LREC 2026 Palma

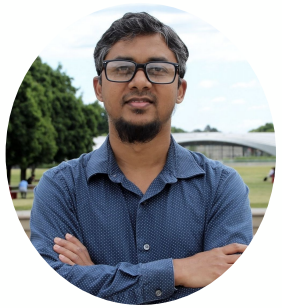
QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

YORK UNIVERSITÉ UNIVERSITY

Multilingual and Multimodal LLMs in the Wild: Building for Low-Resource Languages

Firoj Alam
Senior Scientist, QCRI

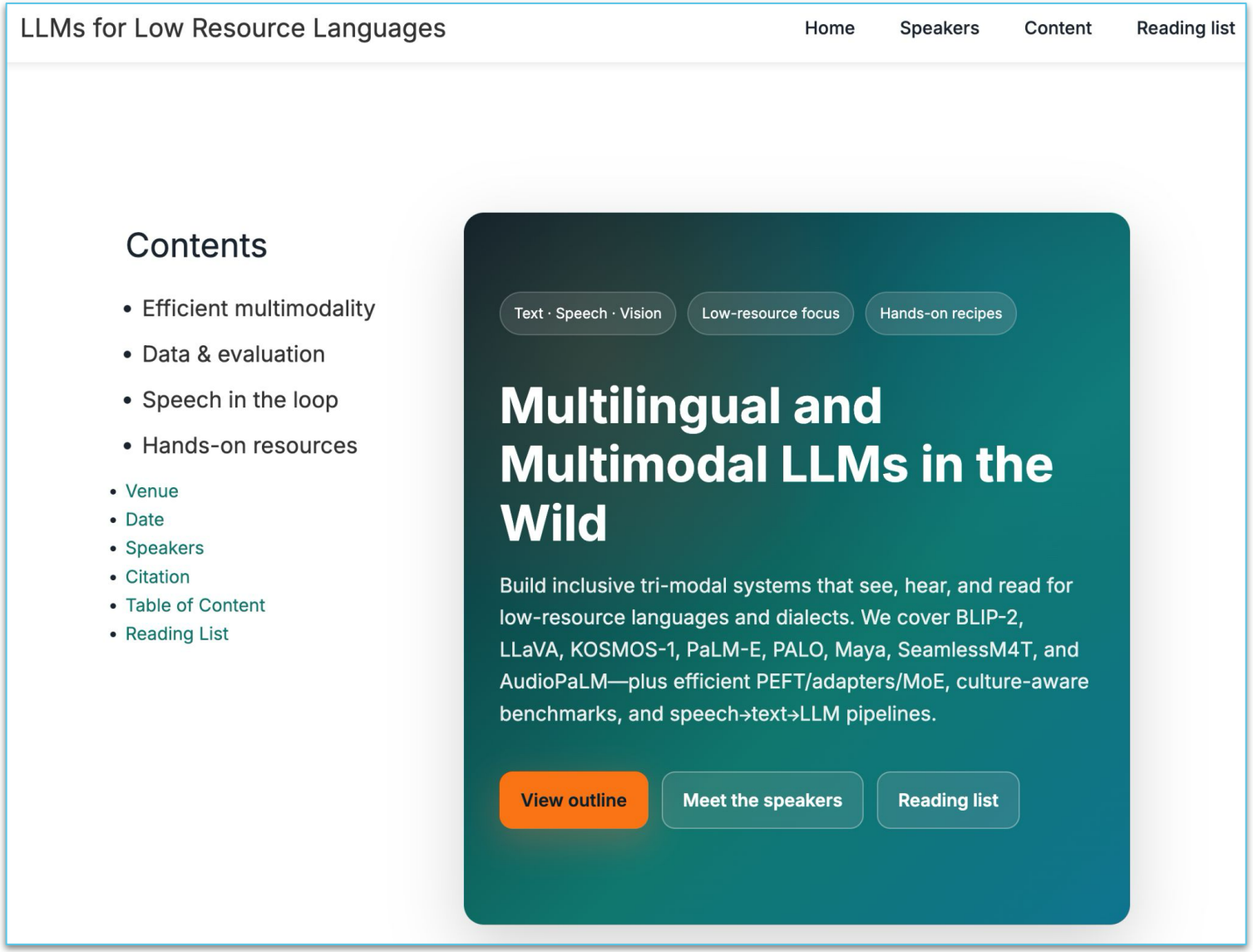
Shammur Chowdhury
Scientist, QCRI



Enamul Hoque Prince
Associate Professor,
York University

LREC 2026, 12th May, 2026

Tutorial Resources



https://mm-llms-in-the-wild.github.io

Tutorial Outlines

Introduction
(low-resource and challenges)

Multilingual and Multimodal Models
(Capabilities for low-resource languages)

Multilingual & Multimodal Resource Development
(Low-cost pipeline, training, and benchmarking datasets)

Architectures and Efficient Training
(Adapter-based and parameter-efficient methods)

Speech-centric LLMs
(State-of-the-art models and curated resources)

Evaluation, Benchmarking Tools, Analysis
(Evolution of Evaluation metrics, benchmarking tools and resources)

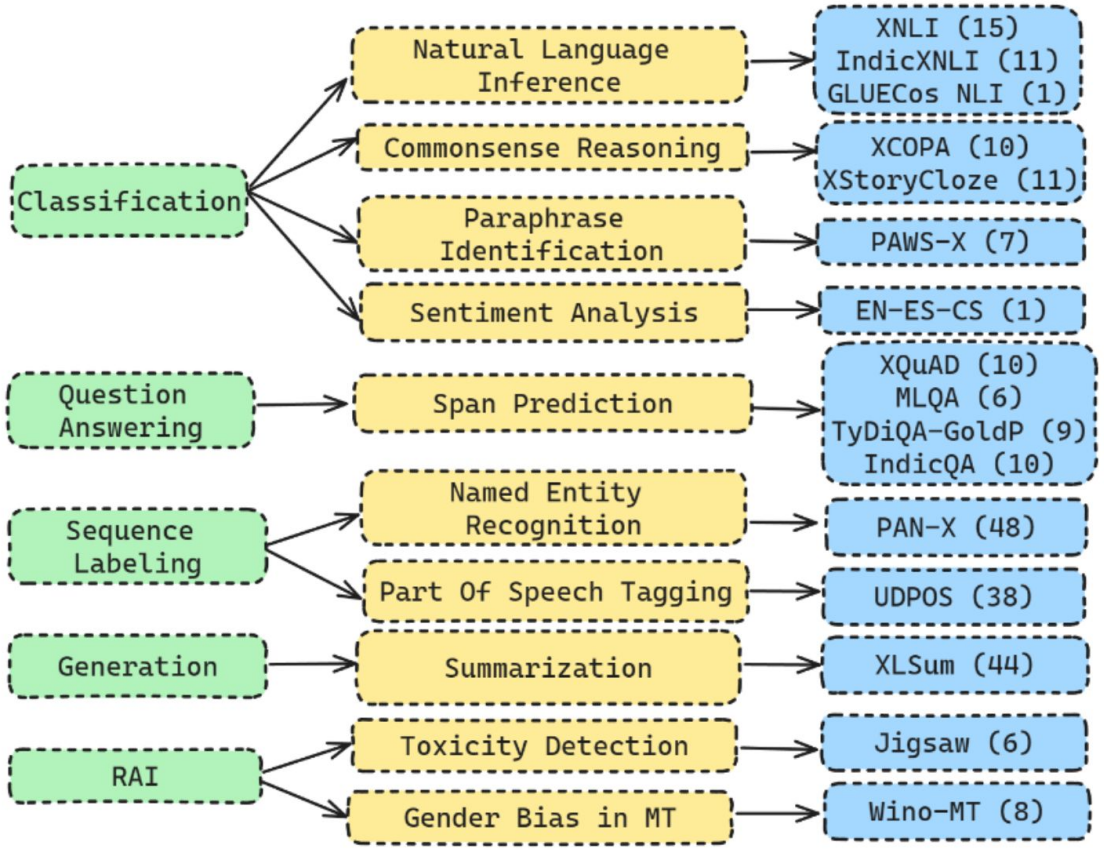
Tutorial Schedule

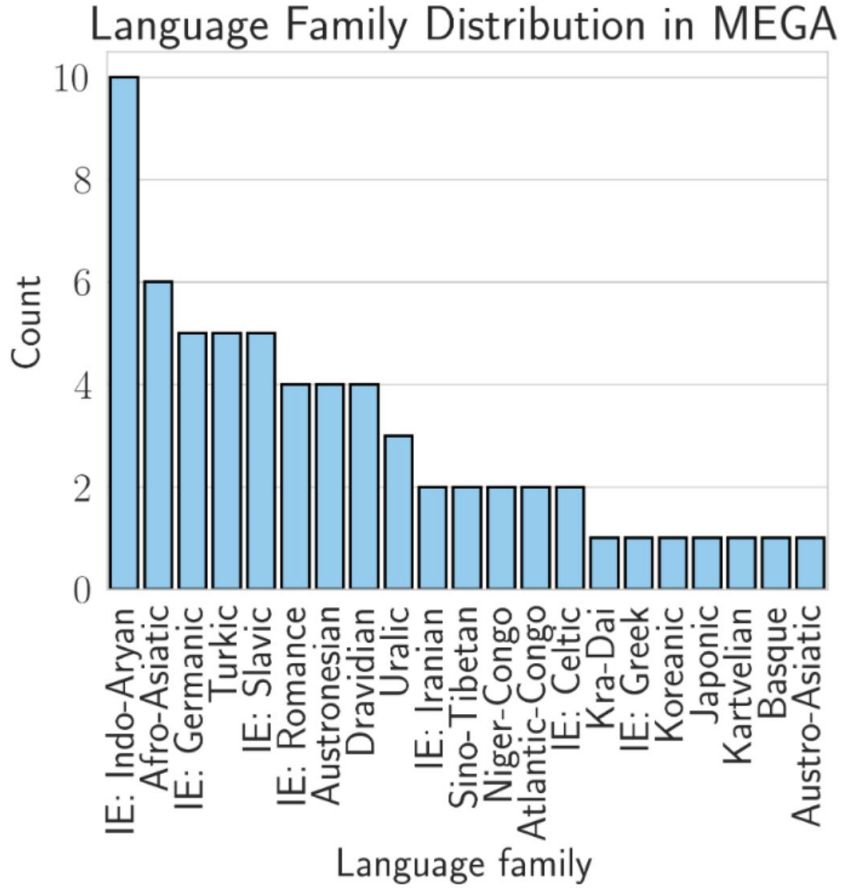
Content	Speaker	Time (Minutes)	Accumulative Time (minutes)	Start	End
Introduction	Firoj Alam	25	25	14:00	14:25
Multilingual & Multimodal Resource Development - 1	Firoj Alam	35	60	14:25	15:00
QA		5	65	15:00	15:05
Multilingual & Multimodal Resource Development - 2	Enamul Hoque Prince	30	95	15:05	15:35
QA		5	100	15:35	15:40
Architectures and Efficient Training	Firoj Alam	20	120	15:40	16:00
Coffee break		30	150	16:00	16:30
Speech-centric LLMs	Shammur Absar Chowdhury	45	195	16:30	17:15
QA		5	200	17:15	17:20
Tools and Resources for Evaluation and Analysis	Firoj Alam	25	225	17:20	17:45
QA		10	235	17:45	17:55
Closing remarks	Firoj Alam	5	240	17:55	18:00

Multilingual LLMs

Multilingual Models & Capabilities

MEGA evaluates models on standard NLP benchmarks, covering 16 NLP datasets across 70 typologically diverse languages

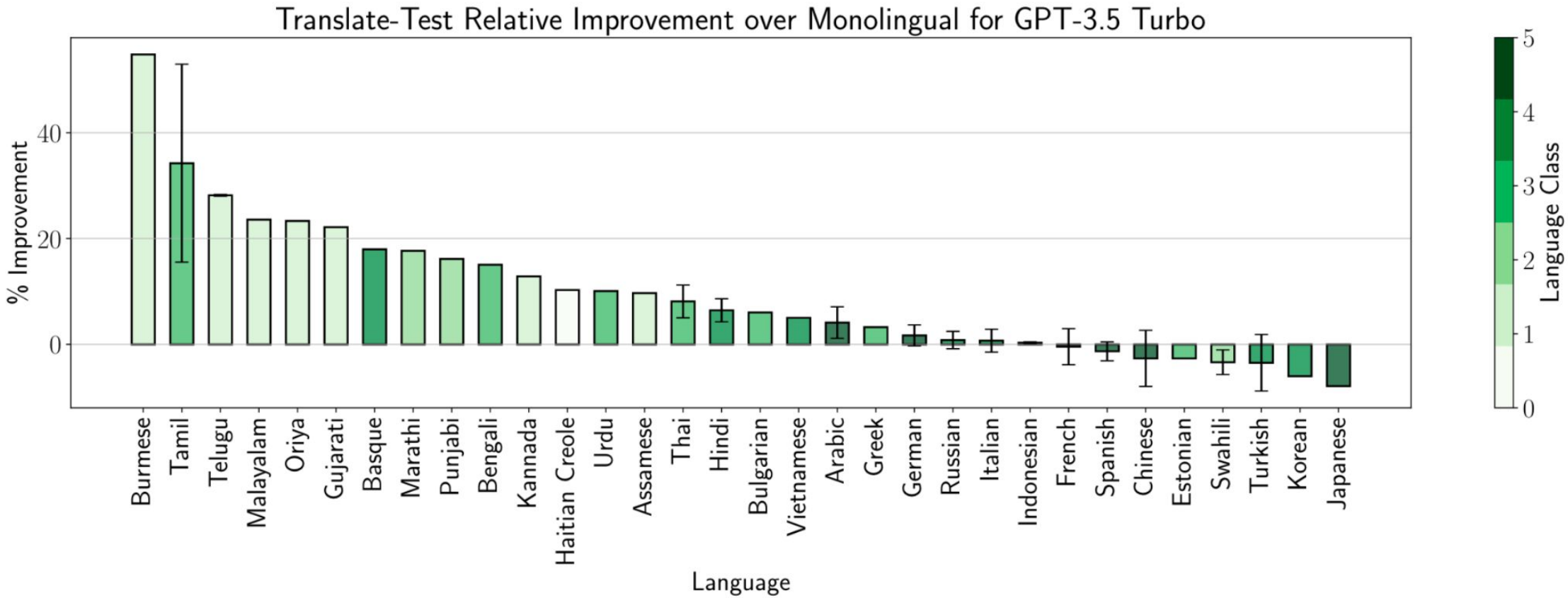




Ahuja et al., MEGA: Multilingual Evaluation of Generative AI, 2024

Multilingual Models & Capabilities

- LLMs still vastly underperform on (especially low-resource) non-English languages



Ahuja et al., MEGA: Multilingual Evaluation of Generative AI, 2024

Multilingual Models & Capabilities

21 datasets covering 8 different common NLP application tasks

- ChatGPT fails to generalize to low-resource and extremely low-resource languages (e.g., Marathi, Sundanese, and Buginese).
- It shows more weakness in inductive reasoning than in deductive or abductive reasoning
- It suffers from the hallucination problem

Bang et al., A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity, IJCNLP-AACL 2024

Multilingual Models & Capabilities

37 diverse languages, characterizing high-, medium-, low-, and extremely low-resource languages

- ChatGPT’s zero-shot learning performance is generally worse than the SOTA
- The importance of task-specific models is higher
- ChatGPT’s performance is generally better for English than for other languages, especially for higher-level tasks that require more complex reasoning abilities

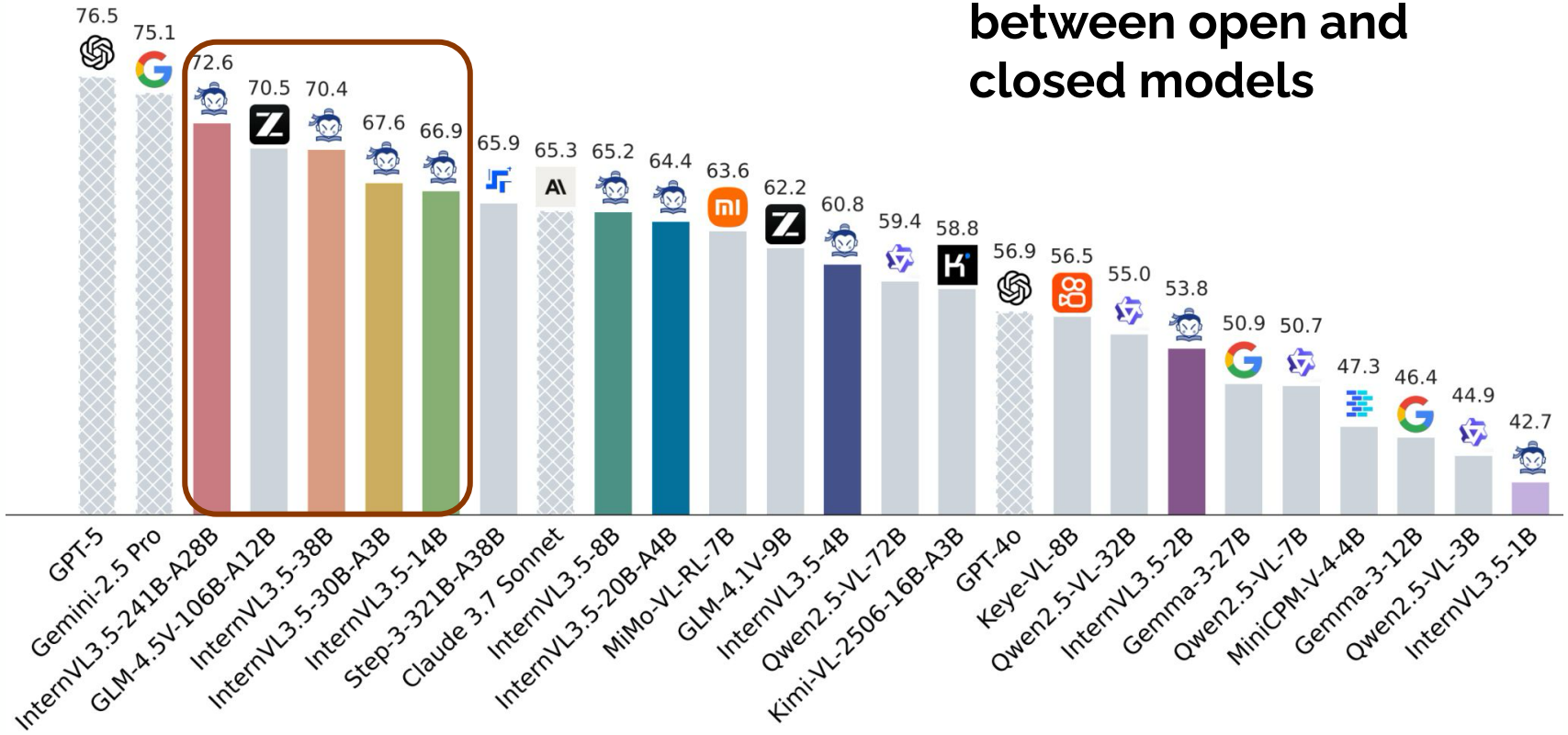
Bang et al., A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity, IJCNLP-AACL 2024

Multimodal LLMs

Models & Capabilities

InternVL3.5

Gaps are closing between open and closed models



InternVL3.5 is SOTA or near-SOTA among open-source MLLMs, especially on general multimodal, reasoning, text, GUI, and embodied tasks.

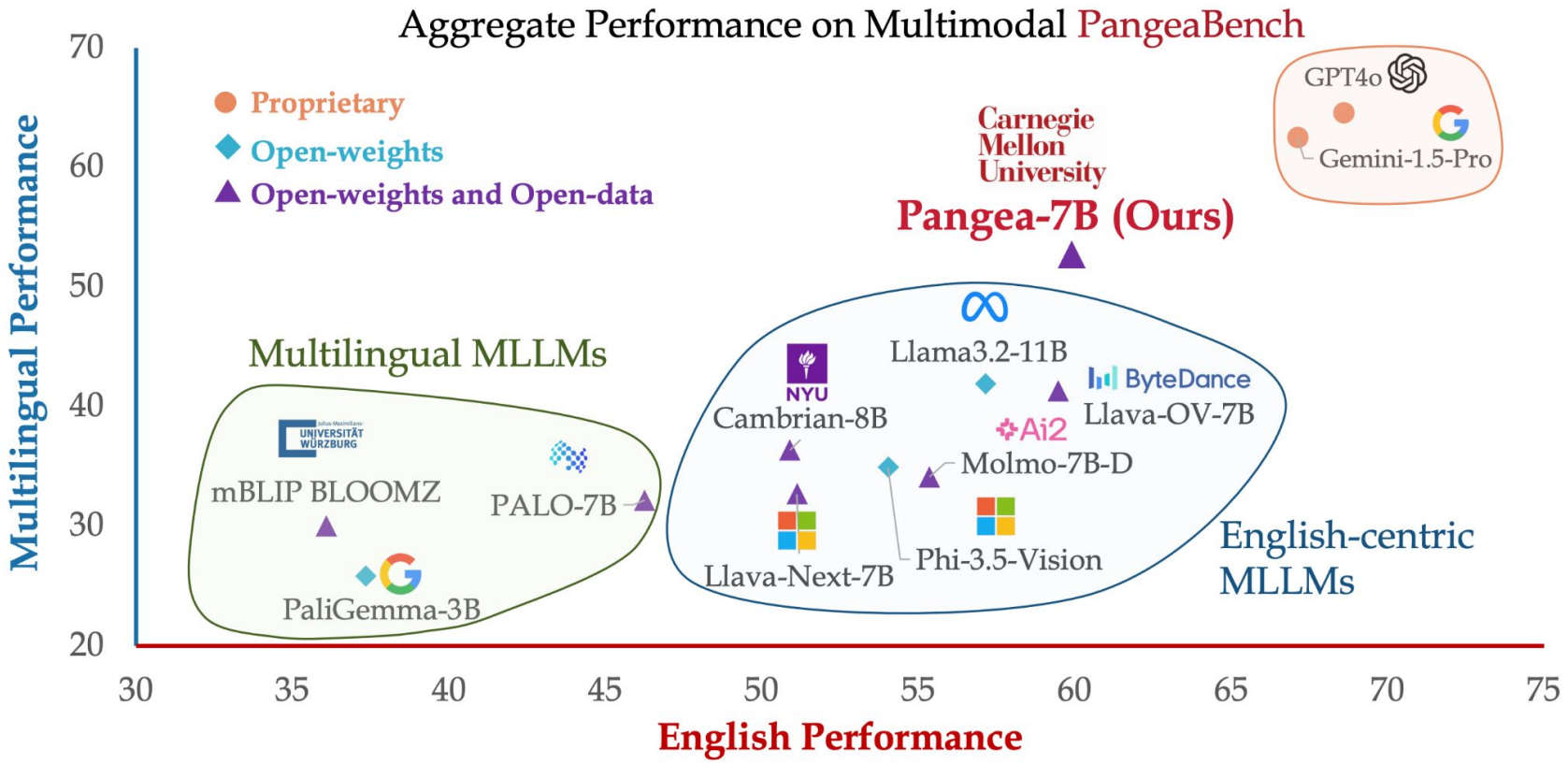
It substantially improves over InternVL3, especially in reasoning

Average scores on a set of multimodal general, reasoning, text, and agentic benchmarks: MMBench v1.1 (en), MMStar, BLINK, HallusionBench, AI2D, OCRBench, MMVet, MME-RealWorld (en), MVBench, VideoMME, MMMU, MathVista, MathVision, MathVerse, DynaMath, WeMath, LogicVista, MATH500, AIME24, AIME25, GPQA, MMLU-Pro, GAOKAO, IFEval, SGP-Bench, VSI-Bench, ERQA, SpaCE-10, and OmniSpatial.

Wang et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, 2025

Models & Capabilities

Pangea



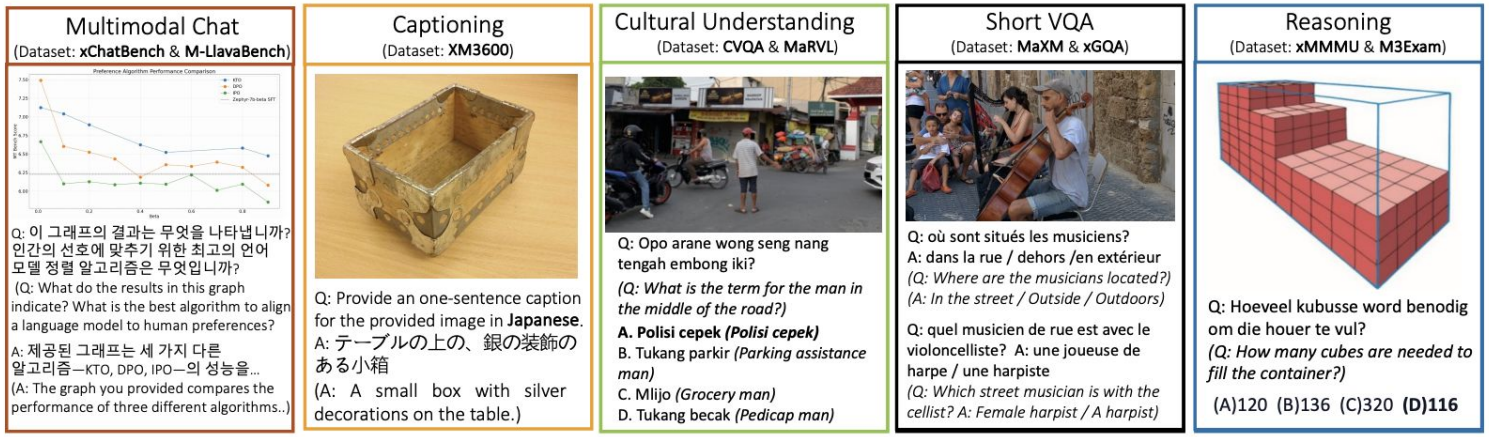
Aggregate performance of various multimodal LLMs on PANGEABENCH

Yueet al. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages, 2025

Models & Capabilities

PangeaBench

- 5 multimodal and 3 text tasks
- 14 datasets
- 47 languages across multimodal chat, captioning, VQA, cultural reasoning, multimodal reasoning, QA, translation, and math.



Category	Tasks	Datasets	Forms	Size	Languages	Metric
Multimodal	Multimodal Chat	xChatBench	Long	400	zh,en,hi,id,ja,rw,ko,es	LLM-as-Judge
		M-LlavaBench	Long	600	ar,bn,zh,fr,hi,ja,ru,es,ur,en	LLM-as-Judge
	Captioning	XM100	Long	3.6K	36 languages	ROUGE-L
	Cultural Understanding	CVQA	MC	21K	en,zh,ko,mn,ja,id,jv,min,su	Accuracy
		MaRVL	Short	6K	id,sw,ta,tr,zh	Accuracy
	Multilingual VQA	xGQA	Short	77K	en,de,pt,ru,id,bn,ko,zh	Accuracy
		MaXM	MC	2K	hi,th,zh,fr,en,iw,ro	Accuracy
	Reasoning (Multi-subject)	xMMMU	Short/MC	3K	en,ar,fr,hi,id,ja,pt	Accuracy
		M3Exam	MC	3K	en,zh,it,pt,vi,th,af	Accuracy
Text-only	QA	TyDiQA	Short	5.1K	ar,ru,bn,te,fi,sw,ko,id,en	Accuracy
	Translation	FLORES-Sub	Long	18K	ar,en,fr,de,hi,id,iw,ja,pt,ro,tr	ChrF
	Reasoning (Multi-subject, Commonsense, Math)	MMMLU	MC	197K	ar,bn,de,es,fr,hi,id,it,ja,ko,pt,sw,yo,zh	Accuracy
		XStoryCloze	MC	21K	en,ar,es,eu,hi,id,my,ru,sw,te,zh	Accuracy
		MGSM	Open	3K	bn,de,en,es,fr,ja,ru,sw,te,th,zh	Accuracy

Yueet al. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages, 2025

Models & Capabilities

Pangea

Models	AVG (all)		Multimodal Chat				Cultural Understanding			
			xChatBench		M-LlavaBench		CVQA		MaRVL	
	en	mul	en	mul	en	mul	en	mul	en	mul
Gemini-1.5-Pro	67.1	62.5	67.0	54.4	103.4	106.6	75.9	75.7	76.4	72.0
GPT4o	68.6	64.6	71.0	64.4	104.6	100.4	79.1	79.4	81.4	82.1
Llava-1.5-7B	45.4	28.4	28.5	11.8	66.1	40.8	48.9	36.5	56.2	53.7
Llava-Next-7B	51.1	32.7	40.5	18.9	78.9	50.7	55.7	42.6	62.8	50.9
Phi-3.5-Vision	54.0	35.0	38.5	13.2	70.8	58.0	56.3	42.3	72.1	56.5
Cambrian-8B	50.9	36.4	27.5	11.3	78.4	61.8	59.7	47.5	75.4	61.8
Llava-OV-7B	59.5	41.3	51.0	28.5	89.7	55.3	65.2	53.7	72.7	57.5
Molmo-7B-D	55.4	34.1	**49.5**	21.1	**95.9**	13.8	59.4	48.3	65.3	54.9
Llama3.2-11B	57.2	41.9	49.0	27.8	93.9	58.2	**70.2**	**61.4**	64.5	58.1
PaliGemma-3B	37.3	25.8	6.0	3.5	32.1	31.9	52.9	42.9	56.5	52.2
PALO-7B	46.3	32.2	27.0	11.8	68.9	71.2	50.9	39.2	63.3	54.2
mBLIP mT0-XL	35.1	29.8	2.5	0.5	32.7	28.2	40.5	37.5	67.3	66.7
mBLIP BLOOMZ	36.1	30.0	4.0	1.6	43.5	41.0	44.9	36.9	62.3	58.6
PANGEA-7B (Ours)	**59.9**	**52.8**	46.0	**35.8**	84.2	89.5	64.4	57.2	**87.0**	79.0
Δ over SoTA Open	+0.4	+10.9	-3.5	+7.3	-11.7	+18.3	-5.8	-4.2	+11.6	+12.3

Models	Captioning		Short VQA				Multi-subject Reasoning			
	XM100		xGQA		MaXM		xMMMU		M3Exam	
	en	mul	en	mul	en	mul	en	mul	en	mul
Gemini-1.5-Pro	27.6	19.1	54.2	48.7	56.4	63.5	65.8	57.7	77.4	64.7
GPT4o	27.7	19.1	55.8	51.0	60.7	65.4	69.1	58.3	68.0	61.0
Llava-1.5-7B	28.6	1.1	62.0	30.6	49.8	20.4	36.2	31.5	32.3	29
Llava-Next-7B	29.3	9.4	**64.8**	37.8	54.9	21.4	36.7	34.3	36.5	28.4
Phi-3.5-Vision	30.2	5.2	64.7	38.4	**55.3**	25.0	42.6	38.8	55.8	37.2
Cambrian-8B	20.6	9.9	64.6	39.8	**55.3**	28.7	41.8	33.2	34.7	33.4
Llava-OV-7B	**30.6**	7.0	64.4	48.2	54.9	34.8	46.3	41.0	60.4	**45.8**
Molmo-7B-D	22.1	9.1	51.5	43.0	52.9	37.5	44.5	40.4	57.1	39.1
Llama3.2-11B	27.6	4.5	55.6	45.4	**55.3**	43.9	**46.5**	41.4	51.8	36.6
PaliGemma-3B	18.7	0.8	59.7	30.5	47.9	19.9	26.3	25.2	36.0	25.6
PALO-7B	30.4	0.8	60.5	37.8	51.4	16.3	33.1	30.5	30.8	27.8
mBLIP mT0-XL	31.9	3.1	44.2	39.9	44.7	36.8	29.3	30.4	22.8	25
mBLIP BLOOMZ	22.5	10.3	43.3	36.9	44.7	24.8	29.2	30.8	30.3	29.5
PANGEA-7B (Ours)	30.4	**14.2**	64.7	**60.2**	**55.3**	**53.3**	**45.7**	**43.7**	**61.4**	42.1
Δ over Best Open Model	-0.2	+3.9	-0.1	+12.0	0.0	+9.4	-0.8	+2.3	+1.0	-3.7

PANGEA-7B outperforms existing open-source MLLMs in multilingual and culturally diverse settings

Yueet al. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages, 2025

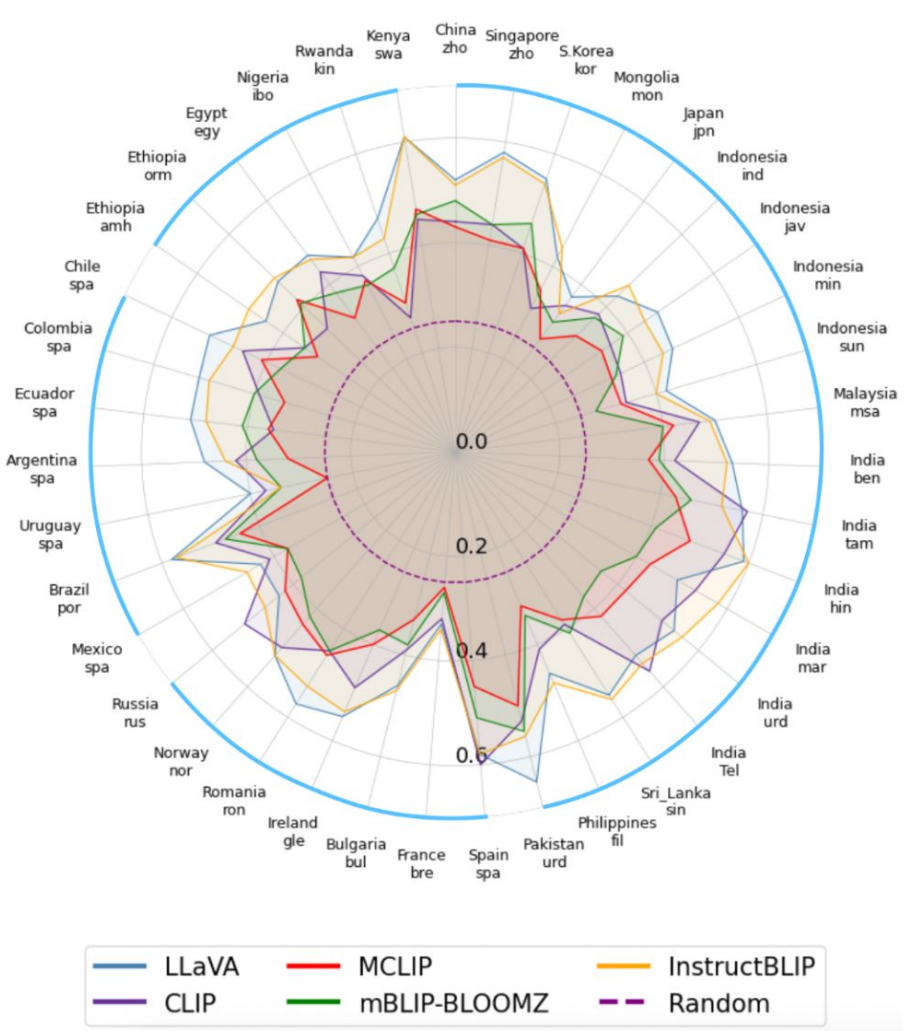
Models & Capabilities

CVQA

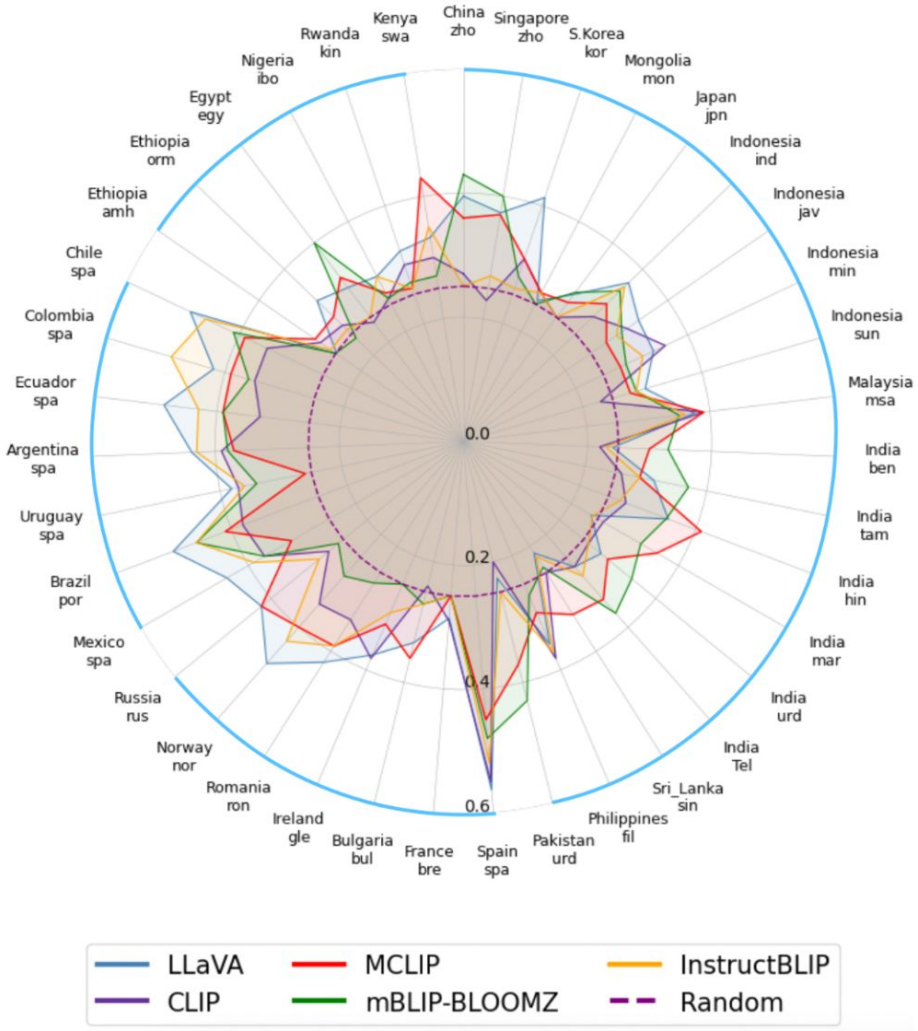
- **CVQA** is challenging benchmark for current MLLMs, especially open-source models.
- **Open models usually score below 50%**; LLaVA-1.5-7B reaches about 49.6% with English prompts but drops to 35.5% with local-language prompts.
- **Closed models perform better:** GPT-4o reaches 75.4% with English prompts and 74.3% with local-language prompts.
- **Performance drops sharply for underrepresented languages** such as Breton, Javanese, and Mongolian.
- Food and pop-culture questions are especially difficult because they require local cultural knowledge.
- Removing multiple-choice options reduces performance further, showing weaker real-world cultural understanding.

Romero et al., CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark, 2024

Models & Capabilities



Using English-only question-option pairs



Using local-language question-option pairs

Romero et al., CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark, 2024