LREC 2026 Palma

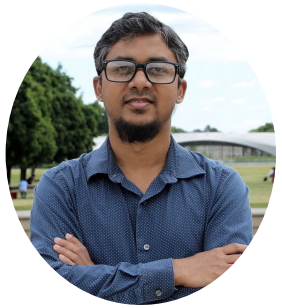
QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

YORK UNIVERSITÉ UNIVERSITY

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

Firoj Alam
Senior Scientist, QCRI

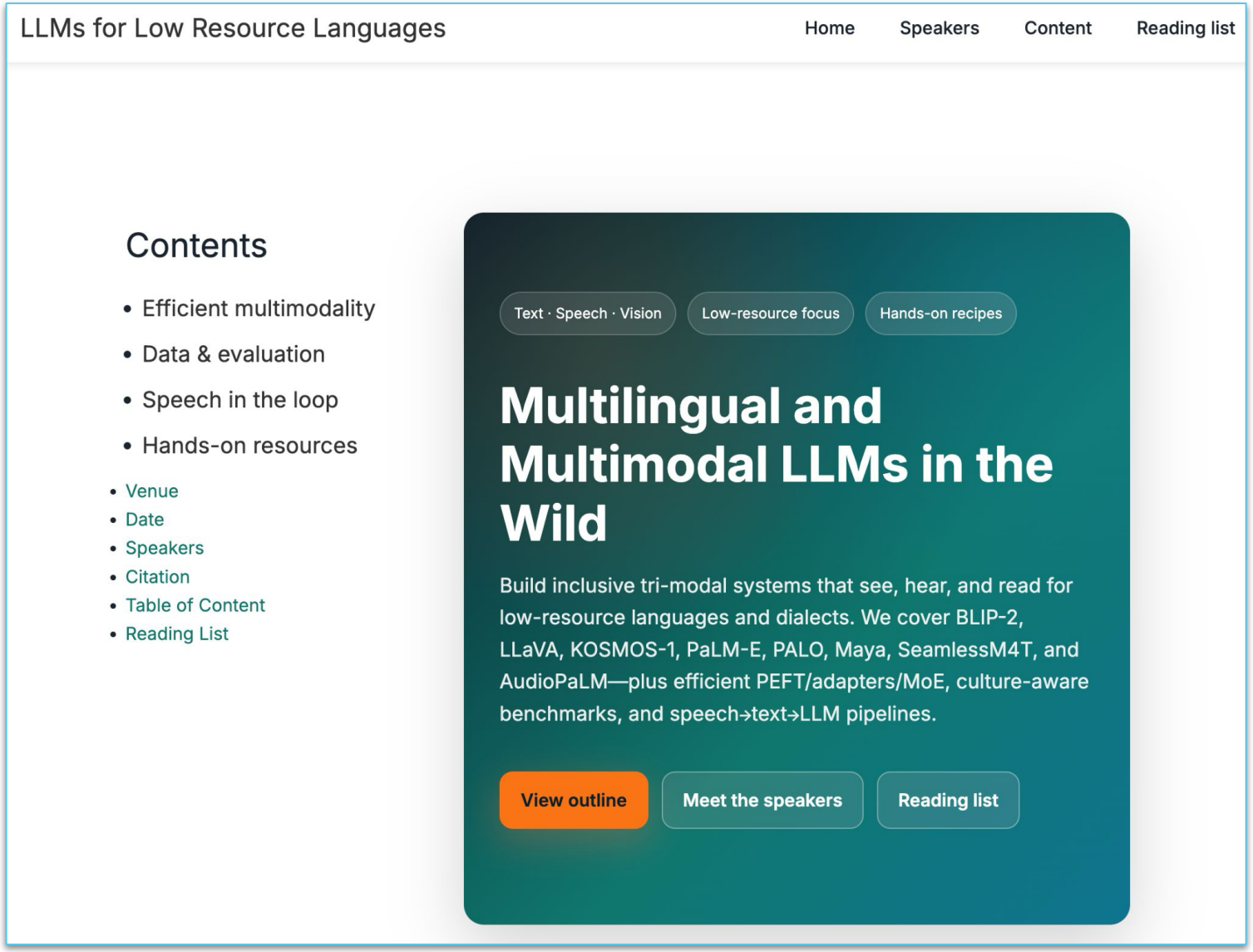
Shammur Chowdhury
Scientist, QCRI



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

Introduction

Introduction

What do we mean by low resource?

Data Scarcity

NLP & Modalities

- Lack of labeled/annotated task-specific datasets.
- Minimal or no raw unlabeled text for pre-training.
- Severe scarcity in Speech & Multimodal pairs (Image-Text/Video-Text).
- High development costs for high-quality manual data.

Compute Limits

Hardware Constraints

- Insufficient GPU/TPU hours for LLM/MLLM full pre-training.
- Memory bottlenecks for large model parameter counts.
- Need for parameter-efficient fine-tuning (PEFT) methods.

Inference & Dev

Operational Low Resource

- Inference cost.
- Lack of evaluation benchmarks for rare languages/modalities.
- Difficulty in scaling MLLM dataset development pipelines.
- Sparse human feedback for RLHF in new domains.

Introduction

Language-specific challenges

Data Scarcity

Not enough labelled speech-text pairs for training robust models.

Domain Mismatch

Public read speech ≠ conversational, phone, classroom, or medical speech.

Dialect & Speaker Bias

Datasets underrepresented regions, genders, ages, and diverse accents.

Orthography

Multiple spellings, lack of standard scripts, and frequent code-switching.

Evaluation Instability

Small test sets result in unreliable and inconsistent performance metrics.

Cultural Mismatch

Benchmarks may test tasks that are irrelevant for local user needs.

Understanding these barriers is critical for developing inclusive and effective multilingual LLMs.

Introduction

In the wild?

Noisy perception

Blurred signs, low-light photos, old scans, overlapping speakers, low-bandwidth audio.

Language variation

Dialect, code-switching, non-standard spelling, morphology, scripts, and transliteration.

Cultural grounding

Food, dress, local landmarks, social practices, regional signs, and community-specific context.

Deployment constraints

Small GPUs, mobile latency, privacy, missing labels, fragile OCR/ASR pipelines.

Models must handle messy real-world language and perception: typos, informal spelling, code-switching, dialects, mixed scripts, OCR errors, blurred or low-light images, noisy video frames, overlapping speakers, accented speech, and low-bandwidth audio.

Introduction

Why do we need multilingual and multimodal LLMs?

Real-world environments are inherently multimodal.

Utilization of diverse channels leads to ***better*** knowledge acquisition.

DIVERSE SENSORY CHANNELS

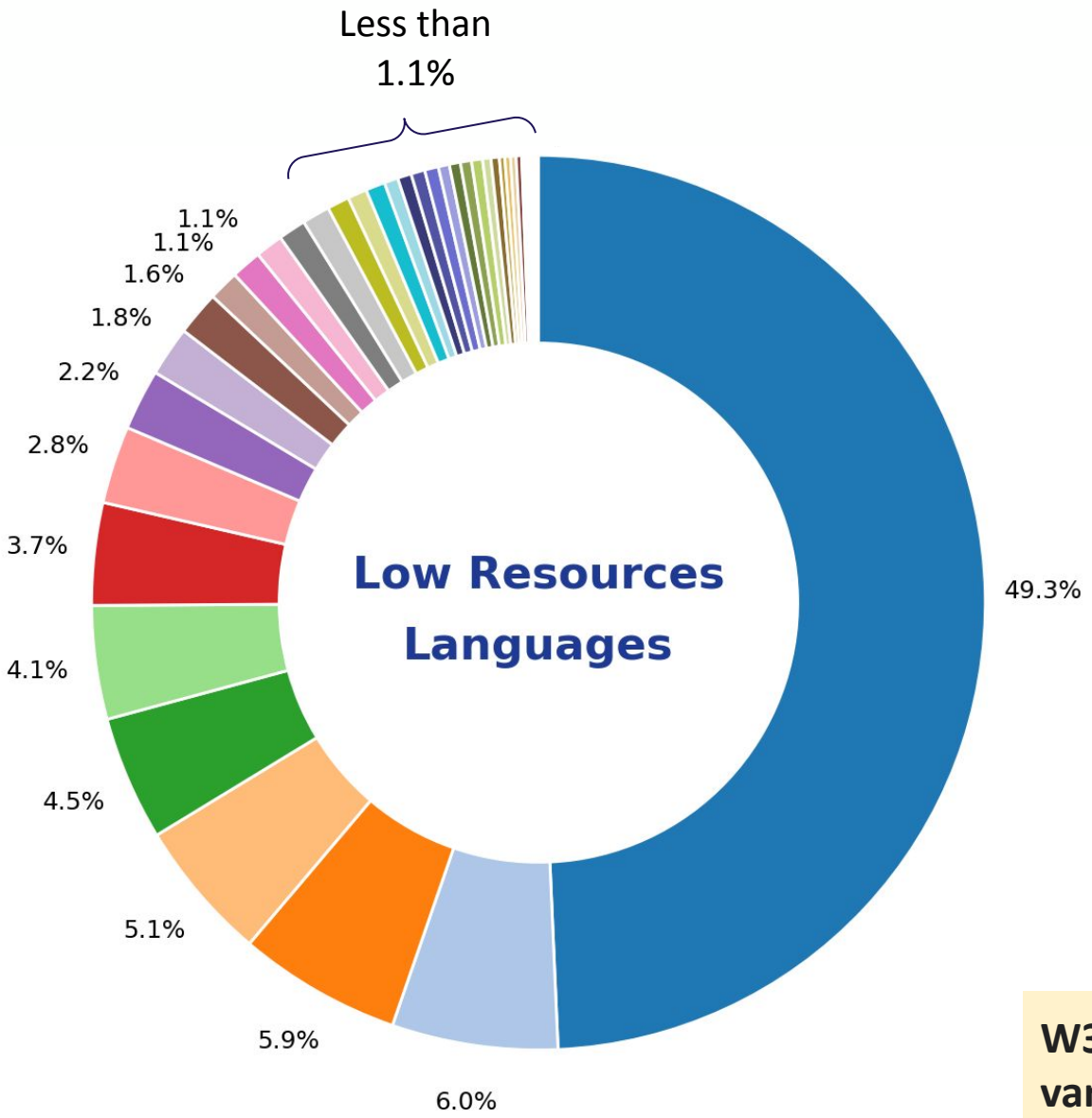
Introduction

Low Resources Languages

- Approximately ~7,000 languages
- Majority of the internet content are in English
- Mostly categorized as lack of
  - labeled/annotated datasets
  - unlabelled datasets

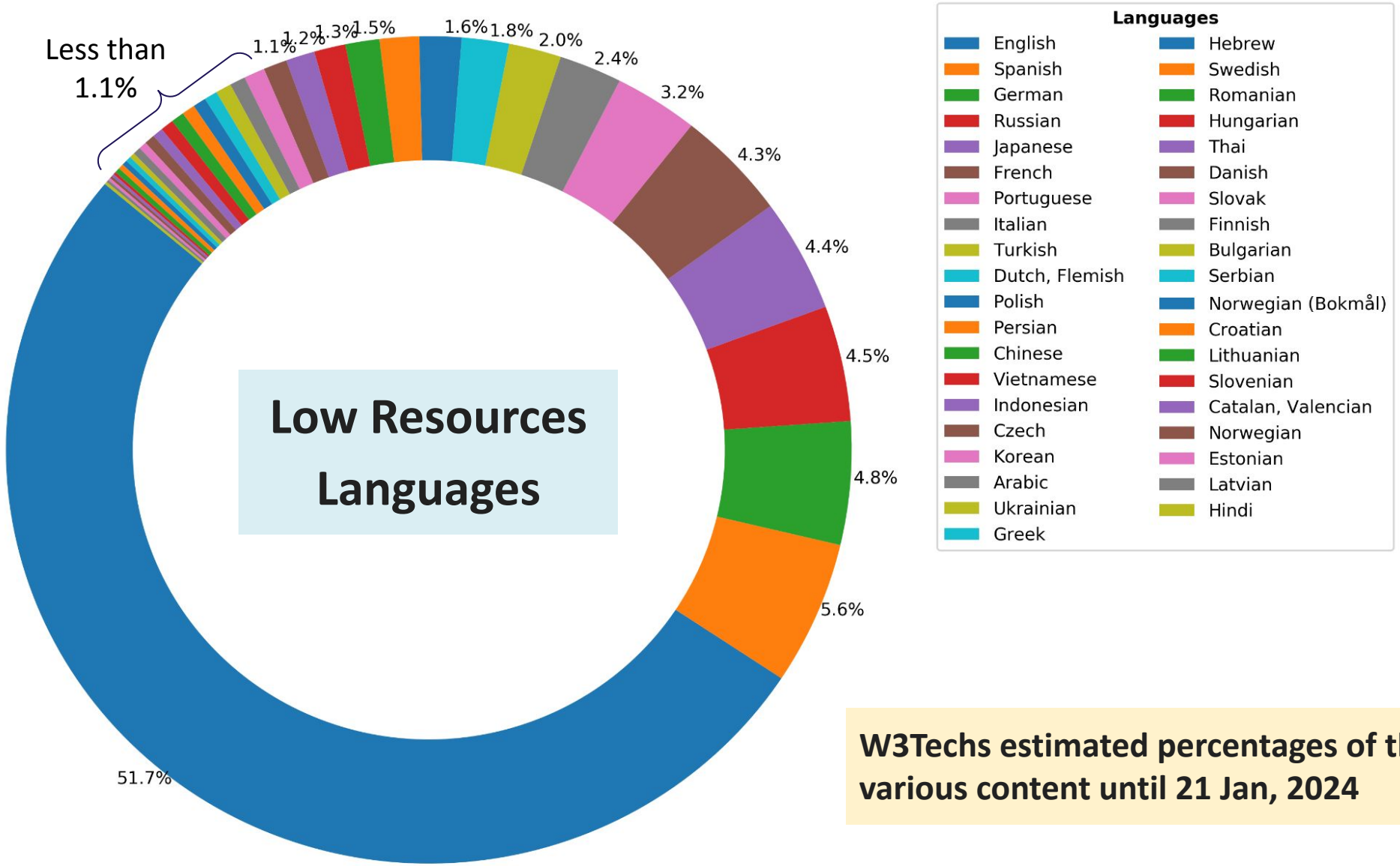
Image: DALL-E

Introduction



W3Techs estimated percentages of the various content until 6 May, 2026

Introduction



W3Techs estimated percentages of the various content until 21 Jan, 2024

Introduction

Low Resources Languages: Categorization

0- The Left-Behinds (exceptionally limited resources: impossible effort to lift them up in the digital space)

1- The Scraping-Bys (some amount of unlabeled data)

2- The Hopefuls (small set of labeled datasets)

3- The Rising Stars (strong web presence, a thriving cultural community online)

4- The Underdogs (serious amounts of resource, a large amount of unlabeled data, dedicated NLP communities)

5- The Winners (dominant online presence, massive effort to develop resources and technologies)

Joshi et al., The State and Fate of Linguistic Diversity and Inclusion in the NLP World, 2020

Introduction

Low Resources Languages: Categorization

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Number of languages, number of speakers, and percentage of total languages for each language class.

Joshi et al., The State and Fate of Linguistic Diversity and Inclusion in the NLP World, 2020

Introduction

Low Resources Languages: Categorization

High (H, > 1%)

Language	Code	Pop. (M)	CC Size (%)	Cat.
English	en	1,452	45.8786	H
Russian	ru	258	5.9692	H
German	de	134	5.8811	H
Chinese	zh	1,118	4.8747	H
Japanese	jp	125	4.7884	H
French	fr	274	4.7254	H
Spanish	es	548	4.4690	H
Italian	it	68	2.5712	H
Dutch	nl	30	2.0585	H
Polish	pl	45	1.6636	H
Portuguese	pt	257	1.1505	H
Vietnamese	vi	85	1.0299	H

Medium (M, > 0.1%)

Turkish	tr	88	0.8439	M
Indonesian	id	199	0.7991	M
Swedish	sv	13	0.6969	M
Arabic	ar	274	0.6658	M
Persian	fa	130	0.6582	M
Korean	ko	81	0.6498	M
Greek	el	13	0.5870	M
Thai	th	60	0.4143	M
Ukrainian	uk	33	0.3304	M
Bulgarian	bg	8	0.2900	M
Hindi	hi	602	0.1588	M

Low (L, > 0.01%),
Extremely-Low (X, < 0.01%)

Bengali	bn	272	0.0930	L
Tamil	ta	86	0.0446	L
Urdu	ur	231	0.0274	L
Malayalam	ml	36	0.0222	L
Marathi	mr	99	0.0213	L
Telugu	te	95	0.0183	L
Gujarati	gu	62	0.0126	L
Burmese	my	33	0.0126	L
Kannada	kn	64	0.0122	L
Swahili	sw	71	0.0077	X
Punjabi	pa	113	0.0061	X
Kyrgyz	ky	5	0.0049	X
Odia	or	39	0.0044	X
Assamesese	as	15	0.0025	X

Languages, language codes, numbers of speakers (first and second), data ratios in the CommonCrawl corpus and language categories.

Lai et al., ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning, 2023

Introduction

Multilingual Challenges in Multimodal LLMs

English- and Western-centric Foundations

Most MLLM training data and benchmarks remain English-anchored, leaving the world's other languages and cultures under-represented.

(Pangea: Yue et al. 2025; Lupascu et al. 2026)

Translation is Not Localization

Extending datasets via machine translation keeps images Western and concepts English, keeping cultural representation narrow.

(CVQA: Romero et al. 2025; CulturalGround: Nyandwi et al. 2025)

Large & Quantified Low-Resource Gap

Frontier LLMs lose up to **24.3% absolute accuracy** on LR languages; open VLMs sit below 50% on native-language VQA.

(MMLU-ProX Xuan et al. 2025; CVQA: Romero et al. 2025)

Introduction

Multilingual Challenges in Multimodal LLMs

Cultural & Factual Bottlenecks

Hardest failures are factual: long-tail cultural entities (food, customs, architecture) the model has never seen at scale.

(CulturalGround: Nyandwi et al. 2025; ALM-Bench: Vayani et al. 2025; WorldCuisines: Winata et al. 2025)

Dialects & Script Diversity

Arabic dialects, code-switching, and orthographic variations break models. ALM-Bench alone spans 24 distinct scripts.

(JEEM: Kadaoui et al. 2026; ALM-Bench: Altakrori et al. 2026; EverydayMMQA: Alam et al. 2025)

Uneven Modality Coverage

Low-resource research focuses on text+image. Speech and video remain critically under-served for non-English users.

(Lupascu et al. 2025; Typhoon-Audio: Manakul et al. 2025)

Introduction

> *"How do we build multilingual multimodal LLMs that are robust to diverse input conditions, measurable and safe?"*

Evolution

Task-specific architectures | **No more task-specific architectures**

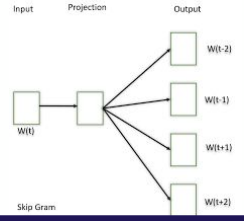
Lots of task specific data | **Small amount of task specific data** | **Zero to few examples**

Neural Language Model (2003)

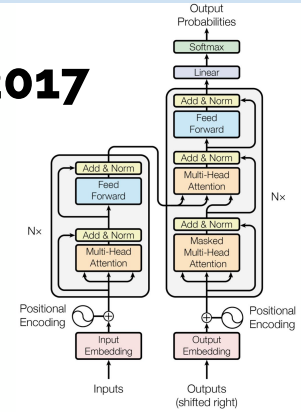
Multitask learning for NLP tasks (2008)

2010

Word Embedding
Deep Learning:
CNN



2017



GPT-3
Few-shot learners

2022

LLM + Prompting

VLLMs
Omni
AudioLLM

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

Bag of words
SVM, RF, Logistic Regression, CRF etc..

2013

2014
Sequence to sequence learning

Transformer
Attention methods

2020

ChatGPT
GPT-3.5

GPT-4
Llama 2,
Falcon,
Palm2

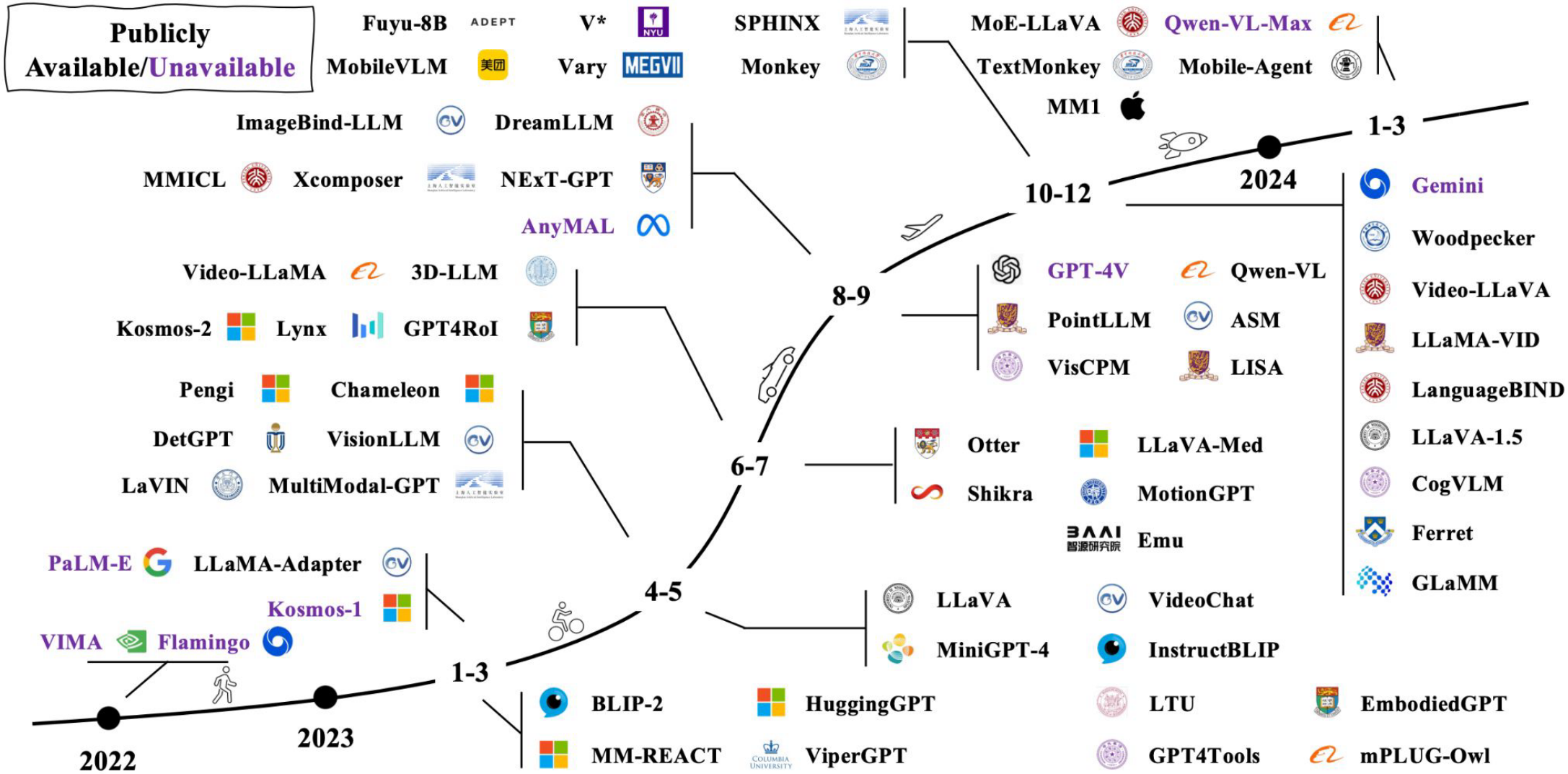
Bloom, Palm, Llama,

Feature Engineering | Network Architectures Engineering | Prompt Engineering

Objective Engineering

Machine learning

Evolution: Recent Models



A Survey on Multimodal Large Language Models (Shukang et al. 2024)

Evolution: Recent Models

A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks (Wang et al. 2024)