasets	≈
P3 [Sanh et al., 2022]	12M	1	100%	Template generation given applied to English datasets	✓
Flan 2022 [Longpre et al., 2023a]	15M	60	100%	Template generation applied to English datasets	✓
xP3 [Muennighoff et al., 2023c]	81M	46	39%	Template generation applied to English datasets	✓
Sweinstruct [Holmström & Doostmohammadi, 2023]	68K	1	0%	Machine translation English IFT datasets	≈
Okapi [Dac Lai et al., 2023]	158K	26	45%	Machine translation English IFT datasets	✓
Bactrian-X [Li et al., 2023a]	3.4M	52	2%	Machine translation + synthetic data generation	≈
Aya Dataset	204K	65	2%	Original IFT Human-annotations	✓
Aya Collection	513M	114	3.5%	Template Generation and translating existing datasets	✓

Aya Dataset

Data Card for the Aya Dataset

The **Aya** Dataset is a multilingual instruction fine-tuning dataset curated by an open-science community. The dataset contains a total of 204,114 annotated prompt-completion pairs.

- Curated by: 2,007 contributors from 110 countries
- Language(s): 65 languages
- License: Apache 2.0
- Repository: https://huggingface.co/datasets/CohereForAI/aya_dataset

Authorship

Publishing Organization:
Cohere For AI

Industry Type:
Not-for-profit - Tech

Contact Details:
https://aya.for.ai/

Example of Data Points

The dataset contains multilingual prompts and completions in the following format: `{prompt: "What day is followed by Saturday?", completion : "Saturday is followed by Sunday.", language: "English" }`

Aya Collection

Data Card for the Aya Collection

The **Aya** Collection incorporates instruction-style templates from fluent speakers and applies them to a curated list of 44 datasets. It also includes translations of 19 instruction-style datasets into 101 languages. This collection provides 513,579,625 instances of prompts and completions covering a wide range of tasks..

- Curated by: 2007 contributors from 110 countries
- Language(s): 114 languages
- License: Apache 2.0
- Repository: https://huggingface.co/datasets/CohereForAI/aya_collection

Authorship

Publishing Organization:
Cohere For AI

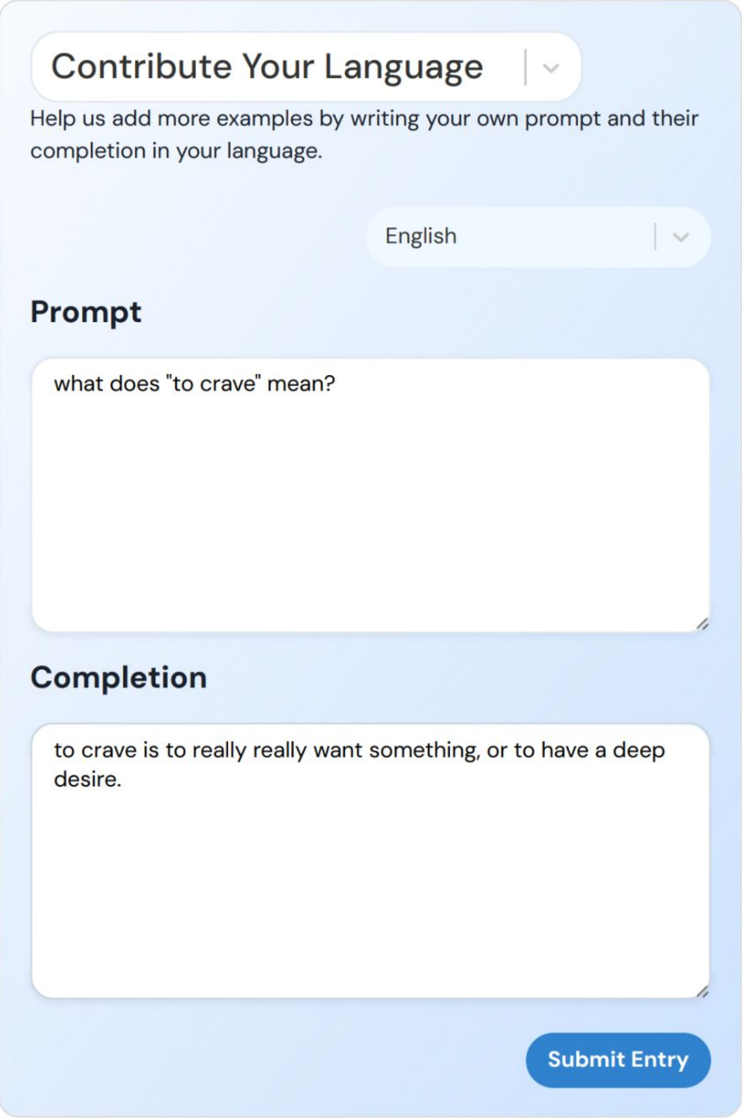
Industry Type:
Not-for-profit - Tech

Contact Details:
https://aya.for.ai

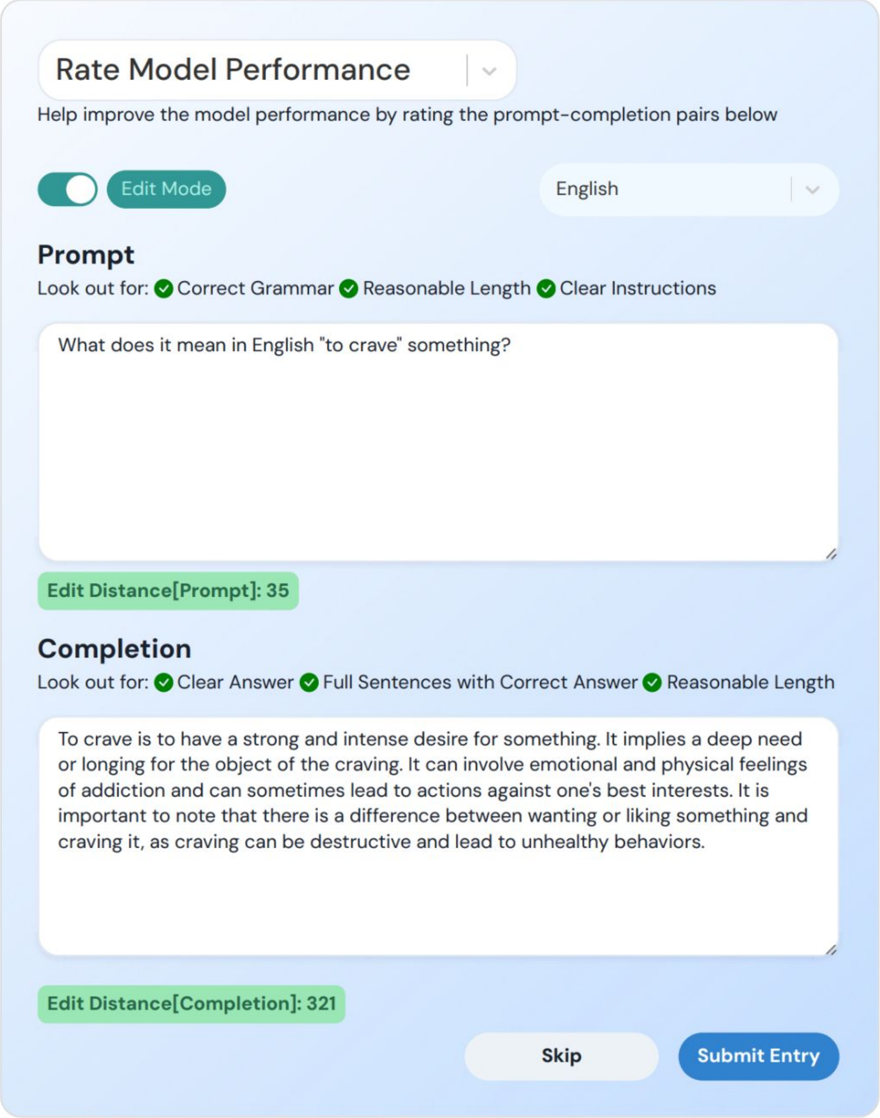
Example of Data Points

The dataset contains multilingual prompts and completions in the following format: `{'prompt': "Generate an article for the given headline: {{headline}}", 'completion': "{{news_article}}", 'lang': "English" }`

Aya Annotation Platform



(a) Example of an original annotation contribution.



(b) Example of a re-annotation contribution.

Aya Annotation Platform

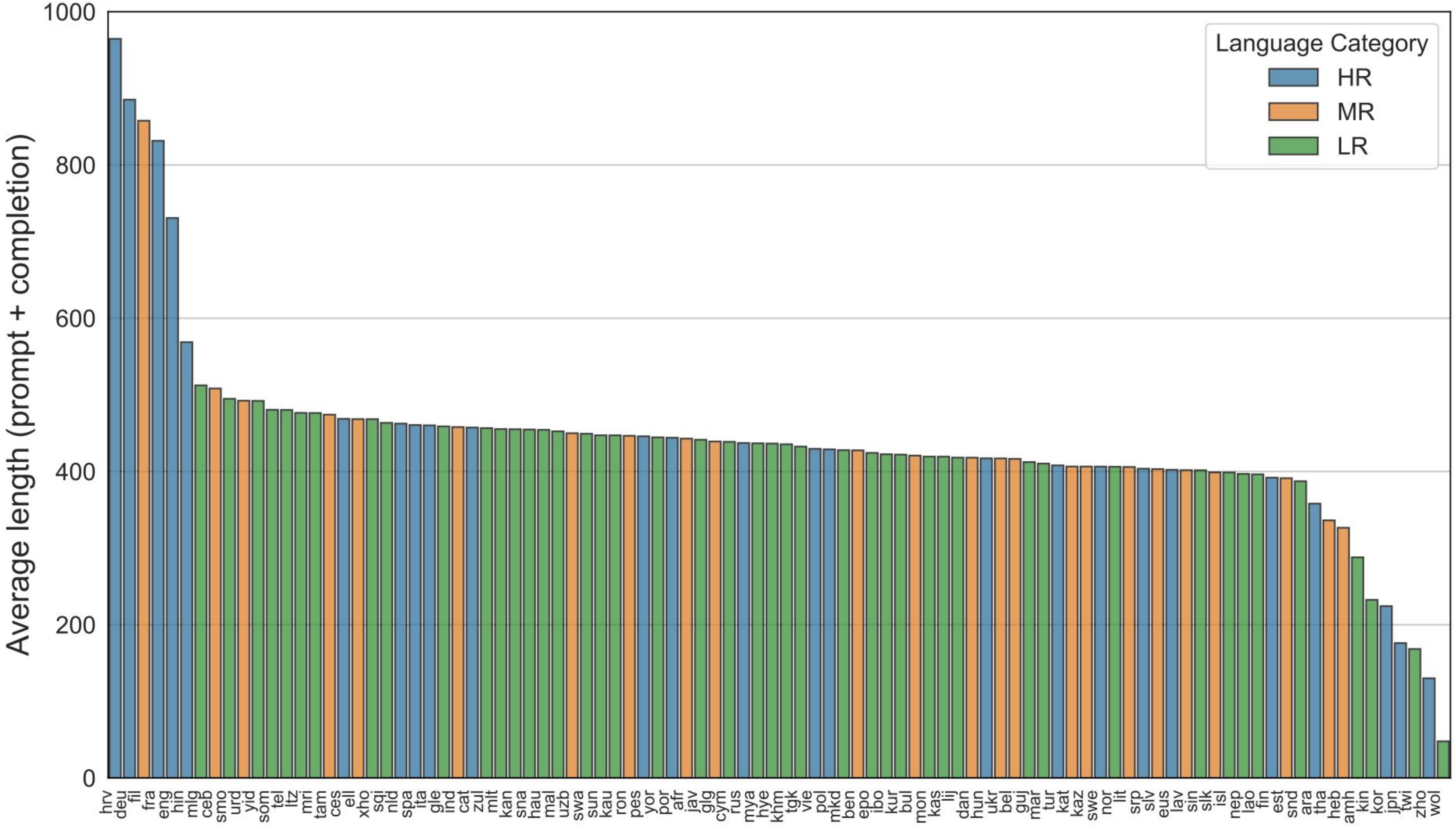


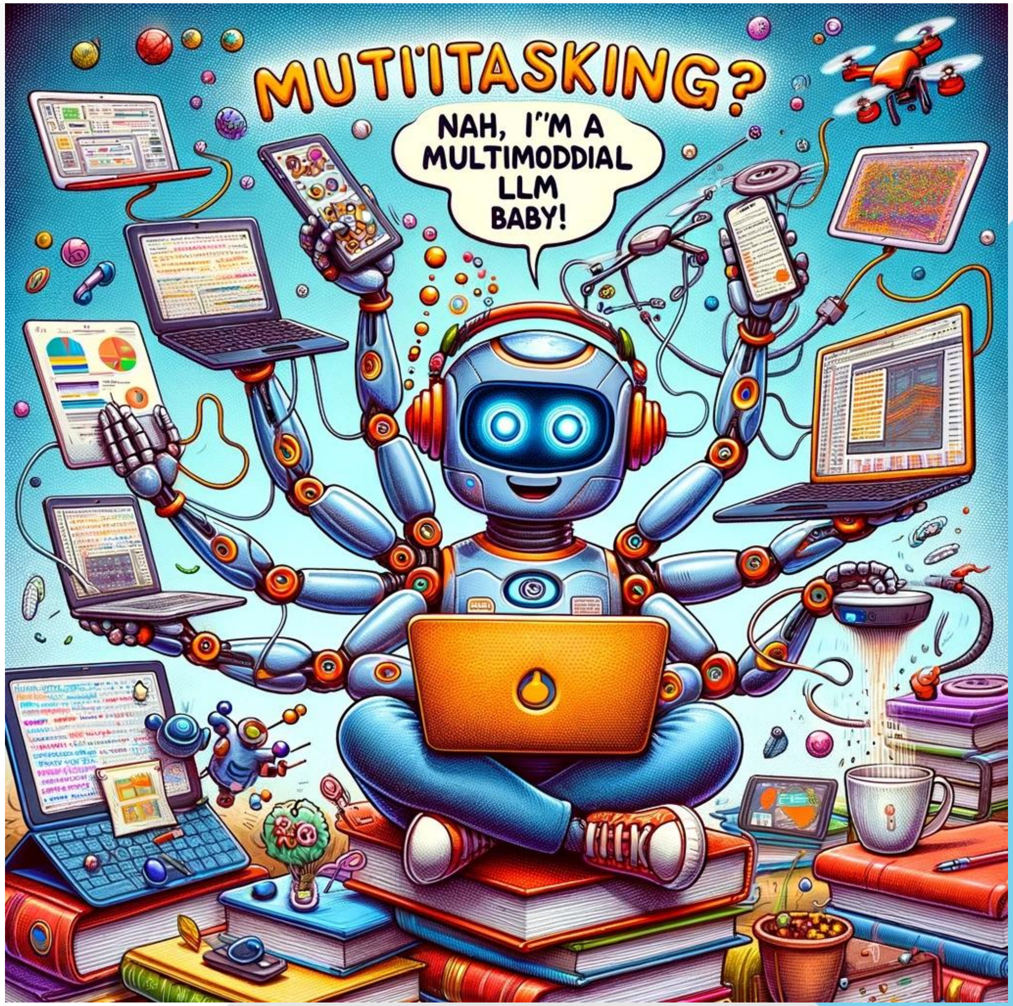
Figure 15: The average length of prompts and completions for high (HR), medium (MR) and low-resource (LR) languages in **Aya** Collection.

Multimodal LLMs

You

Generate a fun meme about multimodal LLMs like yourself

DALL·E



Why we need multimodal?

- Real World Environment inherently multimodal
- Utilization of Diverse channel: speech, sound, vision, touch among others for ***better*** knowledge acquisition

Why we need multimodal?

- The high-quality representation present in pretrained (uni)modal **Foundation models**
- The cognitive power of **LLMs**
- To empower various **MM tasks**

Harness the power of Multimodal LLMs for better understanding, reasoning and generation capabilities!

Capabilities and Modalities

Core tasks MMLLMs focus on are:

Understanding

- Image + Text ⟶ Text
- Video + Text ⟶ Text
- Audio/Speech + Text ⟶ Text
- 3D + Text ⟶ Text
- Many ⟶ Text

Generation

- Image + Text ⟶ Image + Text
- Speech/Audio + Text ⟶ Speech/Audio + Text
- Many ⟶ Image + Text
- Many ⟶ Many

MM-LLMs: Recent advances in multimodal large language models (Zhang, Duzhen, et al. arXiv 2024)

Why we need multimodal?

- **Multimodal LLMs (MMLLMs) harness**

  - The high-quality representation present in pretrained unimodal **Foundation models**

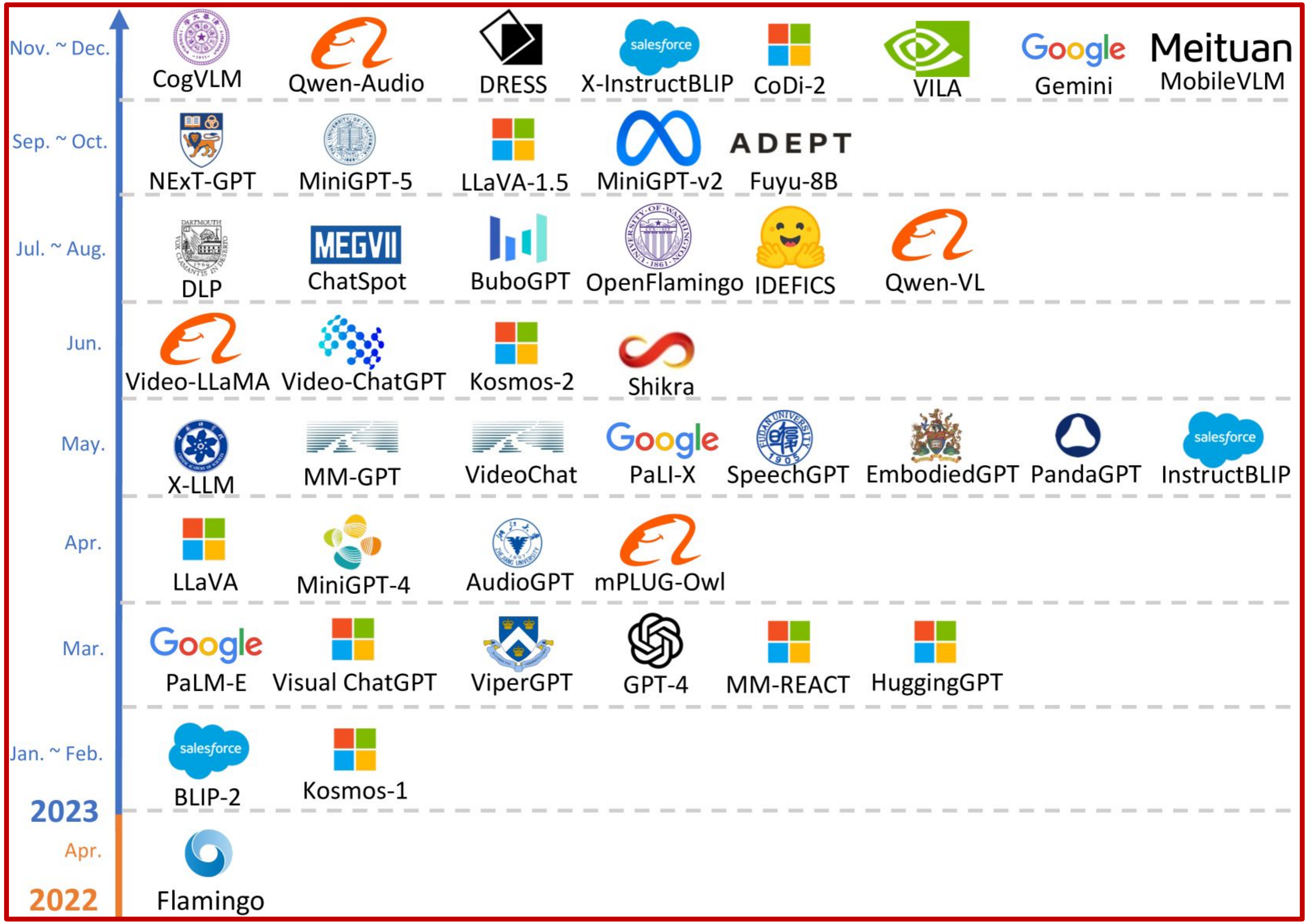
  - The cognitive power of **LLMs**

  - To empower various **MM tasks**

- **Core Challenge:** How to **connect** the LLM with other modalities for understanding and generation capabilities?

Refining Alignment between different Modalities and the Text-LLMs!

Overview of MMLLMs



2024
Jan. ~ March

DeepSeek-VL, ASMv2, AnyGPT, VisLingInstruct, ViGoR, SPHINX-X, CogCoM, Video-LaVIT, VLGuard, LLaVA-NeXT, MoE-LLaVA LLaVA-MoLE, InternLM-XComposer2 WebVoyager, Yi-VL, Vary-toy, KAM-CoT, RPG, MLLM-Tool, SkyEyeGPT, MM-Interleaved, DiffusionGPT, α-UMi, ModaVerse, GroundingGPT, ..

MM-LLMs: Recent advances in multimodal large language models (Zhang, Duzhen, et al. arXiv 2024)

Research on MMLLMs

Understanding

I+T→T: BLIP-2 (Li et al., 2023e), Kosmos-1 (Huang et al., 2023c), PaLM-E (Driess et al., 2023), ViperGPT (Surís et al., 2023), LLaVA (Liu et al., 2023e), MiniGPT-4 (Zhu et al., 2023a), mPLUG-Owl (Ye et al., 2023b), Otter (Li et al., 2023b), MultiModal-GPT (Gong et al., 2023), PandaGPT (Su et al., 2023), PaLI-X(Chen et al.) LLaVA-Med (Li et al., 2023d), LLaVAR (Zhang et al., 2023h), mPLUG-DocOwl($\mathbf{I_D}$) (Ye et al., 2023a), DLP (Jian et al., 2023), ChatSpot (Zhao et al., 2023b), OpenFlamingo (Awadalla et al., 2023), Chinese-LLaVA (LinkSoul-AI., 2023), ASM (Wang et al., 2023c), BLIVA (hu2, 2023), IDEFICS (IDEFICS, 2023), Qwen-VL (Bai et al., 2023b), Kosmos-2.5 (Lv et al., 2023), InternLM-XComposer (Zhang et al., 2023f), JAM (Aiello et al.), LLaVA-1.5 (Liu et al., 2023d), MiniGPT-v2 (Chen et al., 2023d), Fuyu-8B (Bavishi et al., 2023), CogVLM(Wang et al., 2023b), mPLUG-Owl2 (Ye et al., 2023c), Monkey (Li et al., 2023l), Volcano (Lee et al., 2023), DRESS (Chen et al., 2023i), LION (Chen et al., 2023c), DocPedia($\mathbf{I_D}$) (Feng et al., 2023), ShareGPT4V(Chen et al., 2023f), VIM (Lu et al., 2023b), mPLUG-PaperOwl($\mathbf{I_D}$)(Hu et al., 2023a), RLHF-V (Yu et al., 2023b), Silkie (Li et al., 2023g), Lyrics (Lu et al., 2023a), VILA (Lin et al., 2023), CogAgent (Hong et al., 2023), Osprey (Yuan et al., 2023a), V* (Wu and Xie, 2023), MobileVLM (Chu et al., 2023a), TinyGPT-V (Yuan et al.), DocLLM($\mathbf{I_D}$) (Wang et al., 2023a), LLaVA-ϕ (Zhu et al., 2024c), Yi-VL(Team., 2023) KAM-CoT(Mondal et al.), InternLM-XComposer2 (Dong et al., 2024b), MoE-LLaVA (Lin et al., 2024a), LLaVA-MoLE (Chen et al., 2024), LLaVA-NeXT (Liu et al., 2024b), VLGuard (Zong et al., 2024), MobileVLM V2 (Chu et al., 2024), ViGoR(Yan et al., 2024), VisLingInstruct (Zhu et al., 2024b)
V+T→T: VideoChat (Li et al., 2023f), Video-ChatGPT (Maaz et al., 2023), Dolphins (Ma et al., 2023)
A+T→T: SALMONN (Tang et al., 2023a), Qwen-Audio (Chu et al., 2023b)
3D+T→T: 3DMIT (Li et al., 2024b)
Many→T: Flamingo (Alayrac et al., 2022), MM-REACT (Yang et al., 2023b), X-LLM (Chen et al., 2023b) InstructBLIP (Dai et al., 2023), EmbodiedGPT (Mu et al., 2023), Video-LLaMA (Zhang et al., 2023e), Lynx (Zeng et al., 2023), AnyMAL(Moon et al., 2023), LanguageBind (Zhu et al., 2024a), LLaMA-VID (Li et al., 2023j), X-InstructBLIP (Panagopoulou et al., 2023), InternVL (Chen et al., 2023j)

Generation

I+T→I+T: FROMAGe($\mathbf{I_R}$) (Koh et al., 2023b), Visual ChatGPT (Wu et al., 2023a), DetGPT($\mathbf{I_B}$)(Pi et al., 2023) GILL(Koh et al., 2023a), Kosmos-2($\mathbf{I_B}$) (Peng et al., 2023), Shikra($\mathbf{I_B}$) (Chen et al., 2023e), GPT4RoI($\mathbf{I_B}$) (Zhang et al., 2023g), SEED (Ge et al., 2023), LISA($\mathbf{I_M}$) (Lai et al., 2023), VisCPM(Hu et al., 2023b), CM3Leon(Yu et al., 2023a), LaVIT (Jin et al., 2024), DreamLLM (Dong et al., 2024a), MiniGPT-5 (Zheng et al., 2023b), Kosmos-G (Pan et al., 2023), GLaMM($\mathbf{I_M}$) (Rasheed et al., 2023), LLaVA-Plus(+$\mathbf{I_B}$&$\mathbf{I_M}$) (Liu et al., 2023f), PixelLM($\mathbf{I_M}$) (Ren et al., 2023), VL-GPT (Zhu et al., 2023b), CLOVA(+$\mathbf{I_B}$&$\mathbf{I_M}$) (Gao et al., 2023b), Emu-2 (Sun et al., 2023a), MM-Interleaved (Tian et al., 2024), DiffusionGPT (Qin et al., 2024), RPG(Yang et al., 2024),Vary-toy($\mathbf{I_B}$) (Wei et al., 2024), CogCoM($\mathbf{I_B}$) (Qi et al., 2024), SPHINX-X($\mathbf{I_B}$) (Gao et al., 2024)
A/S+T→A/S+T: SpeechGPT (Zhang et al., 2023a), AudioPaLM (Rubenstein et al., 2023)
Many→I+T: Emu (Sun et al., 2024), BuboGPT($\mathbf{I_M}$) (Zhao et al., 2023d), GroundingGPT($\mathbf{I_B}$) (Li et al., 2024c)
Many→Many: GPT-4 (OpenAI, 2023), HuggingGPT (Shen et al., 2023), AudioGPT (Huang et al., 2023b) NExT-GPT (Wu et al., 2023d), ControlLLM (Liu et al., 2023i), TEAL (Yang et al., 2023a), CoDi-2(Tang et al.) Gemini (Team et al., 2023), ModaVerse (Wang et al., 2024c), MLLM-Tool(Wang et al., 2024a)

Popular: Visual Modality
Major Target Language: English

MM-LLMs: Recent advances in multimodal large language models (Zhang, Duzhen, et al. arXiv 2024)

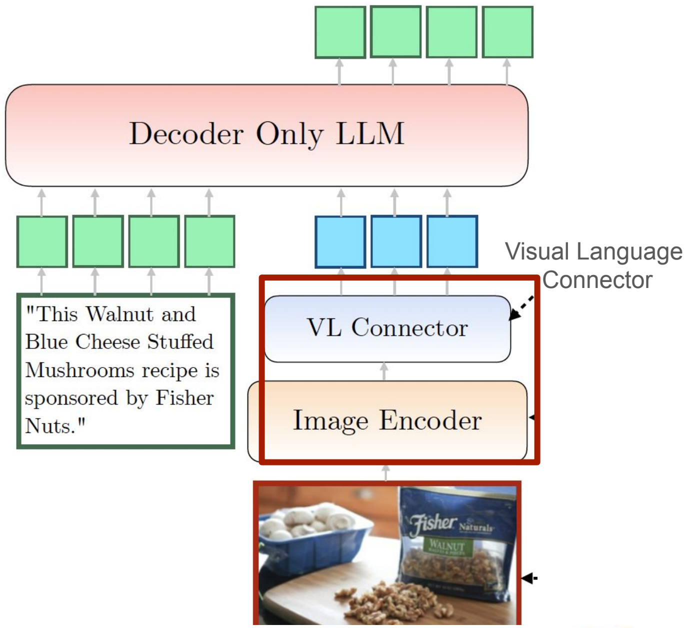
Examples MMLLMs

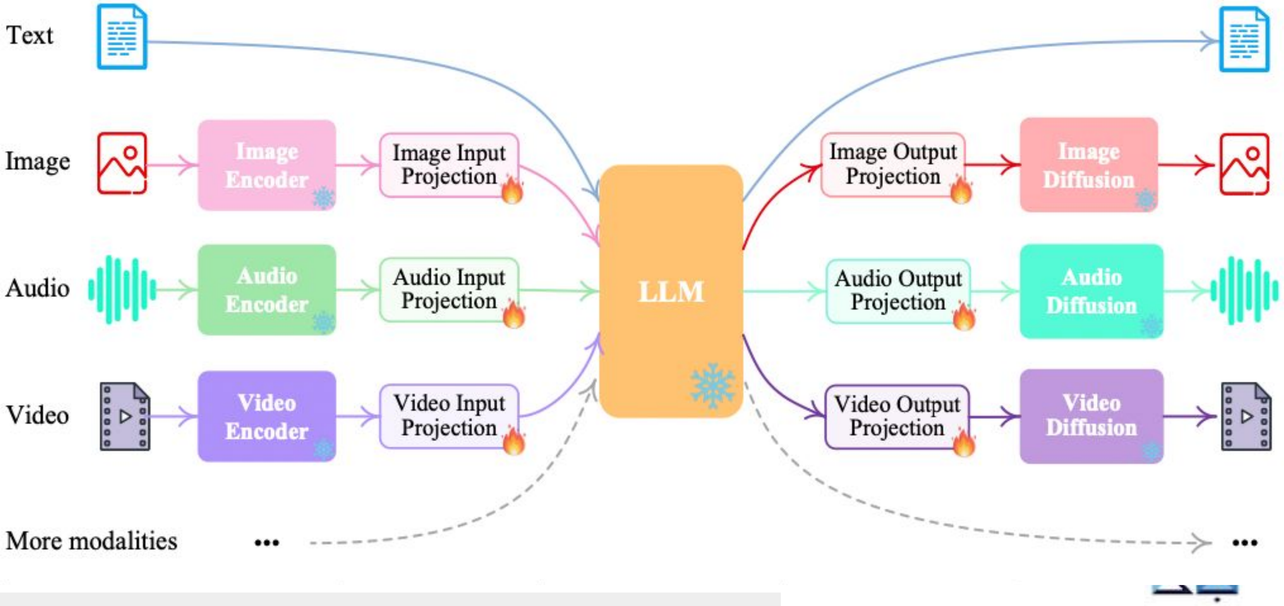
- **Gemini Family** Google Gemini ★
  - Image, Speech, Video, Text understanding → Outputs: Text and Image
  - *Ultra*: State-of-the-art performance in wide variety of complex tasks (e.g. reasoning) and multimodal tasks.
  - *Pro*: Enhanced for performance and deployability at scale.
  - *Nano* (1.8B and 3.25B): on-device application
- **ChatGPT/GPT-4V** GPT-4
  - Image, Speech, Text understanding → Outputs: Text, Image, Speech
  - Speech: Whisper Model (transcription) [Closed Information]

Gemini: a family of highly capable multimodal models. (Team, Gemini, et al., arXiv 2023)
ChatGPT can now see, hear, and speak (https://openai.com/blog/chatgpt-can-now-see-hear-and-speak)
The dawn of LLMs: Preliminary explorations with gpt-4v(ision). (Yang, Zhengyuan, et al. arXiv 2023)

Examples MMLLMs

- **MM1 Family**
    - Image, Text understanding
    - 3B, 7B to 30B, 3BX64 to 7BX32 MOE
    - Multi-image reasoning capability
- **NextGPT** ★
    - Any-to-Any Modality, Semantic understanding and reasoning
    - Text, Images, Videos, and Audios
    - LLM Vicuna (7B) [LoRA 33M]

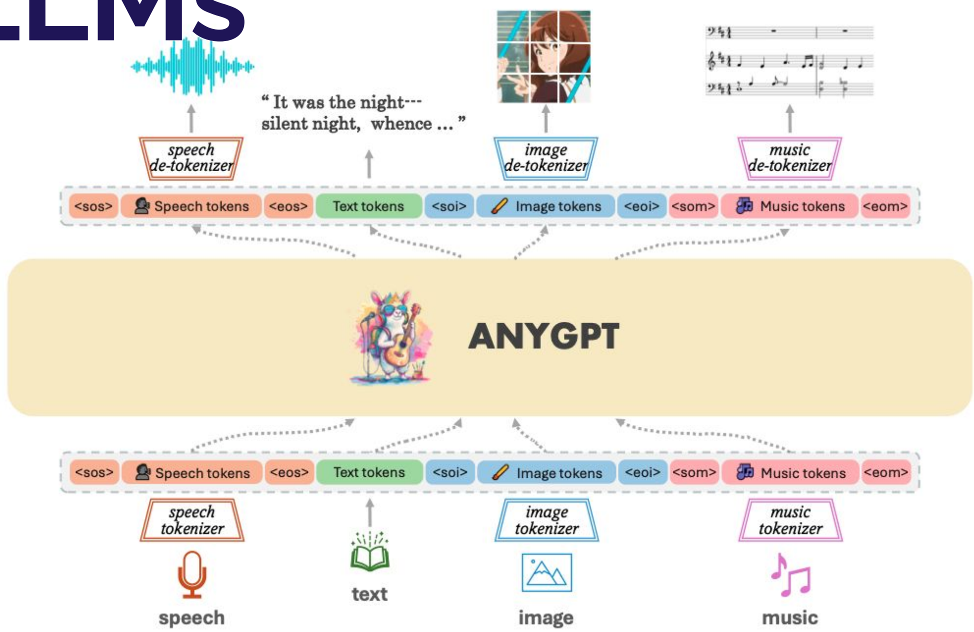


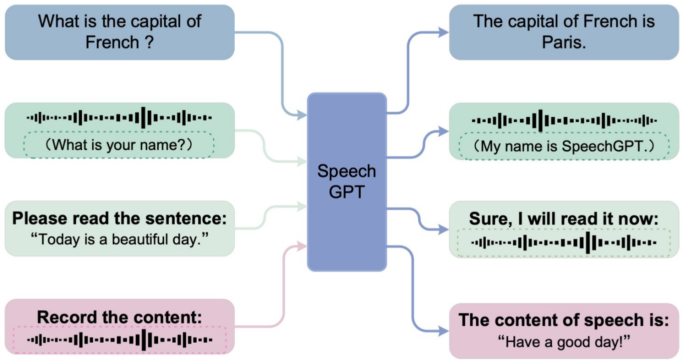


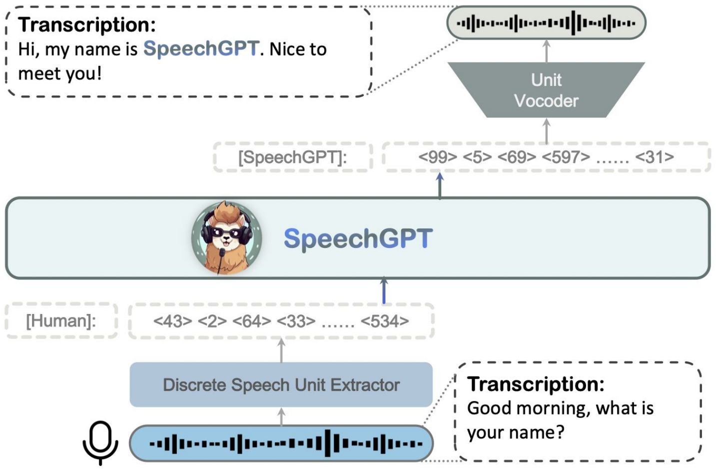
MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training. (McKinzie, Brandon, et al. arXiv 2024)
Next-gpt: Any-to-any multimodal llm (WU, Shengqiong, et al. arXiv 2023)

Examples MMLLMs

- **AnyGPT** ★
  - Any-to-Any Modality
  - Discrete Tokens representation
  - LLM LLaMA-2 7B
- **SpeechGPT** ★
  - Speech/Text → Speech/Text
  - Discrete Tokens representation
  - Spoken dialogue following ability







AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. (Zhan, Jun, et al. arXiv 2024)
SpeechGpt: Empowering large language models with intrinsic cross-modal conversational abilities. (Zhang, Dong, et al. arXiv 2023)

MMLLMs Architectures

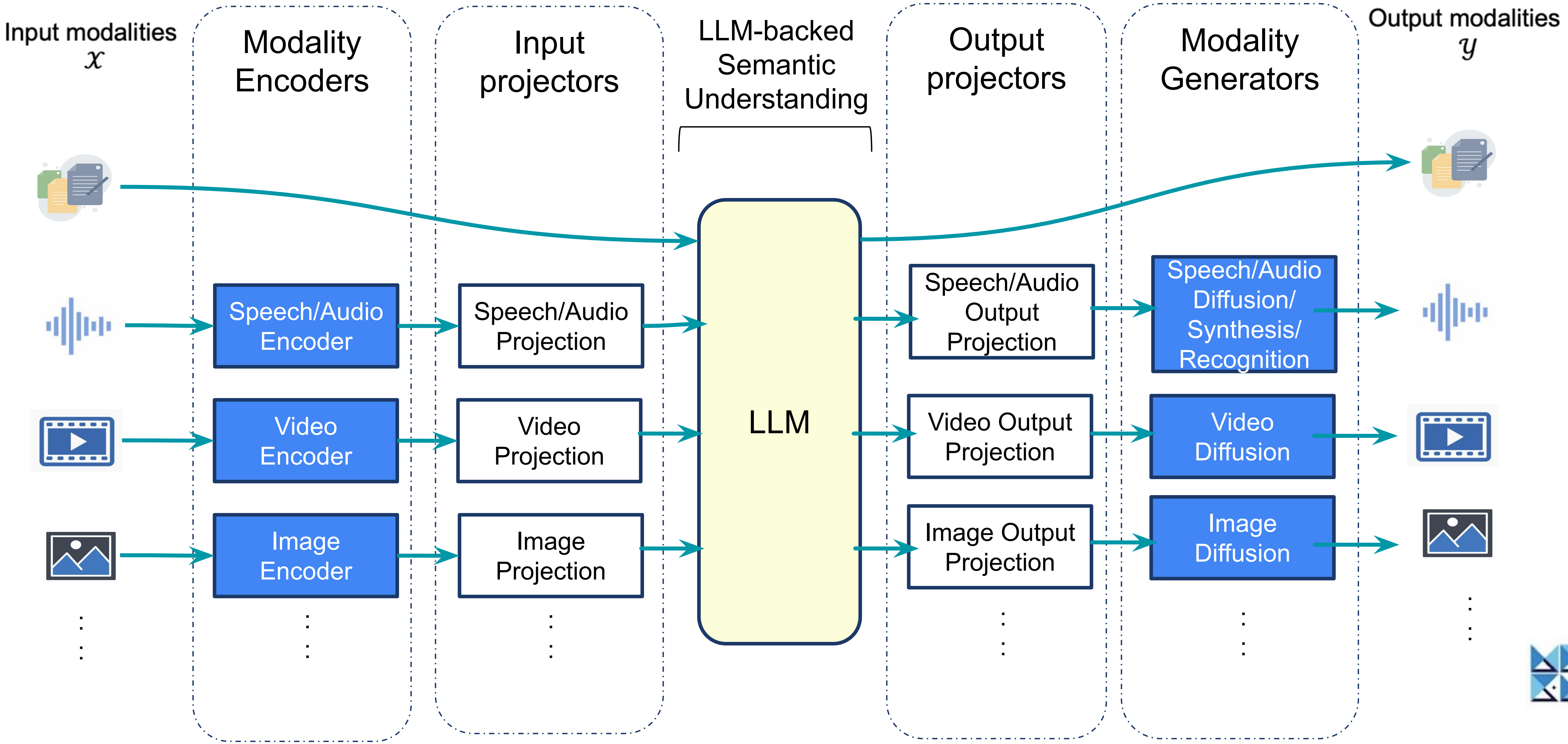
Most widely adapted MMLLMs Model Architectures:

- ★ Modality Encoder
- ★ LLM as Backbone
- ★ Modality Generator

Representation Learning → *Continuous modality representation* or *Discrete token representation*

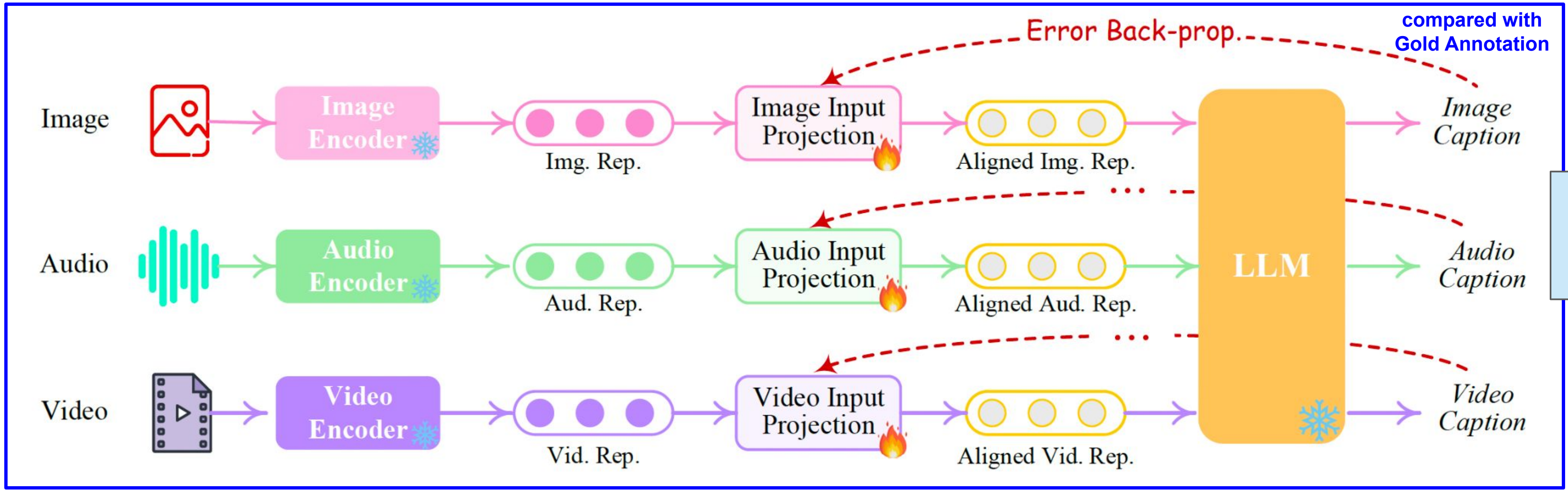
MMLLM Architectures: Continuous Representation

General Overview

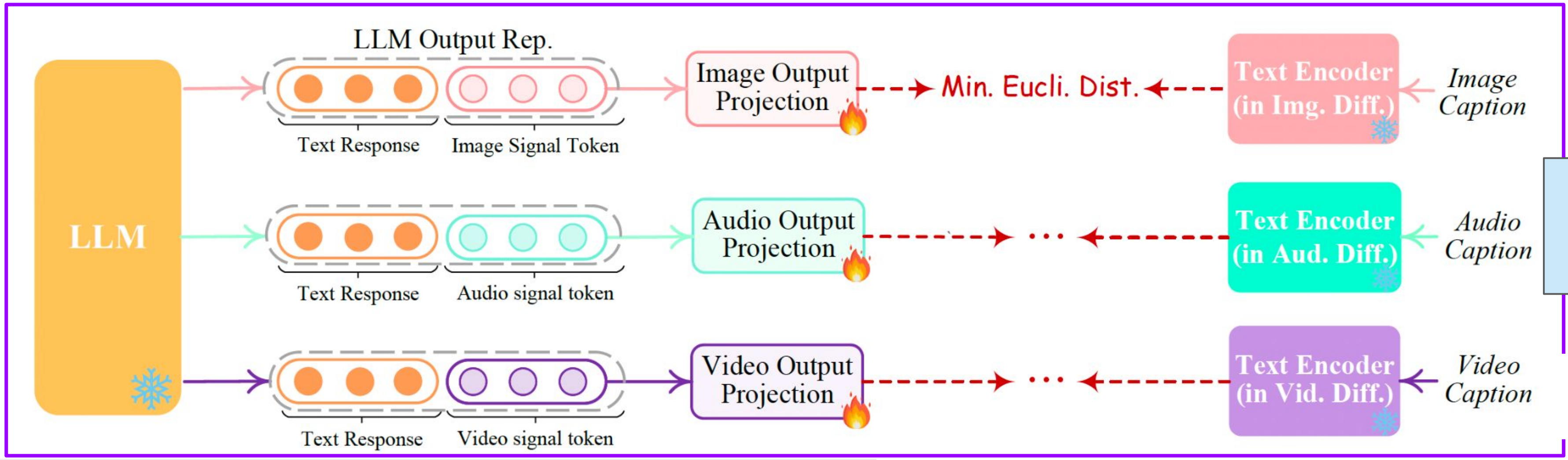


MM-LLMs: Recent advances in multimodal large language models (Zhang, Duzhen, et al. arXiv 2024)

Multimodal Alignment: Next-GPT
Continuous Representation



LLM-centric Alignment



Align Diffusion models with LLMs' output –Expensive :(

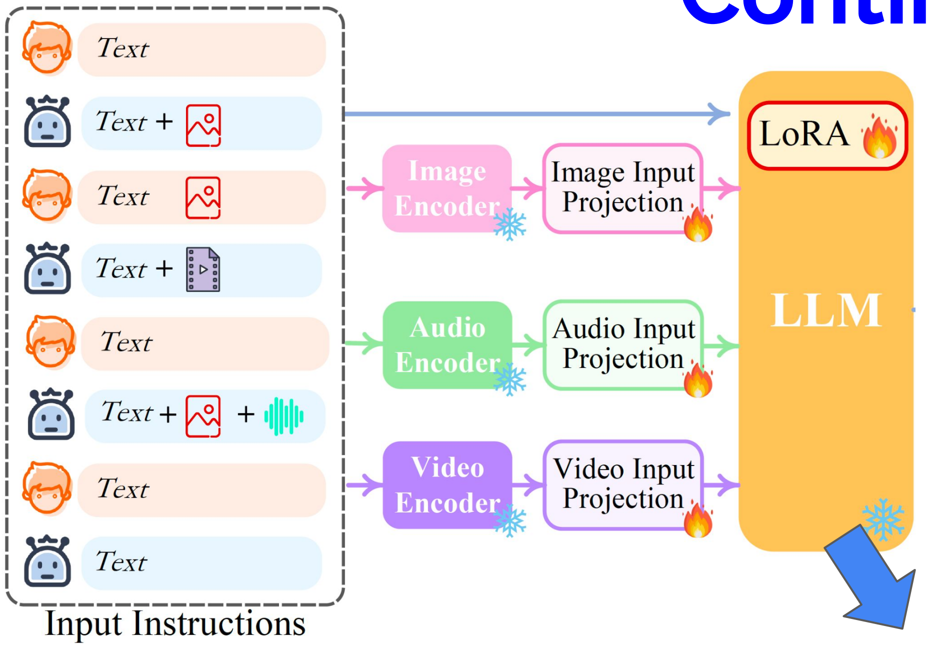
Instruction-following Alignment

Diffusion model solely conditioned on text input

Next-gpt: Any-to-any multimodal llm (WU, Shengqiong, et al. arXiv 2023)

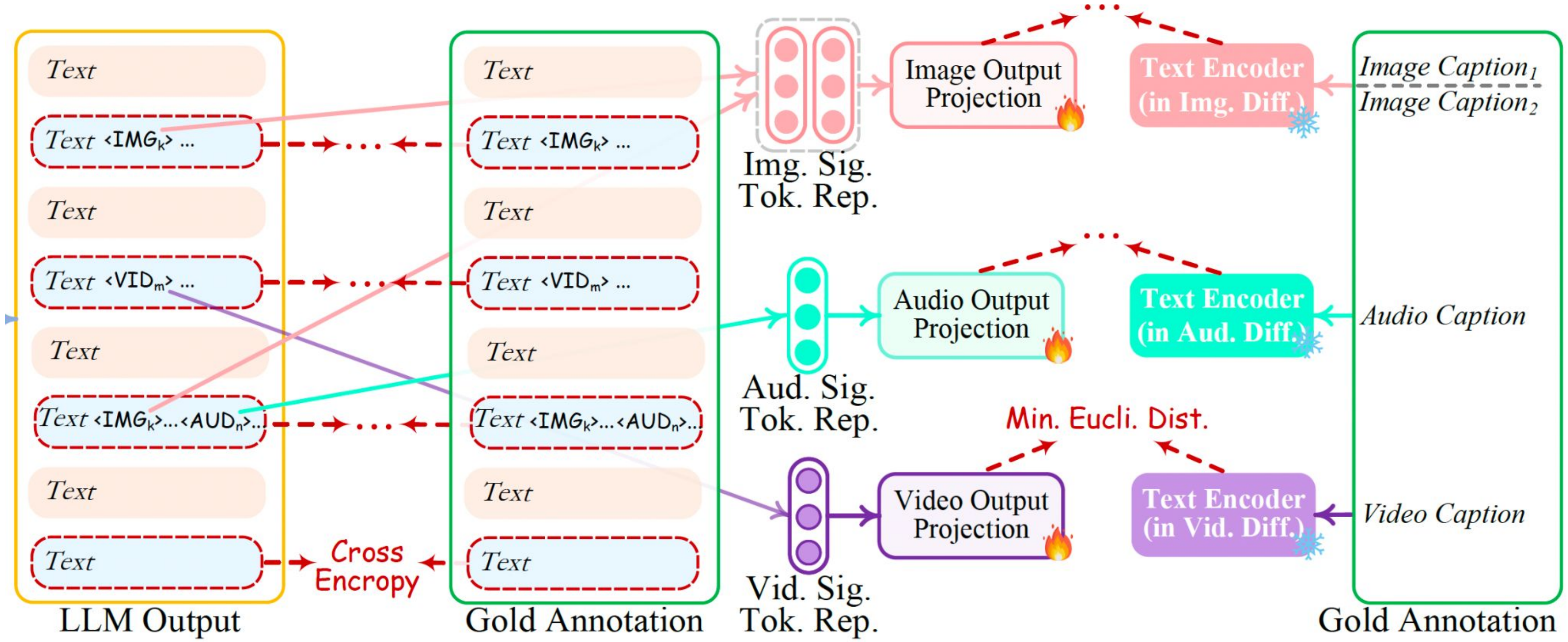
Multimodal Instruction Tuning: Next-GPT

Continuous Representation



But can the model understand and follow instruction??

Modality-switching Instruction Tuning

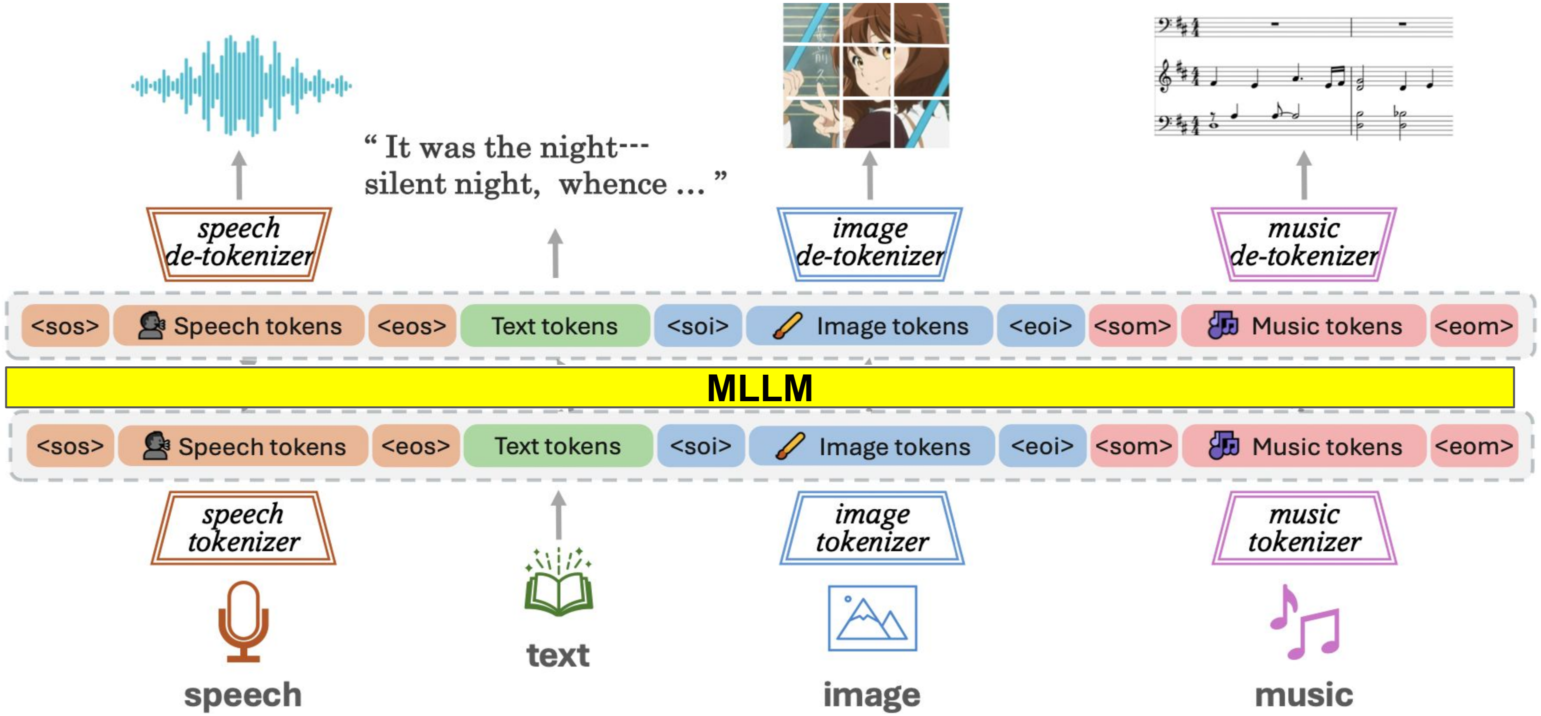


Next-gpt: Any-to-any multimodal llm (WU, Shengqiong, et al. arXiv 2023)

MMLLMs: Discrete Representation

Convert continuous representation to discrete tokens of fixed vocabulary size.

- **AnyGPT**



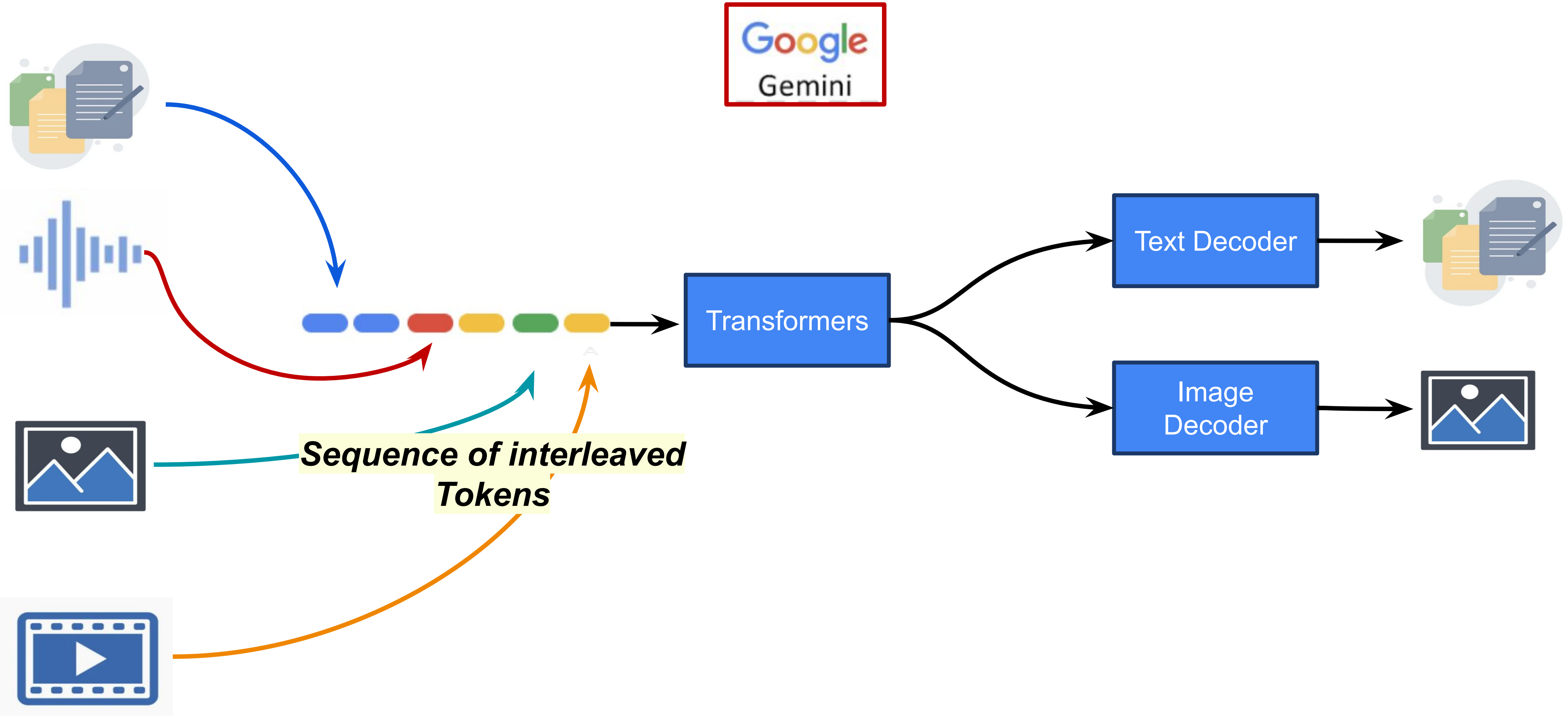
- **Gemini**

Interleaved Tokens

AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. (Zhan, Jun, et al. arXiv 2024)
Gemini: a family of highly capable multimodal models. (Team, Gemini, et al., arXiv 2023)

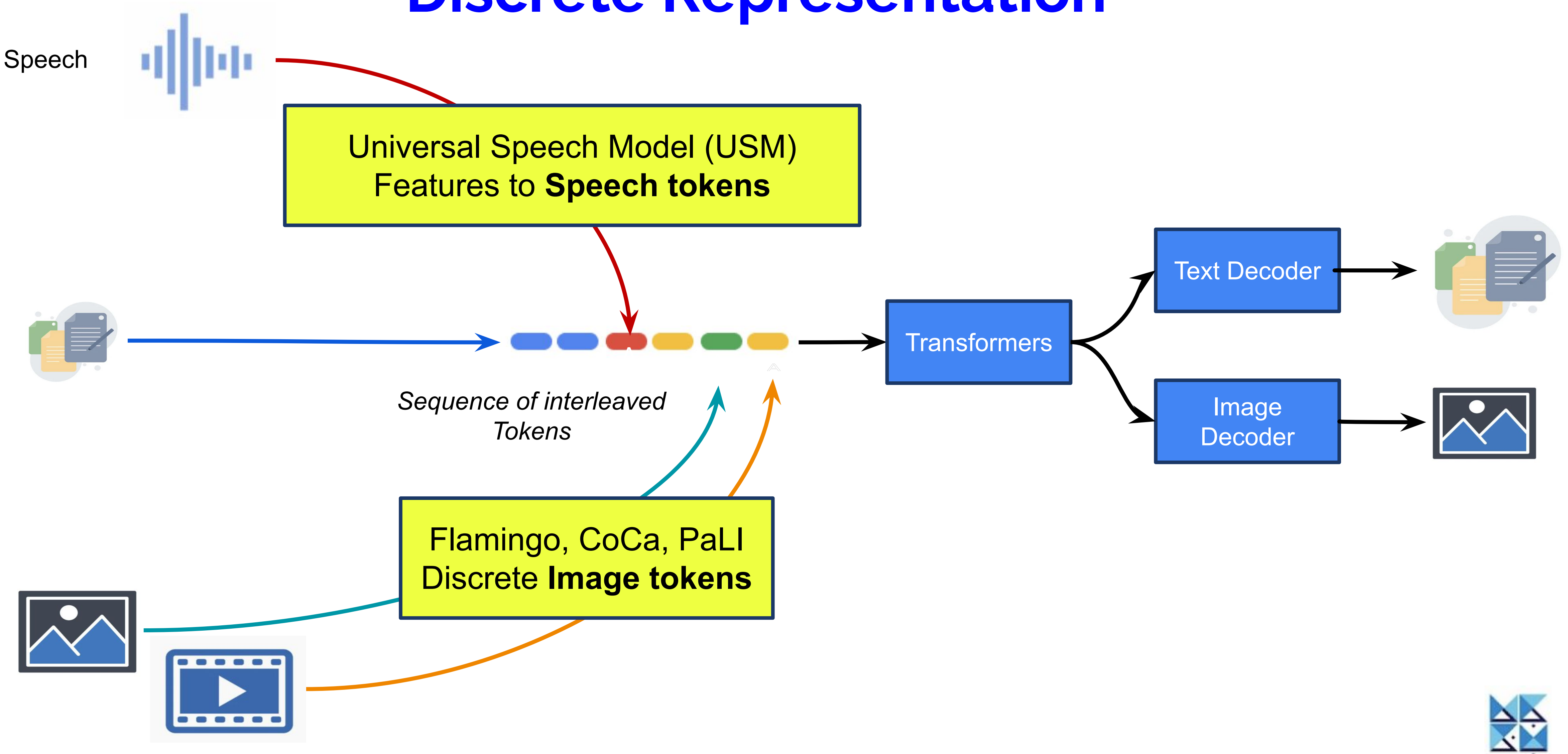
MMLLM Architectures: Gemini (closed)

Discrete Representation



Gemini: a family of highly capable multimodal models. (Team, Gemini, et al., arXiv 2023)

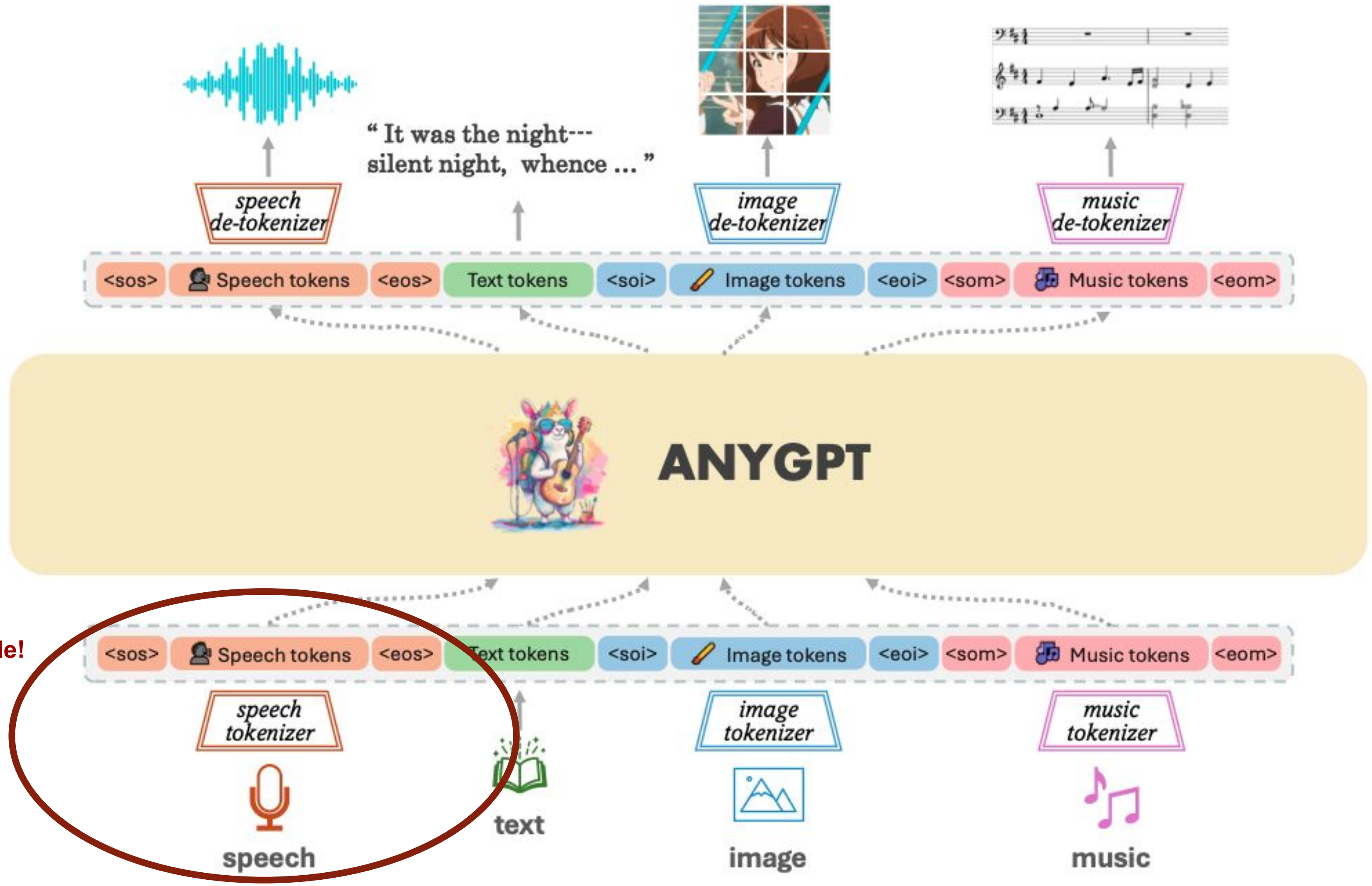
MMLLM Architectures: Gemini
Discrete Representation



Gemini: a family of highly capable multimodal models. (Team, Gemini, et al., arXiv 2023)

MMLLM Architectures: AnyGPT
Discrete Representation

Speech Modality as an example!



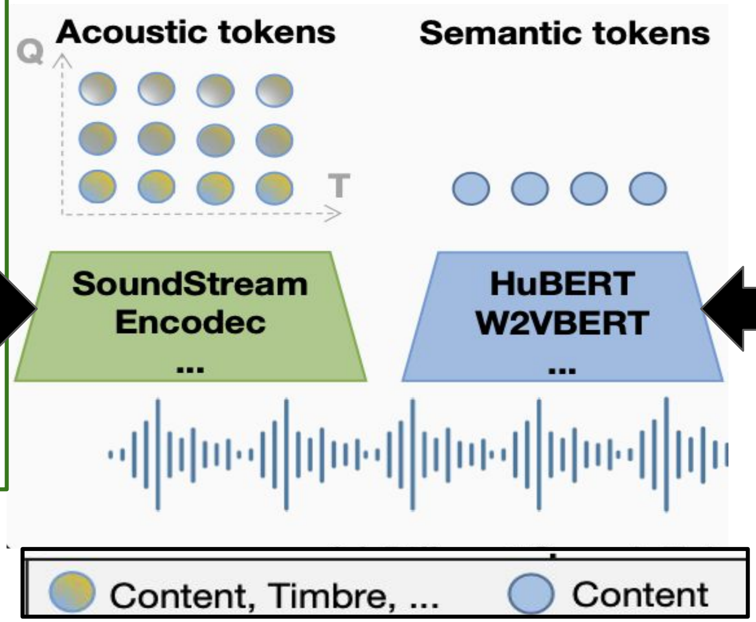
AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. (Zhan, Jun, et al. arXiv 2024)

Modality-based Tokenizers (e.g. Speech)

Acoustic Tokens

Neural audio codecs,
Reconstruction as training objective,
Residual vector quantization (RVQ) with hierarchical quantizers for discretization. Matrices consisting of two dimensions: **timesteps and quantizers.** (Zeghidour et al., 2021; Défossez et al., 2022)

Semantic Accurate Content ☹️
Speech Generation 😃



Semantic Tokens

SSL pretrained model, Masked Language modeling as training objectives and discretized with **k-mean clustering** (Hsu et al., 2021; Baevski et al., 2020; Chung et al., 2021)

Semantic Accurate Content 😃
Speech Generation ☹️

Semantic + Acoustic

Semantic Accurate Content 😃
Speech Generation 😃

<

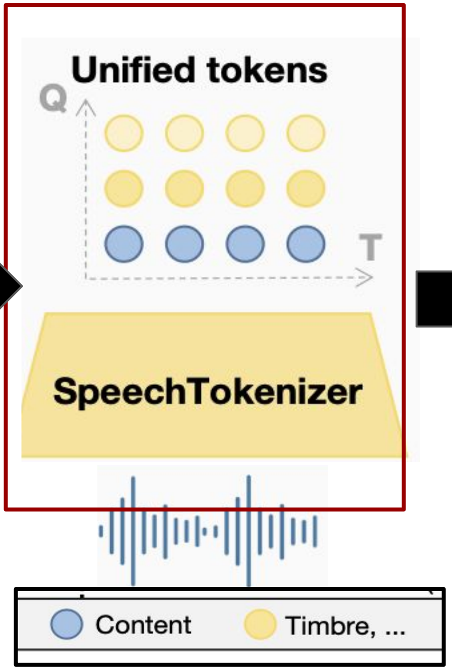
Multi-stage modeling → Complex ☹️
Error accumulation ☹️
Slower processing speed ☹️
Information redundancy ☹️

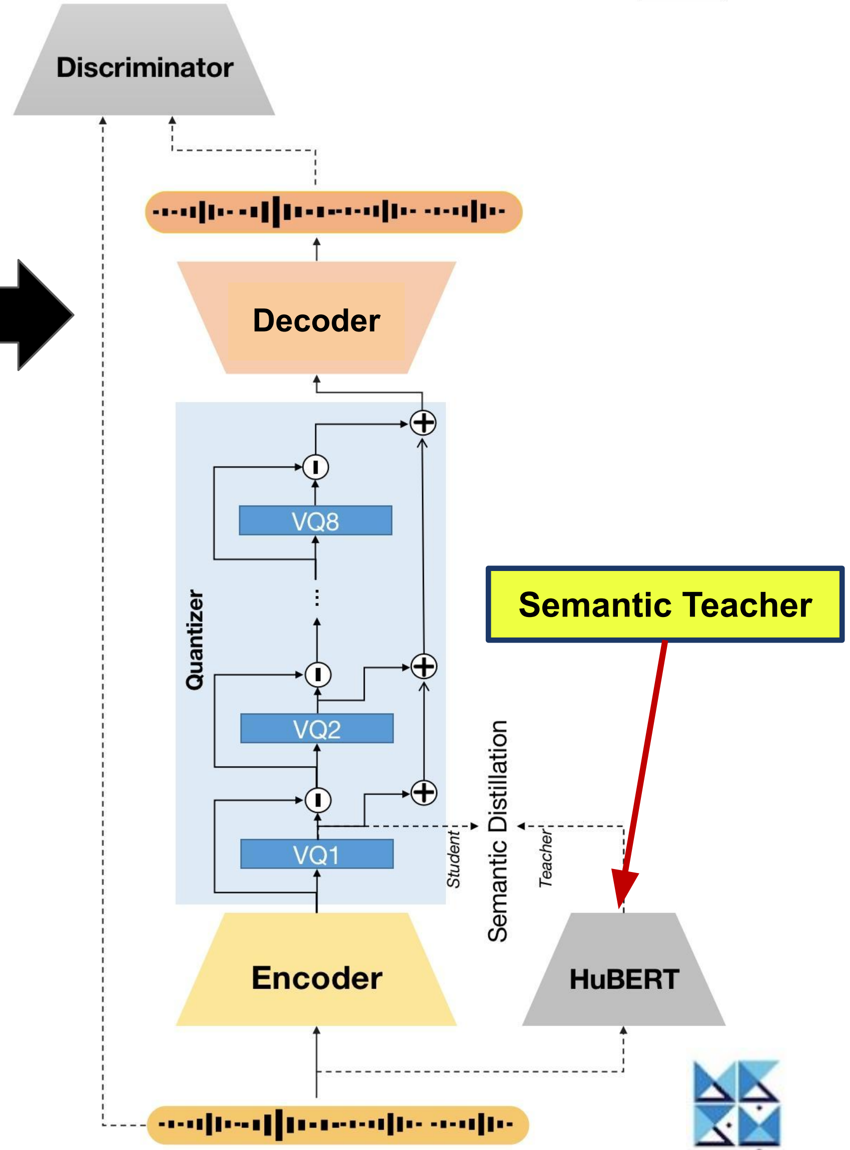
SpeechTokenizer: Unified Speech Tokenizer for Speech Language Models. (Zhang, Xin, et al. 2024)

Modality-based Tokenizers (e.g. Speech)

Unified Tokens

Information disentanglement in the RVQ structure of acoustic tokens. **First RVQ** quantizer capture **semantic tokens**. Subsequent quantizers (VQ2-VQ8) complement the remaining **acoustic/paralinguistic** information.





Speech Reconstruction Result

Tokenizer	Objective		Subjective
	WER↓	VISQOL↑	MUSHRA↑
Groundtruth	4.58	-	91.46
EnCodec	5.11	4.37	79.86
SpeechTokenizer	5.04	4.30	90.55

Content Quality Speech Quality Human acceptability

SpeechTokenizer: Unified Speech Tokenizer for Speech Language Models. (Zhang, Xin, et al. 2024)
High fidelity neural audio compression. (Défossez, Alexandre, et al. arXiv 2022)

MMLLM Architectures: AnyGPT

Modality Generator

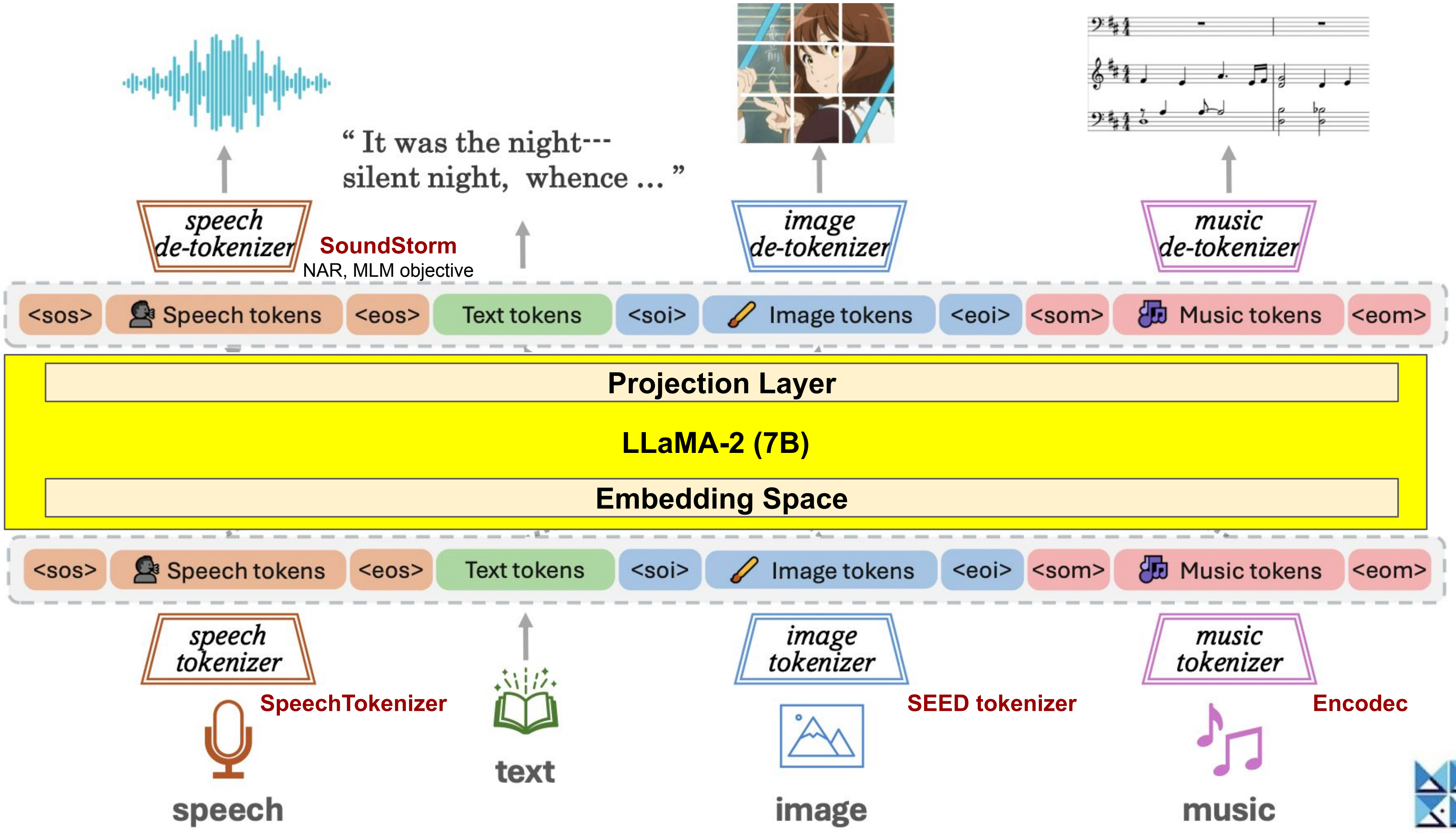
1. ***LLM content (semantic) to X modality content***
2. ***De-Tokenizer (Decoder)***

1. ***V (Text) U V (X)***
2. Initialize *V (X)* embedding randomly.
3. **Trained with Next token Prediction**

Discrete Token of size V (X)

Modality Tokenizer

Modality X



AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. (Zhan, Jun, et al. arXiv 2024)

Modality Encoders

Essence of adding MM in LLMs: Insert modality knowledge effectively

Visual Modality

- NFNet-F6
- ViT
- CLIP ViT
- Eva-CLIP ViT

Speech/Audio

- **HuBERT**
- **MMS**
- **Whisper**
- **USM** ***(close)***
- Wav2Vec2
- BEATs
- C-Former

Multilingual Capabilities!

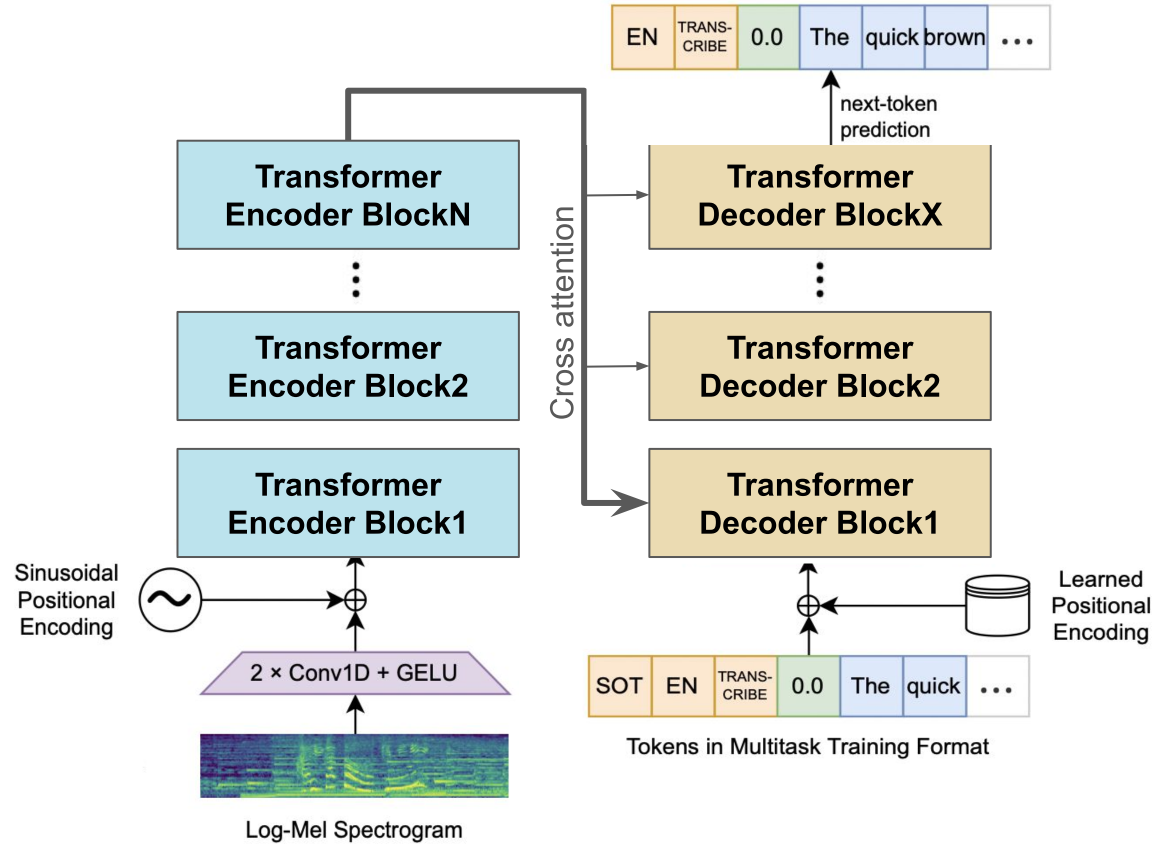
Unified (any2any)

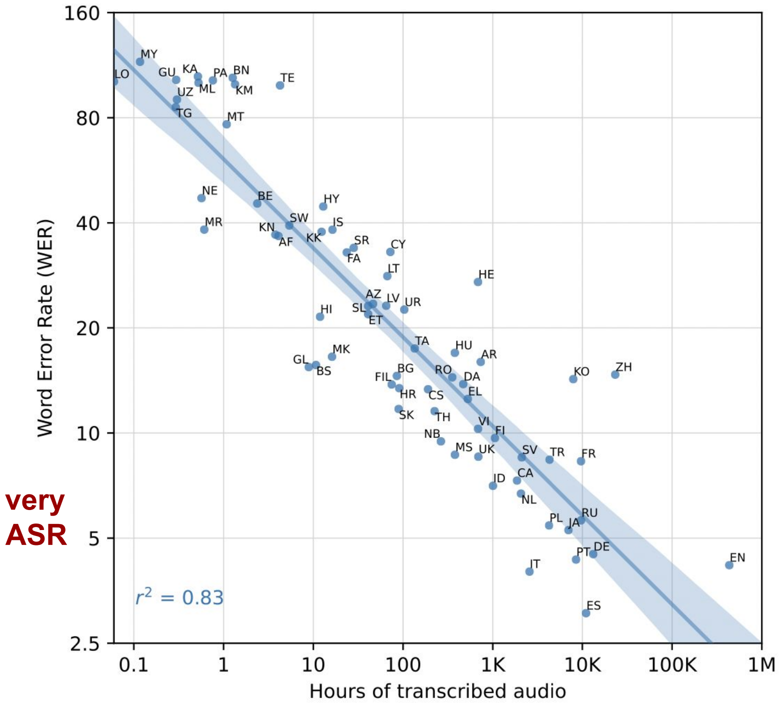
- **ImageBind**
  - Image
  - Video
  - Text
  - Audio
  - Heatmap
  - ...

MM-LLMs: Recent advances in multimodal large language models (Zhang, Duzhen, et al. arXiv 2024)
ML-SUPERB: Multilingual speech universal performance benchmark (Shi, Jiatong, et al. arXiv 2023)

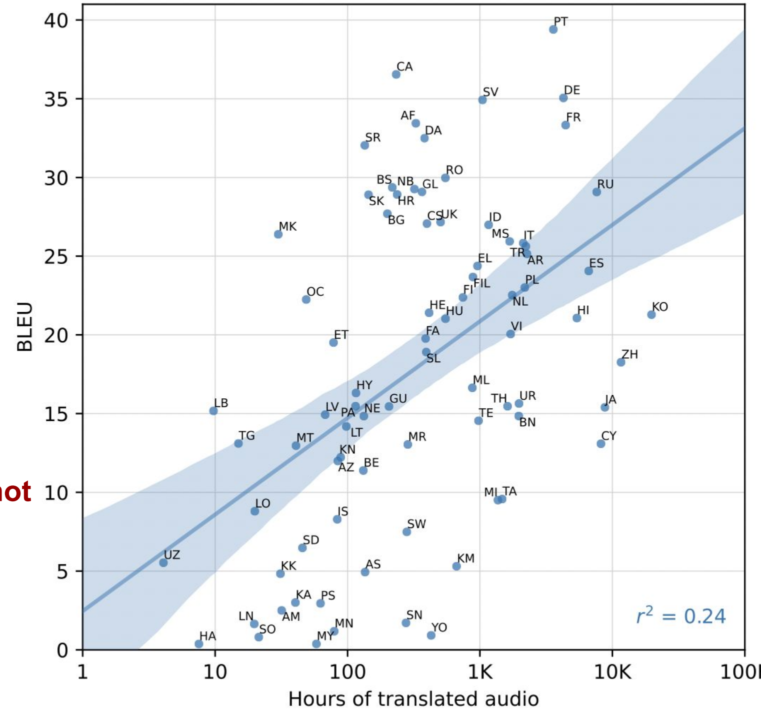
Modality Encoders: Whisper

- **Multitask training (680K hours)**
  - Speech transcription (multilingual), Speech translation (X → En) and Language Identification





Amount of pretraining data is very much predictive for zero-shot ASR performance.

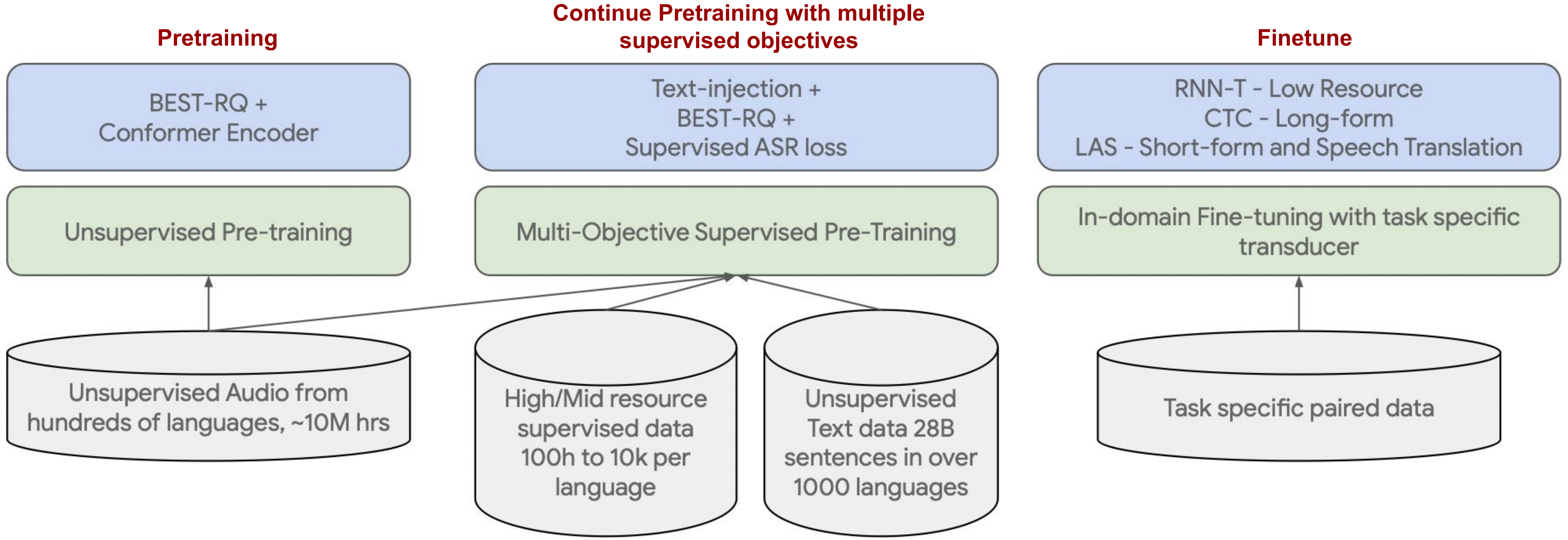


Moderate predictive for zero-shot Translation performance.

Robust speech recognition via large-scale weak supervision.(Radford, Alec, et al. ICML 2023)

Modality Encoders: USM

- **Universal Speech Model (USM)**
  - Speech: 12M hours for 300 languages YT unlabeled data, 429k hours, 51 languages, unlabeled public datasets
  - Text: 2B sentences, 1140 languages
  - Paired Data (Speech, Text):
    - 100k hours, ~100 languages
    - 100k hours en-US pseudo-labeled
    - 10k hours multi-domain en public data



Google usm: Scaling automatic speech recognition beyond 100 languages. (Zhang, Yu, et al. arXiv 2023)

Whisper vs USM

Overall performance comparison: ASR Tasks

Task	Multilingual Long-form ASR				Multidomain en-US	Multilingual ASR	
Dataset	YouTube			CORAAL	SpeechStew	FLEURS	
Langauges	en-US	18	73	en-US	en-US	62	102
Prior Work (single model)							
Whisper-longform	17.7	27.8	-	23.9	12.8		
Whisper-shortform†	-	-	-	13.2‡	11.5	36.6	-
Our Work (single model)							
USM-LAS	14.4	19.0	29.8	**11.2**	**10.5**	**12.5**	-
USM-CTC	**13.7**	**18.7**	**26.7**	12.1	10.8	15.5	-

Google usm: Scaling automatic speech recognition beyond 100 languages. (Zhang, Yu, et al. arXiv 2023)

Whisper vs USM

Low-resource Setting: Standard Arabic vs Dialects and Domain (ASR)

Bilingual (EN, AR) Conformer ASR

Dataset dom./dial.	Models	Zero-Shot	N-Shot (2hrs)	SOTA
MGB2 *Broadcast/MSA*	W.S	46.70	36.8	
	W.M	33.00	-	O: **11.4**
	W.Lv2	26.20	18.8	S:11.9
	USM	15.70	N/A	
MGB3 *Broadcast/EGY*	W.S	83.20	77.5	
	W.M	65.90	-	O: **21.4**
	W.Lv2	55.60	44.6	S: 26.70
	USM	22.10	N/A	
MGB5 *Broadcast/MOR*	W.S	135.20	114.6	
	W.M	116.90	-	O: **44.1**
	W.Lv2	89.40	85.5	S:49.20
	USM	51.20	N/A	

Standard Arabic → High-resource

EGY dialectal Arabic → Mid-resource

MOR dialectal Arabic → Low-resource

Whisper models: W

Dataset dom./dial.	Models	Zero-Shot	N-Shot (2hrs)	SOTA
QASR.CS *Broadcast/Mixed*	W.S	63.60	-	
	W.M	48.90	-	O: **23.4**
	W.Lv2	37.90	31.2^{+}	S: 24.90
	USM	27.80	N/A	
DACS *Broadcast /MSA-EGY*	W.S	61.90	-	
	W.M	48.70	-	O: 15.9
	W.Lv2	34.20	30.4^{+}	S: 21.3
	USM	**14.30**	N/A	
ESCWA.CS *Meeting/Mixed*	W.S	101.50	-	
	W.M	69.30	-	O: 49.8
	W.Lv2	60.00	53.6^{+}	S:48.00
	USM	**45.70**	N/A	
CallHome *Telephony/EGY*	W.S	155.90	152.9	
	W.M	113.70	-	O: **45.8***
	W.Lv2	78.70	64.6	S: 50.90
	USM	54.20	N/A	

LAraBench: Benchmarking Arabic AI with Large Language Models. (Abdelali, Ahmed, et al. EACL 2024)

MLLM (Gemini) vs Whisper and USM

MM + LLMs improve results over Foundation Models?

	Task	Metric	Gemini Pro	Gemini Nano-1	Whisper (OpenAI, 2023; Radford et al., 2023)	USM (Zhang et al., 2023)
Automatic Speech Recognition	**YouTube** (en-us)	WER (↓)	**4.9%**	5.5%	6.5% (v3)	6.2%
	Multilingual Librispeech (en-us) (Pratap et al., 2020)	WER (↓)	**4.8%**	5.9%	6.2% (v2)	7.0 %
	FLEURS (62 lang) (Conneau et al., 2023)	WER (↓)	**7.6%**	14.2%	17.6% (v3)	11.8%
	VoxPopuli (14 lang) (Wang et al., 2021)	WER (↓)	**9.1%**	9.5%	15.9% (v2)	13.4%
Automatic Speech Translation	**CoVoST 2** (21 lang) (Wang et al., 2020)	BLEU (↑)	**40.1**	35.4	29.1 (v2)	30.7

Significant Improvement wrt FM in multilingual space

Gemini: a family of highly capable multimodal models. (Team, Gemini, et al., arXiv 2023)

Modality Generator

Latent Diffusion Models (LDMs)

Visual Modality

- StableDiffusion (Image) (Rombach et al., 2022)
- Zeroscope (Video) (Cerspense et al., 2023)

Speech/Audio

- **AudioLDM**
- **AudioLDM2 (speech, music, sound effect)** (Liu et al., 2023 a, Liu et al., 2023 b)
- **VALL-E**

MM-LLMs: Recent advances in multimodal large language models (Zhang, Duzhen, et al. arXiv 2024)

Sample Pretraining Datasets

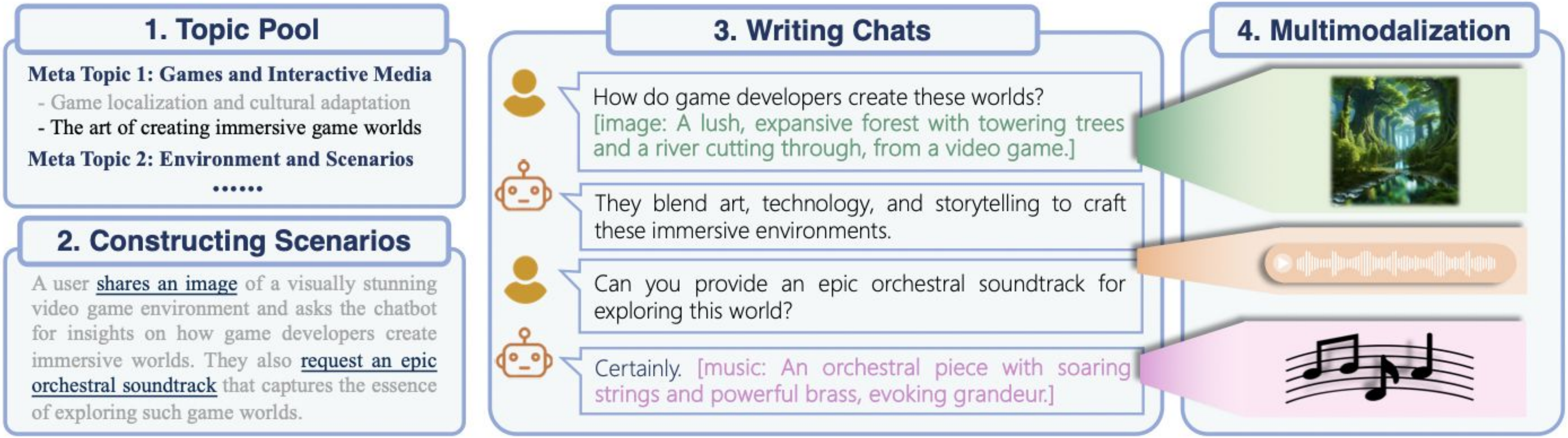
- **Speech, Speech-Text**
  - GigaSpeech, AMI, Tedlium, Multilingual Librispeech (m), CommonVoice (m), QASR (dialectal Ar), AISHELL (Chinese), CSJ (Japanese), Microsoft Speech Corpus (Indian Languages) among many others
- **Music, Music-Text**
  - Youtube-Music-1M, MusicGen-Synthesis
- **Image, Image-Text**
  - LAION-COCO, MMC4-core-ff, JourneyDB (synthetic data - Midjourney), LAION-2B, LAION-Aesthetics ..

Translation for Low-resource languages!

Challenge

Instruction Data

- **AnyInstruct Dataset**
  - Generate text-based conversation with added multimodal element
  - Use the modality description for Text to Modality generation



AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. (Zhan, Jun, et al. arXiv 2024)

Challenge

Instruction Data

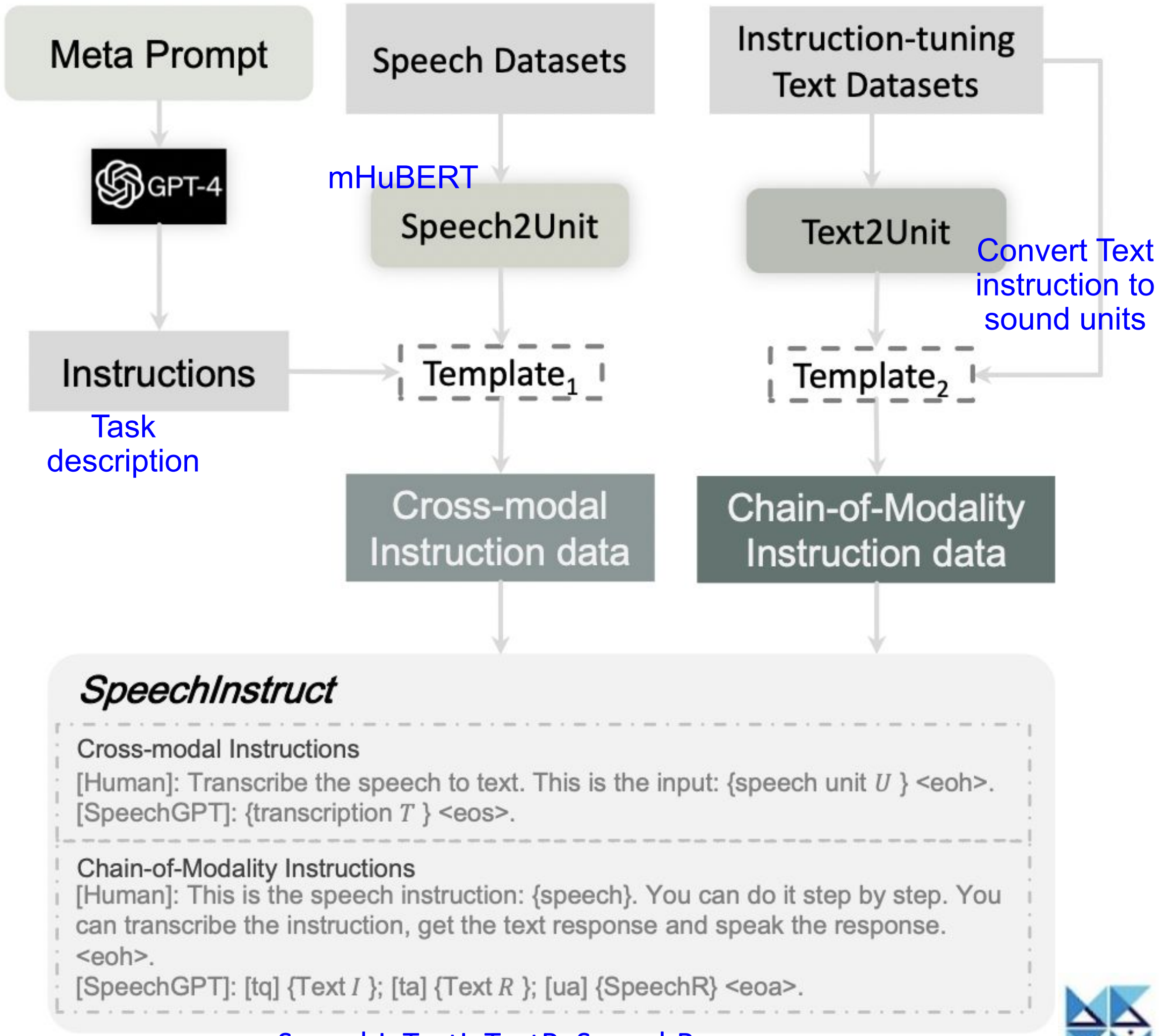
- **Modality-switching Instruction (MosIT) Dataset**
  - Modalities: Image, Audio, Video, Text
  - Supports complex cross-modal understanding, reasoning along with multimodal content generation.
  - Role Design: Human and Machine for various scenarios [more than 100 topics] → GPT4 generate conversations (Multi-turn: 3-7 turns, interleaved with different modalities) (Automatic)
  - For multimodal, best matched content is added from external resources (Manual, Automatic)

Next-gpt: Any-to-any multimodal llm (WU, Shengqiong, et al. arXiv 2023)

Challenge

Instruction Data

- **SpeechInstruct Dataset**
  - Speech-Text cross-modal dataset
  - **Cross-Modal Instruction**
    - Discrete Unit - Text Paired data collection
    - Task description generation
    - Instruction Formatting (<task_description, <units>, <transcription>)
  - **Chain-of-Modality Instruction**
    - Speech instruction generation
    - Instruction formatting

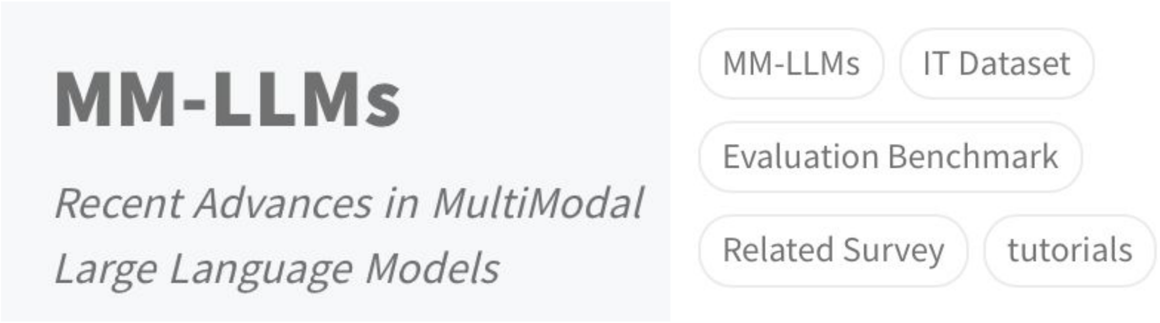


SpeechI, TextI, TextR, SpeechR

SpeechGpt: Empowering large language models with intrinsic cross-modal conversational abilities. (Zhang, Dong, et al. arXiv 2023)

Some Resource

- **Surveys**
  - MM-LLMs: Recent advances in multimodal large language models (Zhang, Duzhen, et al. arXiv 2024)
  - Large Multimodal Agents: A Survey. (Xie, Junlin, et al. arXiv 2024)
  - Multimodal large language models: A survey. (Wu, Jiayang, et al. BigData 2023)
  - A survey on multimodal large language models.(Yin, Shukang, et al. arXiv 2023)
- **https://mm-llms.github.io**



QA

Thank You

https://llm-low-resource-lang.github.io/