LREC 2026 Palma

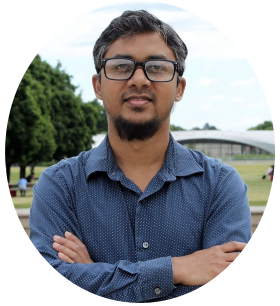
QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

YORK
UNIVERSITÉ
UNIVERSITY

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

Firoj Alam
Senior Scientist, QCRI

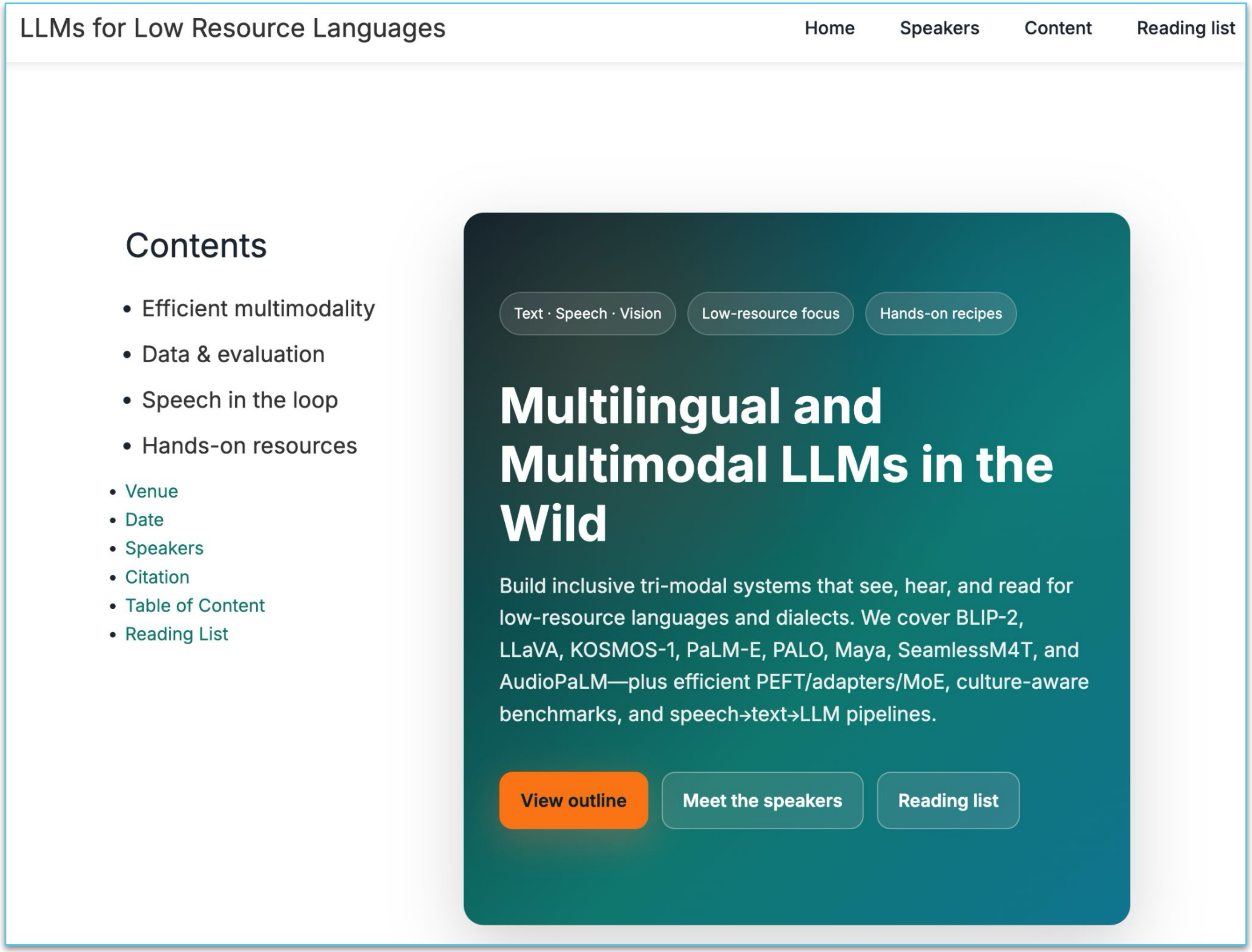
Shammur Chowdhury
Scientist, QCRI



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)

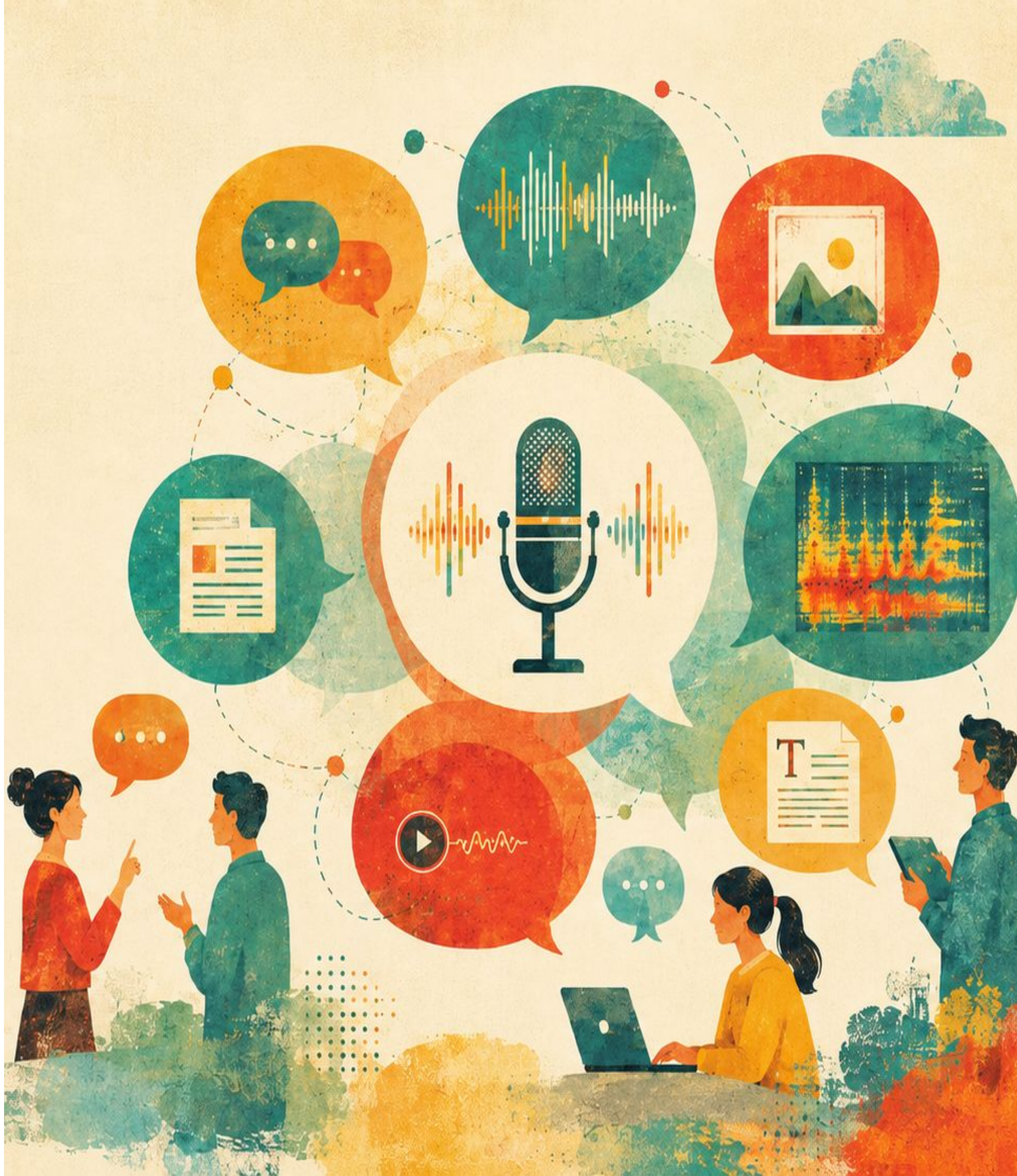
Architectures and Efficient Training
(Adapter-based and parameter-efficient methods)

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

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

Speech-Centric Multimodality

Shammur Chowdhury
Scientist, QCRI



Why Audio LLMs Matter

- Speech is the primary interface for many communities
    - Text-only systems less inclusive: Low literacy, dialectal variation

- Spoken interaction lowers the barrier to using AI.

- Speech carries meaning beyond words

- Text-centric LLMs miss real-world language variation

- AudioLLMs can learn directly from spoken input and enable practical applications

Audio Modalities Types

Speech

Spoken language with semantic content - words, phrases, and meaning conveyed through voice

Non-speech

Crying, laughter, silence, and ambient noise - human sounds beyond words

Sound Events

Door slams, footsteps, machinery, and nature - the sounds of the physical world

Music

Melodies, chords, and instrument timbre - structured auditory art

Challenges

Continuous Waveforms

Unlike text tokens, audio is a dense, unstructured signal requiring aggressive downsampling.

Long Context

A 1-hour meeting is ~200M samples; transformer attention becomes intractable without compression.

Noise & Variation

Real speech is accented, emotional, noisy; models must abstract over immense surface variation.

Multilevel Semantics

Meaning lives at phoneme, word, utterance, and turn-taking levels simultaneously.

Timing-Sensitive

Absolute timing (latency, prosody, rhythm) matters; 'what was said' ≠ 'how it was said'.

Why Speech/Audio Is Low-Resource

Data Scarcity

Only ~1% of the world's 7,000+ languages have **significant speech datasets (labeled)**. **Text corpora are 100-1000x larger** than paired speech-text data.

Annotation Cost

Transcribing speech requires real-time human listening - 1 hour of audio takes 4-10 hours to label. Per-hour costs are 10-50x higher than text.

Domain Fragmentation

Models trained on read speech fail on spontaneous conversation or noisy environments. Each domain needs dedicated data.

Speech models require 10-100x more compute per token of useful supervision than text LLMs

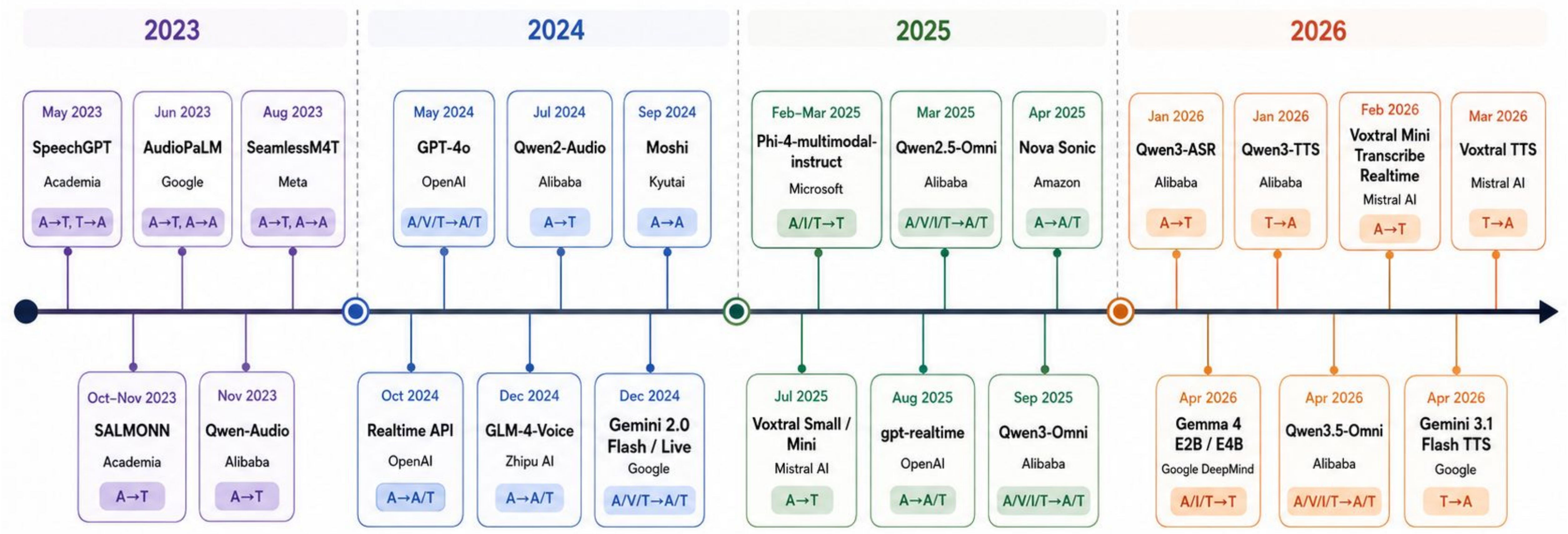
Cascade Pipeline

Pipeline: Traditional speech-to-speech systems chain three separate models in sequence, introducing latency and compounding errors.

Challenges

- High Latency
- Error Propagation
- Loss of Paralinguistic Information

AudioLLM/LALM Timeline



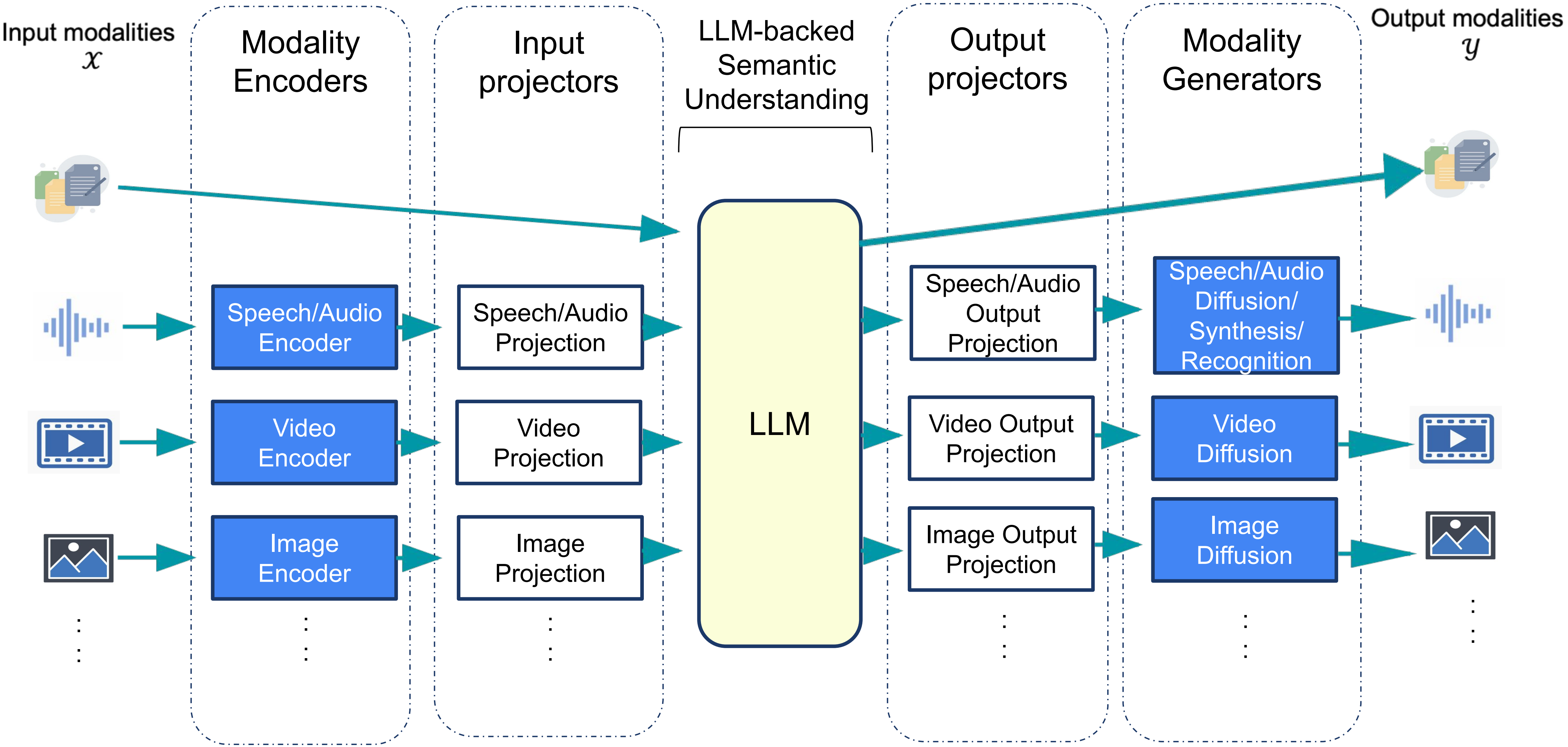
Audio Capable multimodal and real-time speech capabilities

Recent breakthroughs shows Speech I/O is now practical

A→T	speech/audio input to text
A→A/T	speech/audio input to speech or text
T→A	text to speech

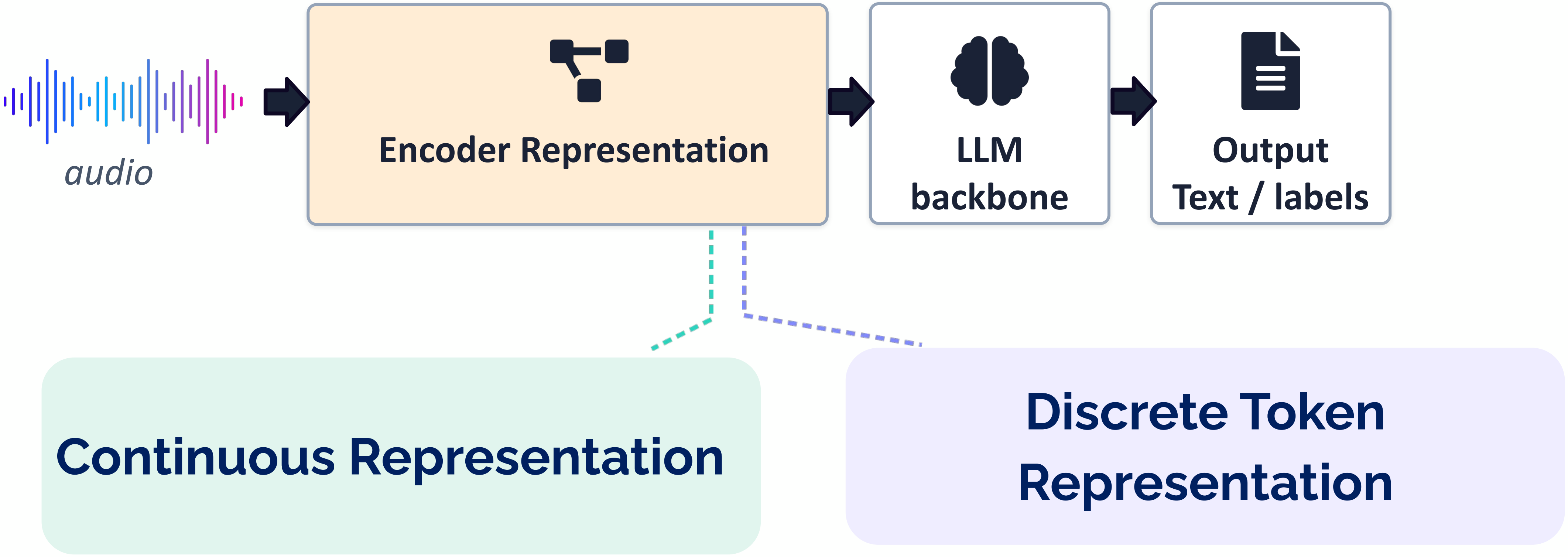
MMLLM Architectures

General Overview

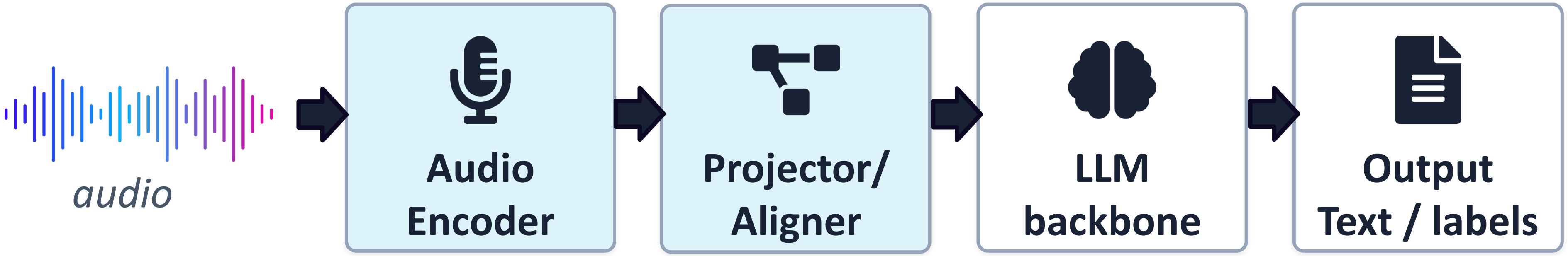


Zhang, Duzhen, et al., MM-LLMs: Recent advances in multimodal large language models, 2024

Simplified Architecture



Continuous Representation



Fast Inference • Preserve Paralinguistic Information • Depends on the Projector Alignment

Speech Encoder Architectures

SSL models

Wav2Vec 2.0, HuBERT, XLS-R, mHuBERT

- Self-supervised learning from raw audio
- No labels required for pretraining
- Discrete units vs. continuous features

Whisper-style

Qwen-Audio, Qwen2-Audio, Qwen2.5 Omni (whisper-L-v3)

- Encoder-Decoder Architecture
- **Trained on ≈5 M h**
- Large weakly-supervised pretrain
- **99-language coverage**
- Off-the-shelf, easy to plug in

AuT (Audio Transformer)

Qwen3-Omni

- Attention-based Encoder-Decoder
- **Trained on ≈20M h audio (v2: 40M)**
- Block-wise streaming
- Stronger audio reps
- ASR and AU tasks
- 80% Chinese and English, 10% ASR other languages, and 10% AU. (8 to 10 lang)

USM-based

Gemma 3n, Gemma 4

- Conformer-based
- Best-RQ: Unsupervised pre-training
- **12M hours** of unlabeled audio and 28B sentences.
- **300+ language**
- Multi-Objective training
- Spacial-Awareness (v2)

Speech-to-LLM Alignment Architectures

Linear Projectors

Simplest form of alignment using a single Linear layer or MLP to map encoder hidden states to LLM embedding space dim.

Used in: Early Audio-LLMs, LLaVA-Audio variants, Low-resource models.

Low Latency

Q-Former / Perceiver

Uses a fixed set of learnable query tokens to cross-attend to encoder features. Significantly reduces sequence length.

Used in: BLIP-2 style adaptations, SALMONN.

Context Efficient

Convolutional Downsamplers

1D Strided Convolutions to shrink the temporal resolution of audio features before LLM injection.

Used in: Qwen2-Audio, Gemini (variants).

Local Modeling

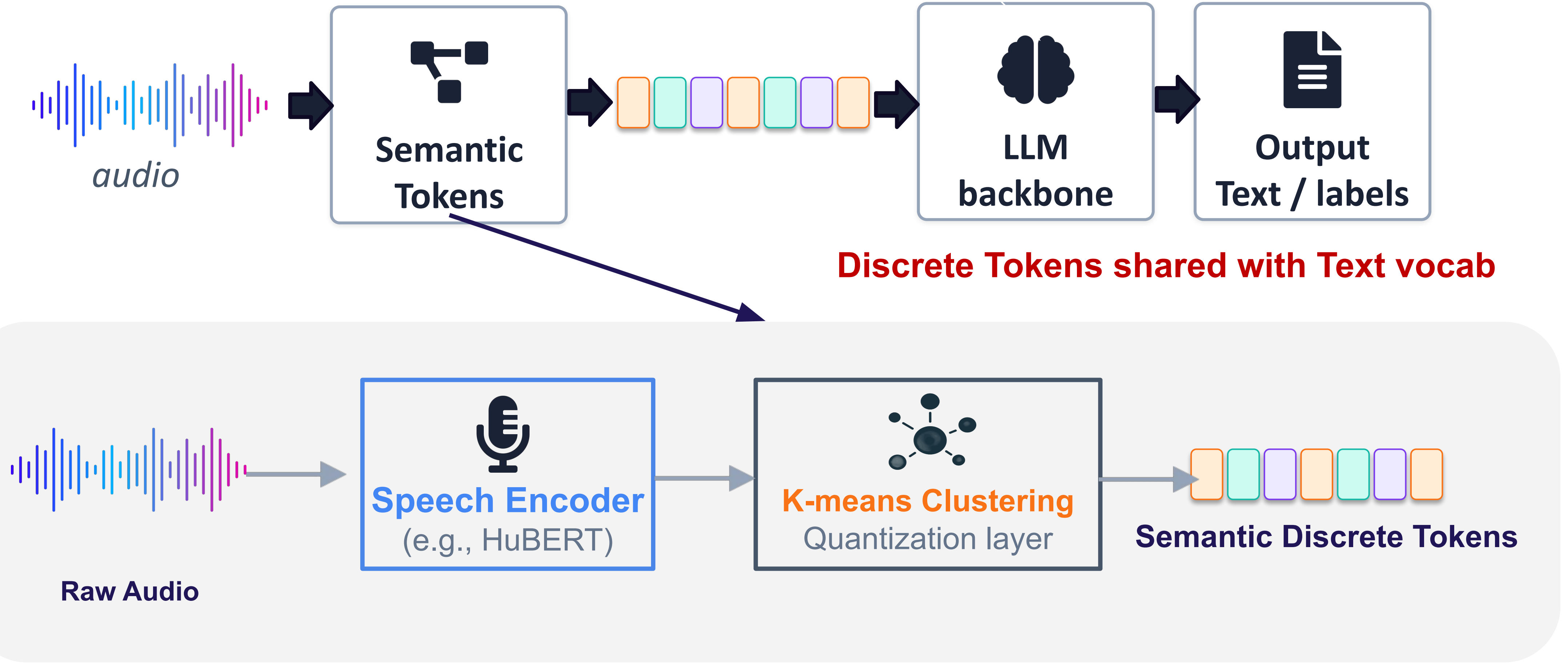
Pooling

Abstracts acoustic features into high-level semantic units via attention-based pooling or clustering.

Used in: LTU.

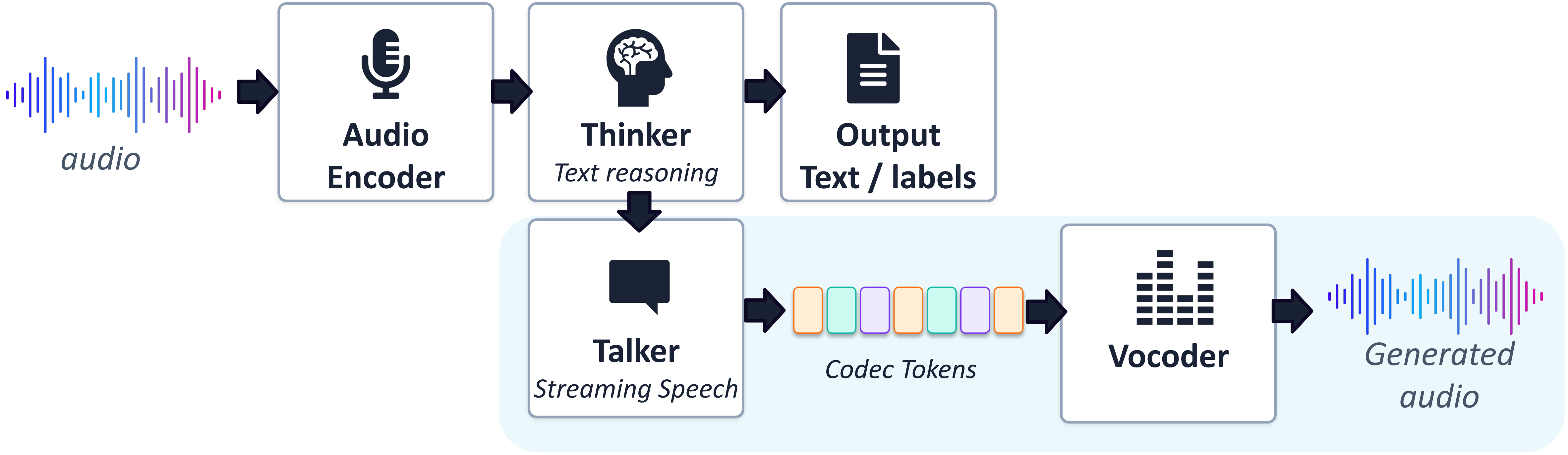
Semantic Rich

Discrete Audio Tokens



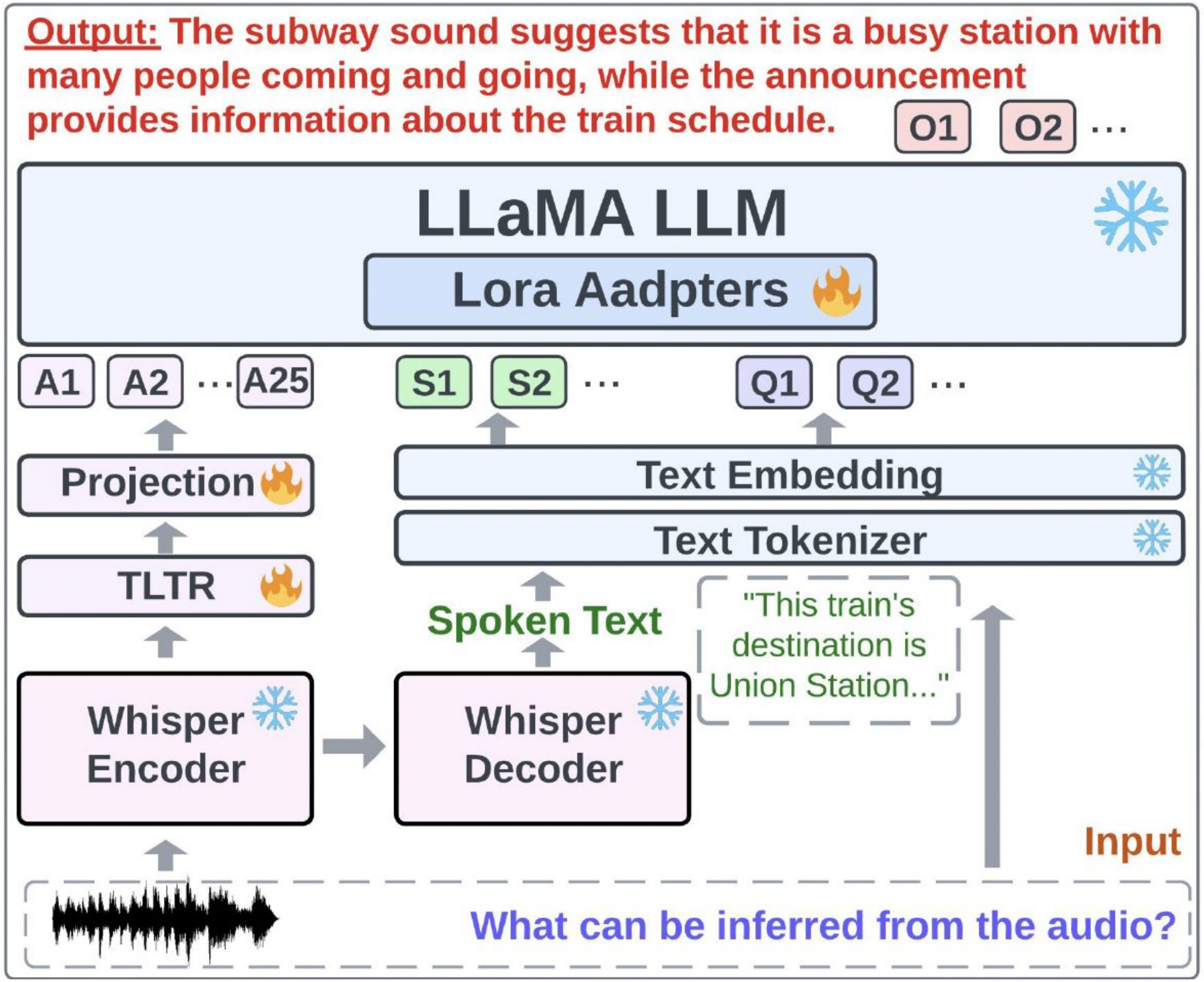
• Codec Quality Bottleneck • Higher Token Overhead

Omni-Model (Thinker-Talker)



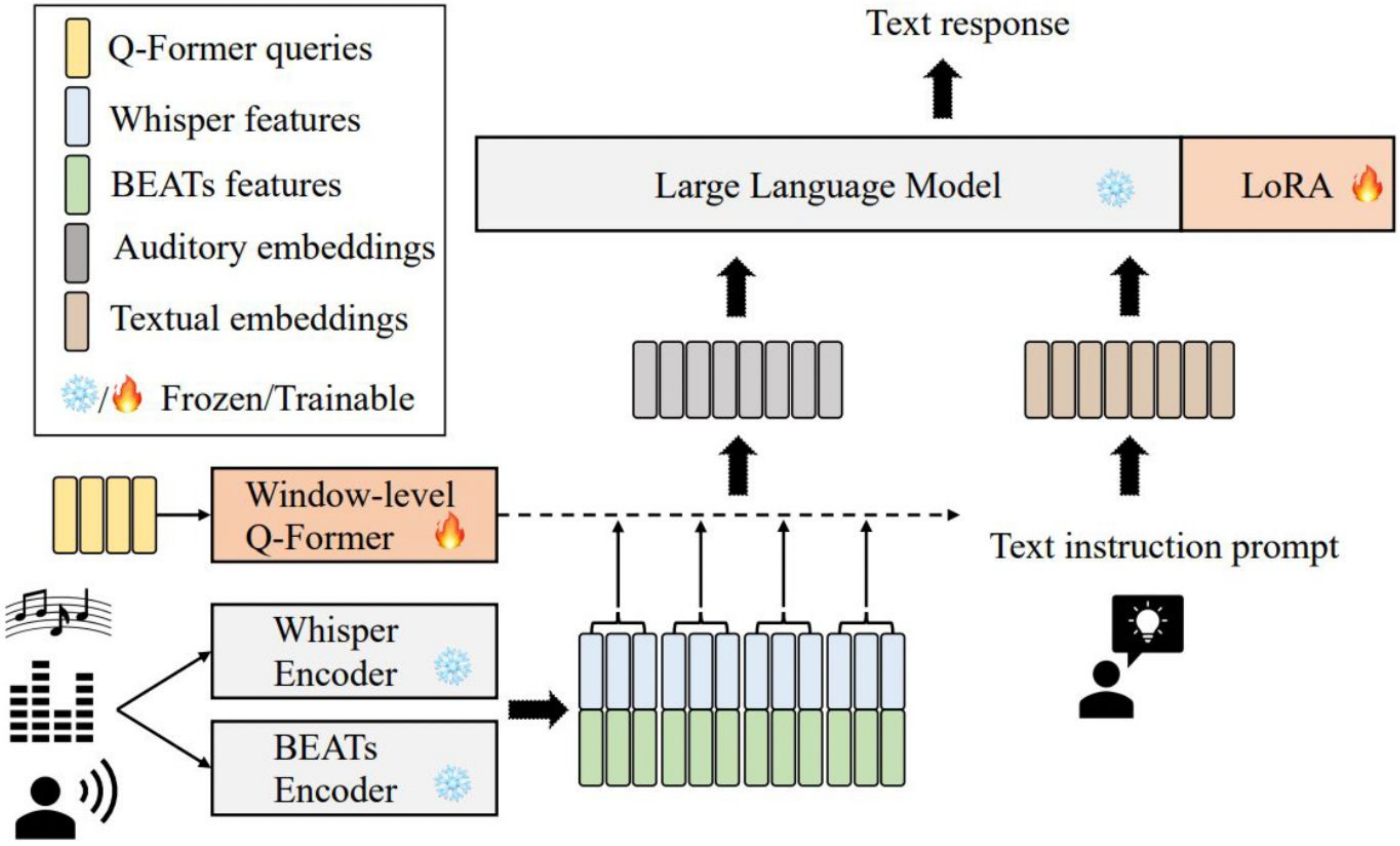
- Low Latency • Expressive • Reduce Error Propagation
- Higher Token Overhead • Need High Resource • Hard to debug

LTU: Listen, Think and Understand



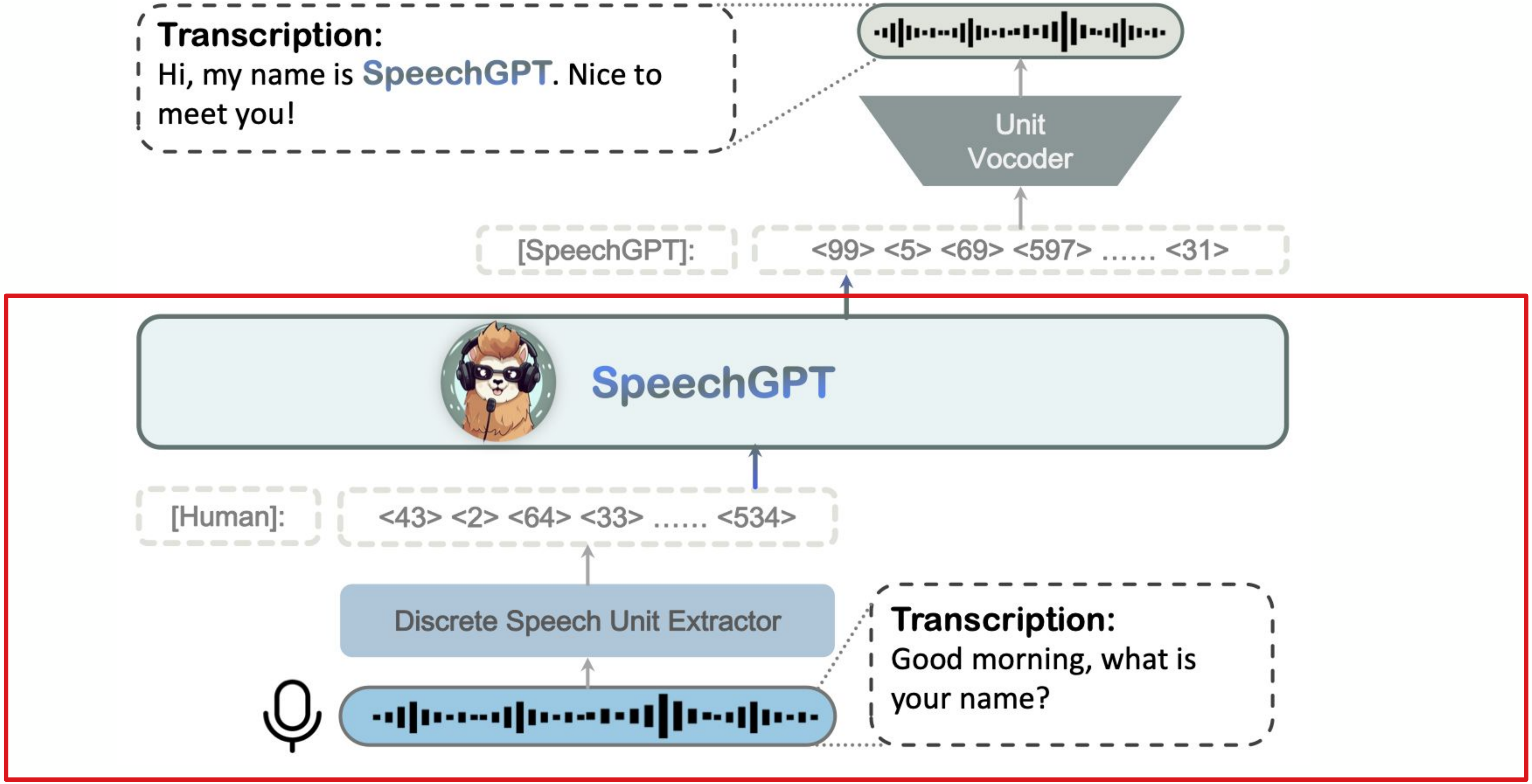
Gong et al., Listen, think, and understand, ICLR 2024

SALMONN



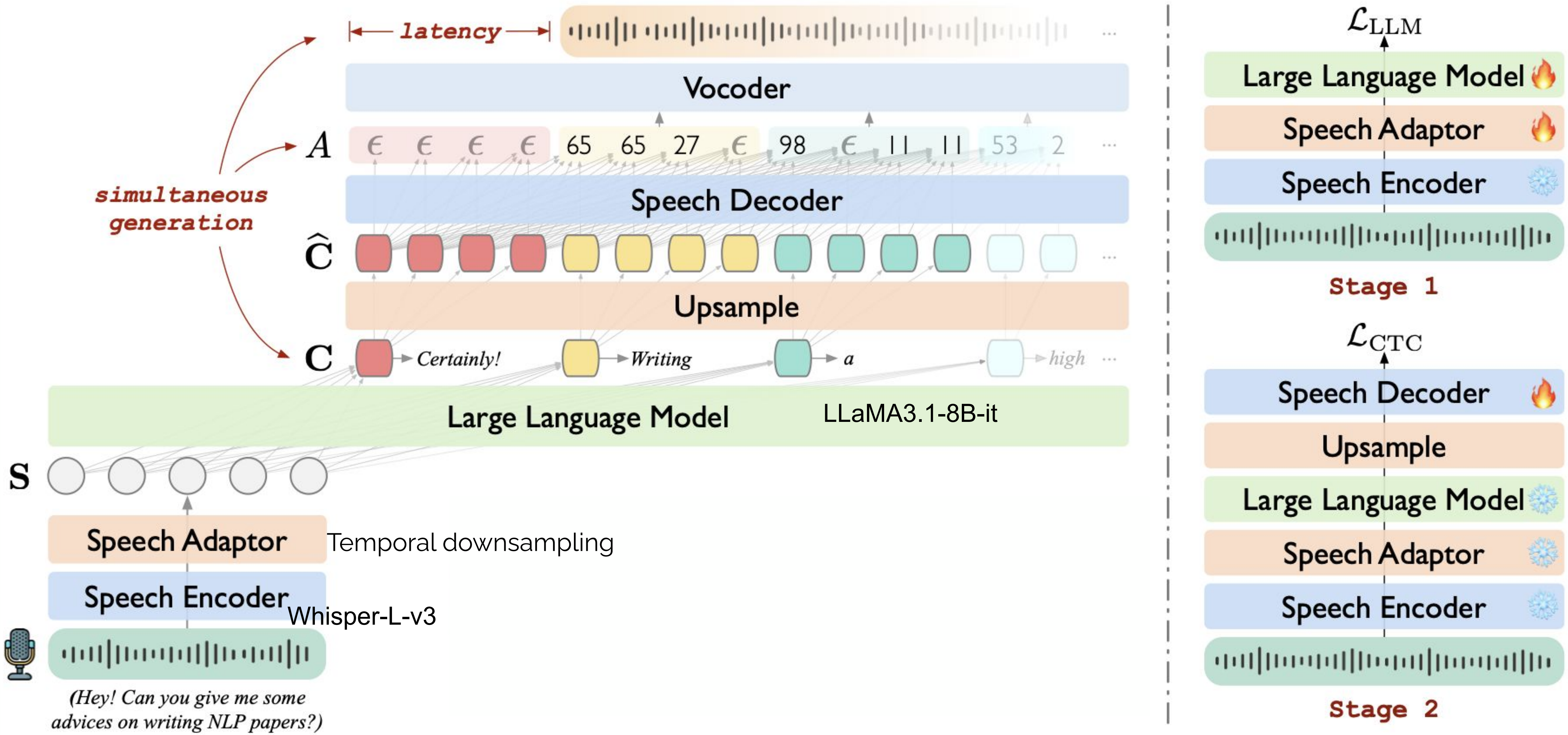
Tang et al., SALMONN: Towards Generic Hearing Abilities for Large Language Models. ICLR 2024

SpeechGPT



Tang et al., SALMONN: Towards Generic Hearing Abilities for Large Language Models. ICLR 2024

LLaMA-Omni

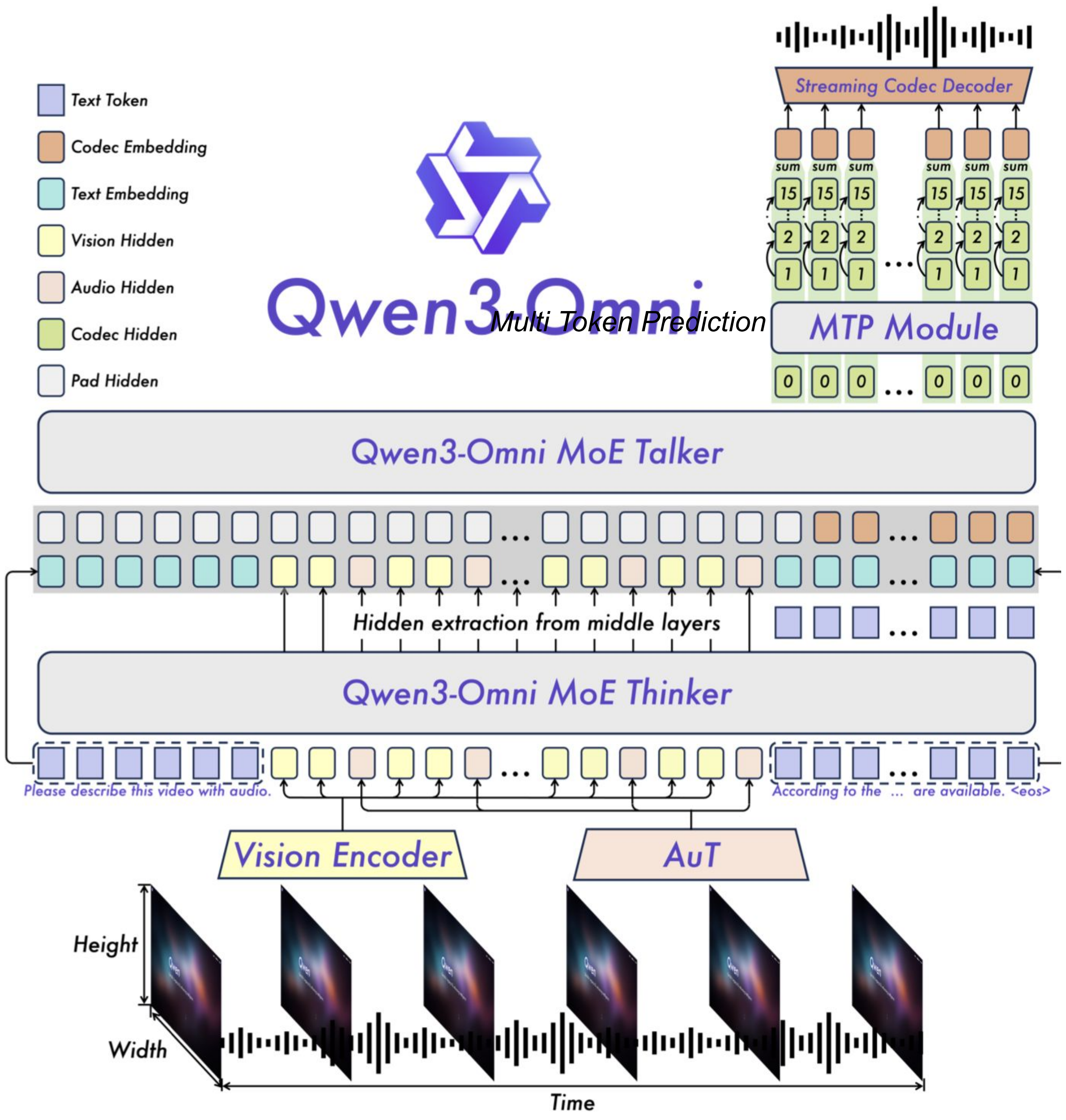


latency
Vocoder
A
ϵ ϵ ϵ ϵ 65 65 27 ϵ 98 ϵ | | | | 53 2 ...
Speech Decoder
Ĉ
Upsample
C
Certainly!
Writing
a
high
simultaneous generation
Large Language Model
LLaMA3.1-8B-it
S
Speech Adaptor
Temporal downsampling
Speech Encoder
Whisper-L-v3
(Hey! Can you give me some advices on writing NLP papers?)
$\mathcal{L}_{\text{LLM}}$
Large Language Model
Speech Adaptor
Speech Encoder
Stage 1
$\mathcal{L}_{\text{CTC}}$
Speech Decoder
Upsample
Large Language Model
Speech Adaptor
Speech Encoder
Stage 2

Qwen3-Omni

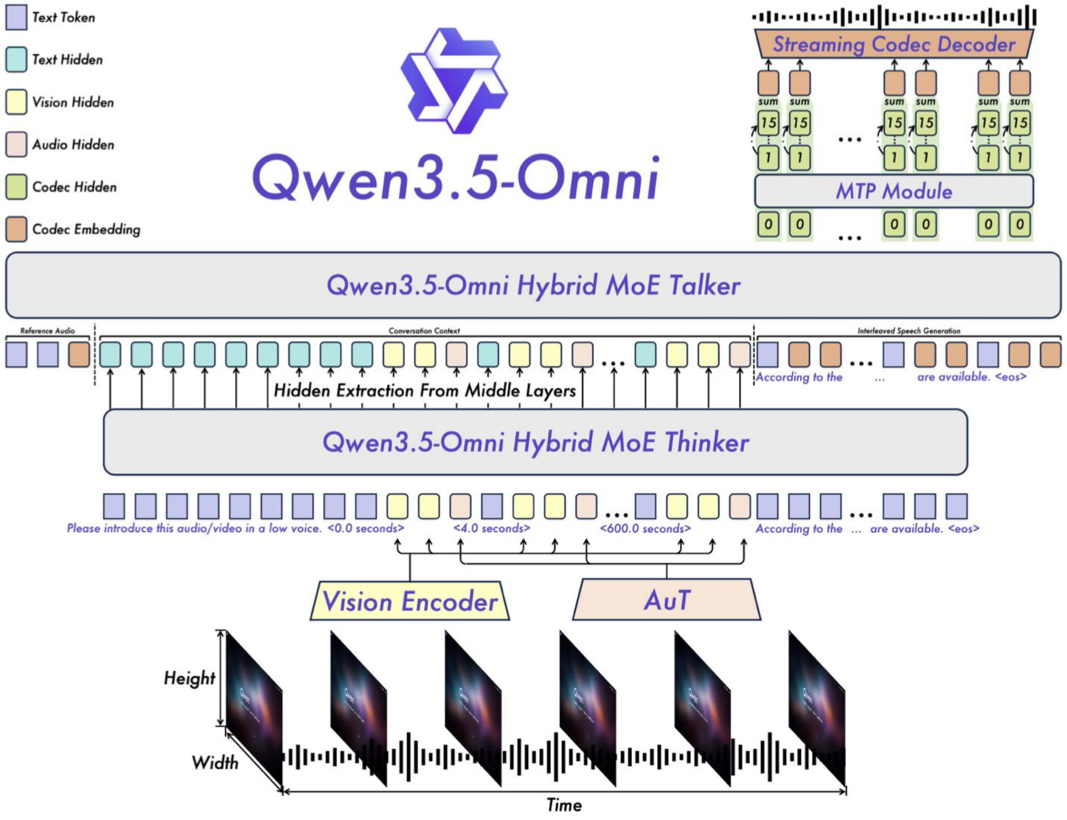
Stages

1. **Encoder alignment:** AuT, vision encoder, modality adapters (Thinker frozen).
2. **Full multimodal pretraining:** Full.
3. **Audio-text / audio-video instruction tuning:** Full. Focus: ASR, speech translation, audio QA, audio-video reasoning.
4. **Talker speech generation training:** Talker, speech codec prediction heads, MTP modules (Thinker frozen). Generate expressive multi-codebook speech tokens.
5. **Code2Wav training:** Lightweight causal ConvNet vocoder (Talker/Thinker frozen).
6. **RL / post-training**: Talker and/or policy heads for better speech stability



Qwen-3.5

1. **AuT encoder pretraining:** 20M → 40M. Build stronger multilingual audio representations.
2. **Full omnimodal pretraining:** Full.
3. **Thinker post-training:** Train the **Thinker** and related modality modules. Use specialist distillation, on-policy audio distillation, and instruction tuning. To improve audio-conditioned reasoning and interaction quality.
4. **Talker general training:** Use multilingual speech data with multimodal context. Generate expressive and context-aware speech.
5. **Talker long-context training:** Extend speech generation context and improve grounding.
6. **Preference / RL alignment:** Use multilingual preference pairs. Goal to improve speech naturalness, stability, and user preference.
7. **Speaker fine-tuning:** Fine-tune lightweight speaker-specific components. Improve voice cloning, speaker similarity, and expressive control.



Instruction Datasets · English Speech & Dialog SFT

Dataset	Languages	Hours	# Examples	Use
SpeechInstruct (SpeechGPT)	English	≈9k h speech	≈2M cross-modal instructions	HuBERT-unit instr SFT
InstructS2S-200K (LLaMA-Omni)	English	≈250 h synthetic	200k spoken-instr → response	TTS-synthesised SFT
LibriSQA	English	≈960 h (LibriSpeech)	107k Part I · 214k total	Free-form spoken QA
AudioChatLlama	English	—	speech-dialog SFT (proprietary)	End-to-end speech chat
SALMONN training mix	English (mostly)	re-uses encoders	600k SQA/AQA + 50k storytelling	Generic-hearing SFT
Distil-Whisper Voice Assistant SFT	English	re-uses Whisper data	≈100k voice-dialog turns	Distilled voice-assistant SFT
VoiceBench train slices	English	synthetic + real	8 task slices · 10s of k prompts	Voice-assistant SFT (incl. safety)

Instruction Datasets · Audio QA & Captioning

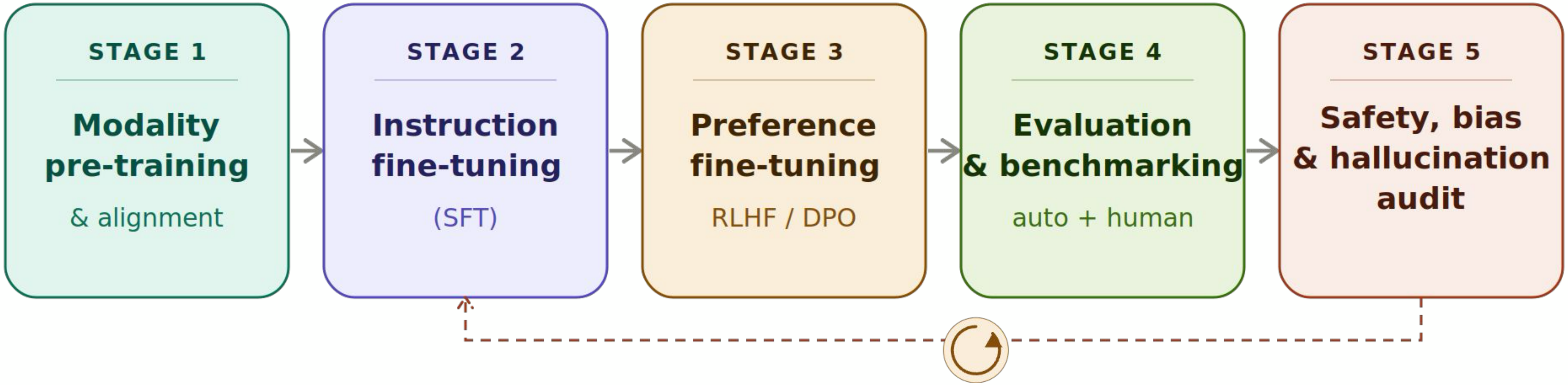
Dataset	Languages	Hours	# Examples	Use
OpenAQA-5M (LTU / LTU-AS)	English	≈5.8k ⟶ 9k h	5.6M ⟶ 9M (audio,Q,A)	Open + closed audio QA
Audio-FLAN	EN-heavy (multi-domain)	multi-source	100M+ instances · 80 tasks	Unified speech / sound / music
WavCaps	English	≈7,568 h	≈403k clip-caption pairs	Weak captions, ChatGPT-filtered
AudioCaps	English	≈128 h	≈46k clip-caption pairs	Human captions for AudioSet
Clotho	English	≈37 h	5,929 × 5 = 29.6k captions	Human-authored, retrieval
AudioSetCaps	English	≈5.8k h	≈1.9M enriched captions	LLM-assisted refinement
MMAU / MMAU-Pro	English	≈28 h	10k QA · 27 tasks	Reasoning eval (+ train mirrors)
BAT (Spatial Audio QA)	English	synth multi-channel	≈30k spatial-audio QA	Listen · spatialise · reason
Audio-Reasoner data	English	multi-source	1.2M structured CoT samples	Audio chain-of-thought
DCASE-2025 AQA	English	small subsets	Bioacoustics + Temporal + Complex	Challenge benchmark

Instruction Datasets · Multilingual

Dataset	Languages	Hours	# Examples	Use
Qwen-Audio / Qwen2-Audio mix	8+ langs (EN · ZH · JA · FR · DE · ES · IT · Cantonese)	proprietary; large	30+ tasks, NL prompts + DPO	Multitask + preference
Audio Flamingo Next (NVIDIA, 2026)	Multilingual	≈1M h audio	≈108M samples	Long-form, multi-talker, safety
Step-Audio 2 mix	Multilingual	proprietary	raw-audio in + reasoning-RL data	End-to-end audio LLM
SIFT-50M (Speech-Instruction FT)	5 langs (EN · ZH · DE · FR · ES)	≈14k h	50M instruction examples	Open multilingual SFT corpus
MLC-SLM 2025 challenge	11 langs (incl. EN · ZH · DE · FR · ES · JA · KO · RU · etc.)	≈1,604 h	real-world conversational dialogues	Multilingual conversational LLM
Emilia	6 langs (EN · ZH · DE · FR · JA · KO)	≈101k h	wild-style speech	Speech generation pretrain · gen SFT
WenetSpeech	Chinese Mandarin	22,400 h (10k labeled)	multi-domain YouTube + podcast	Mandarin ASR / instr backbone
AISHELL-1 · 2 · 3 · 4	Chinese Mandarin	170 + 1k + 85 + 120 h	multi-condition, multi-speaker	Mandarin ASR + speaker mixing
WenetSpeech-Yue	Cantonese (yue)	1k+ h	labelled + weakly-labelled	Cantonese AudioLLM training
KsponSpeech	Korean	969 h	≈2 000 speakers spontaneous	Korean SFT backbone

Recipes: Low-resource Speech Models

Recipes For Low-Resource



Singapore-Centric
Languages: English • Chinese • Malay • Tamil • Vietnamese • Indonesian • Thai • Filipino • Khmer • Lao • Burmese • Javanese • Sundanese

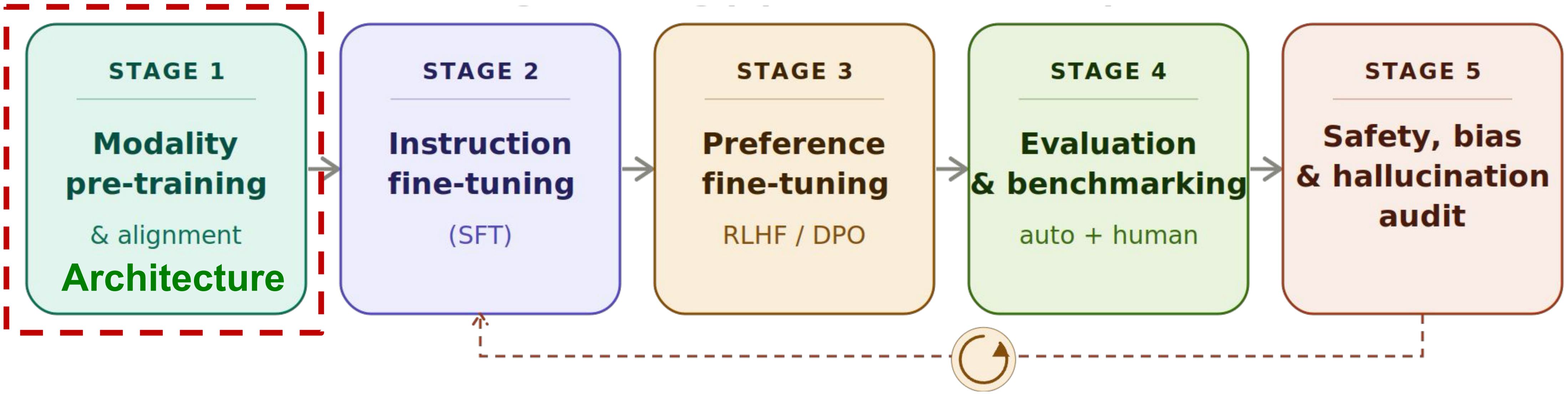
Arabic-Centric
Languages: Modern Standard Arabic (MSA), English
Dialects: Egyptian, Gulf, Levantine, North African

Tibetan (Ti-Audio)
Dialects: Amdo • Ü-Tsang • Kham

Thai-Centric
Languages: Thai (primary) + English (bilingual)

Typhoon 2: A Family of Open Text and Multimodal Thai Large Language Models (2024)
MERaLiON-AudioLLM: Bridging Audio and Language with Large Language Models (2025)
TI-AUDIO: THE FIRST MULTI-DIALECTAL END-TO-END SPEECH LLM FOR TIBETAN. (2026)
Multi-Task Instruction Tuning via Data Scheduling for Low-Resource Arabic AudioLLMs. (2026)

Recipes For Low-Resource



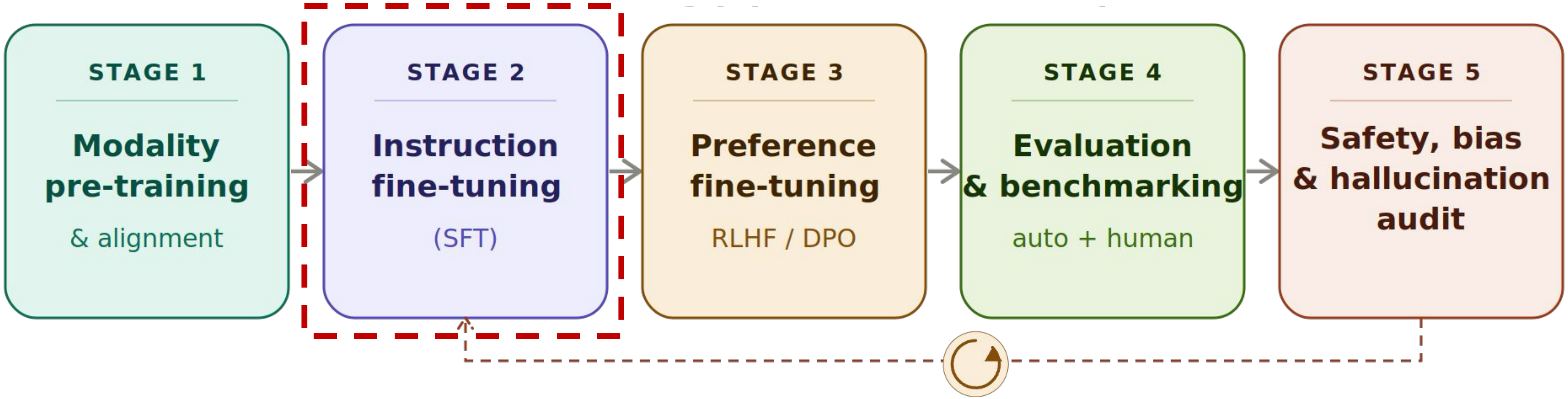
A Multilingual Base

Start from a multilingual foundation model.
Speech Encoder (Whisper, MMS, XLS-R, Seamless, AuT) and LLMs.

B Language Alignment

[Optional] Fine-tuned the model to the target language distribution.
Data Source: ASR, Audio Caption etc.

Recipes For Low-Resource



FT: Instruction Data • Strategies for FT • Balancing Imbalanced class/tasks/labels

Instruction Data Generation Pipelines

LLM-Based Metadata

- **Task Prompting:** Use LLMs to generate text instructions
- **Transcription:** Auto-generate "transcribe the audio" tasks
- **Translation:** Cross-lingual alignment prompts

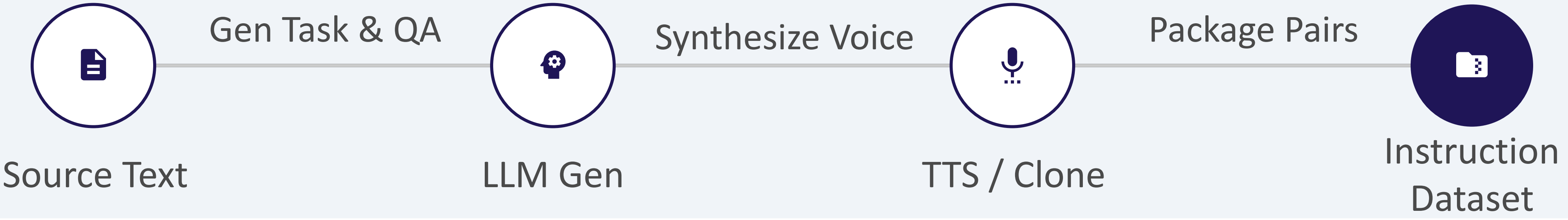
QA Generation

- **Contextual QA:** LLMs generate Q&A pairs from source text
- **Dialogue:** Simulating multi-turn spoken interactions

Synthetic Speech

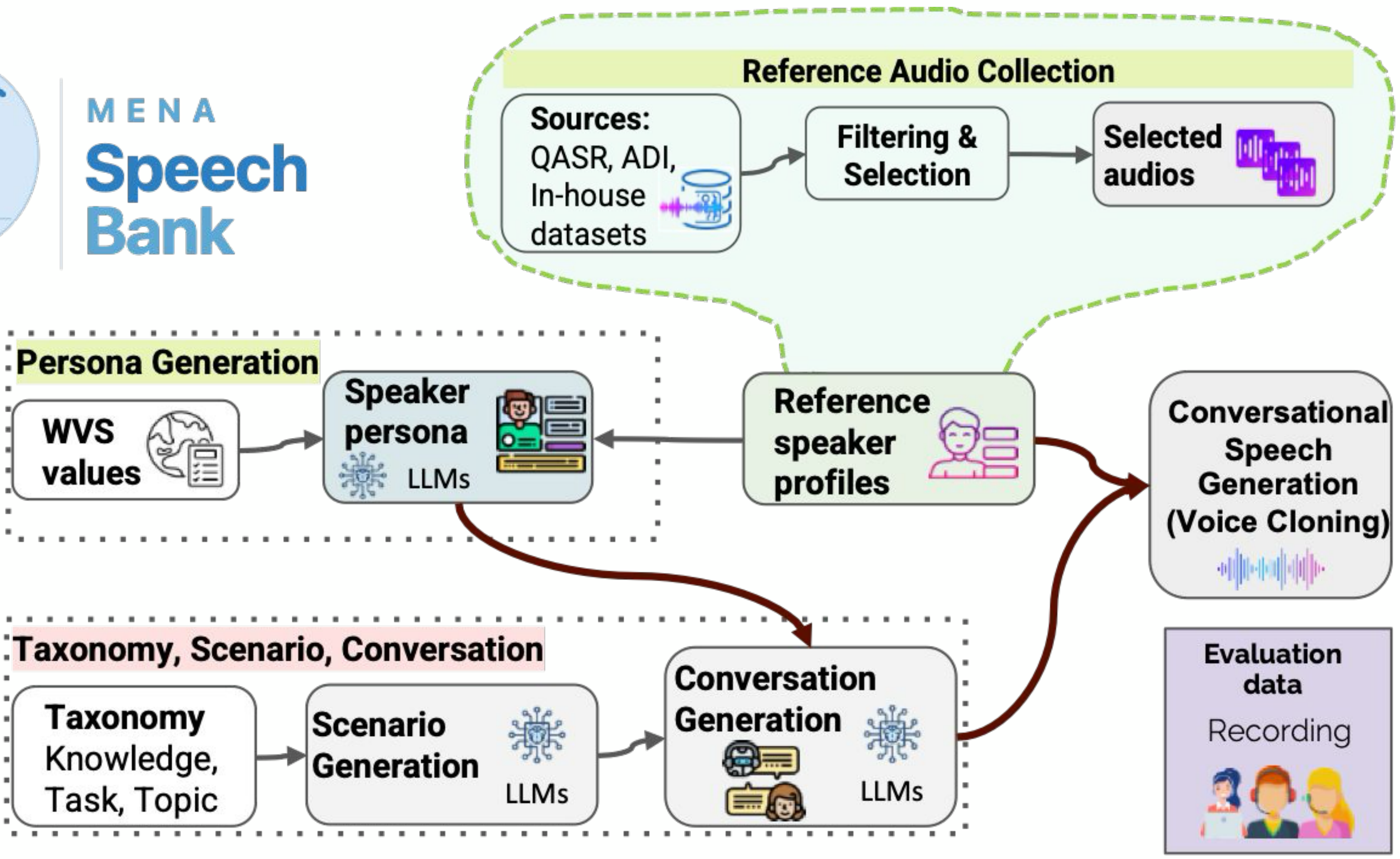
- **TTS Synthesis:** Converting generated QA into audio
- **Voice Cloning:** Using few-shot cloning for diverse speaker profiles

Workflow Diagram



Speech Instruction Data

MENA Speech Bank



17,641 utterances	124 unique speakers	18 countries	417K conversations	5K taxonomy driven scenarios

https://huggingface.co/datasets/QCRI/MenaSpeechBank

Ali et al. MENASpeechBank: A Reference Voice Bank with Persona-Conditioned Multi-Turn Conversations for AudioLLMs, 2026

Demographics

Abdulaziz

Age: 34
Location: Jeddah, Saudi Arabia
Marital status: Single
Living situation: With parents

Technology Access
- Budget smartphone
- Relies on public Wi-Fi
- Reliable power (electricity) access

Occupation
- HR Manager

Language & Culture
- Native Arabic speaker
- Muslim
- Family-oriented social structure

AI Literacy & Usage
- Intermediate: Familiarity with AI tools
- Explains bureaucratic procedures
- Assists with daily planning tasks

Topic: Hiring and Workforce Management
Scenario: A manager needs to quickly find temporary staff to cover for employees on sick leave this week

مرحبًا عبدالعزيز، أتمنى أن يكون يومك جيدًا. هل تحتاج إلى مساعدة في إيجاد موظفين مؤقتين لتغطية غياب بعض الزملاء هذا الأسبوع؟
Hello Abdulaziz, I hope you're having a good day. Do you need any help finding temporary staff to cover for some colleagues who will be absent this week?

نعم، لدي عدد من الموظفين في إجازة مرضية، وأحتاج إلى تعويض غيابهم بسرعة. ما أفضل طريقة للبحث عن موظفين مؤقتين في الوقت الحالي؟
Yes, I have several employees on sick leave, and I need to fill their absences quickly. What's the best way to find temporary staff right now?

مم... هل لديك قائمة سابقة بأسماء أشخاص عملوا معكم سابقًا؟ أحيانًا يكون التواصل معهم أسرع من البدء في البحث من جديد.
Hmm... Do you have a previous list of names of people who have worked with you before? Sometimes it's faster to contact them than to start searching all over again.

للأسف، ليس لدي قائمة جاهزة الآن. أغلب التوظيفات السابقة كانت دائمة. هل هناك منصات إلكترونية موثوقة للبحث عن موظفين مؤقتين في جدة؟
Unfortunately, I don't have a list ready right now. Most of my previous hires were permanent. Are there any reliable online platforms for finding temporary employees in Jeddah?

نعم، هناك بعض المنصات الإلكترونية المحلية التي تقدم خدمات التوظيف المؤقت، مثل منصة "فرصة" ومنصة "توظيف". يمكنك إنشاء إعلان يوضح نوع الوظيفة وعدد الأيام المطلوبة.
Yes, there are some local online platforms that offer temporary employment services, such as "Foras" and "Tawzeef." You can create an advertisement specifying the type of job and the number of days required.

......

Speech Instruction Data

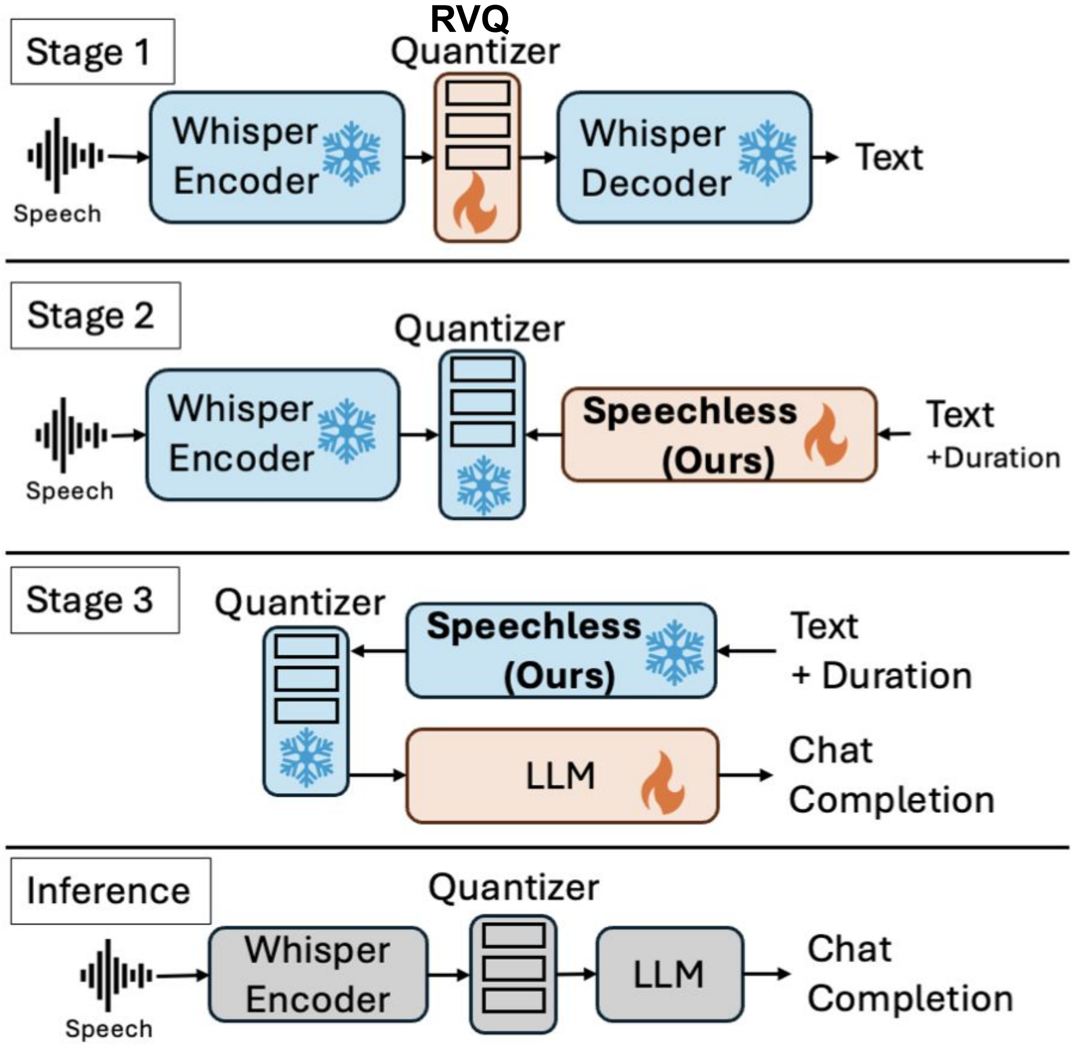
- Trained using: 330k multiturn (Avg. 5 turn/conversation)
- Human Test: 200 Conversation (recorded)
- **Average Rubric Score (ARS)**, computed as the mean pass rate across rubric checks, and
- **Average Pass Rate (APR)**, computed as the fraction of turns that satisfy the required rubric checks jointly (i.e., an over all pass).

Models	Synthetic				Human			
	Conversation		*Turn*		*Conversation*		*Turn*	
	APR	ARS	APR	ARS	APR	ARS	APR	ARS
Large Closed Models								
Gemini-2.5 pro	0.155	0.288	0.391	0.814	0.044	0.399	0.274	0.691
GPT-audio	0.214	0.449	0.423	0.893	0.005	0.314	0.247	0.681
GPT-audio (STT) + GPT-5.2-chat	0.694	0.302	0.573	0.928	0.680	0.288	0.548	0.930
Open and Fine-tuned Models								
Qwen2.5-Omni-7B	0.010	0.099	0.048	0.503	0.005	0.100	0.046	0.490
Qwen2.5-Omni-7B FT	0.000	0.252	0.137	0.599	0.015	0.271	0.128	0.589
Arabic-centric Models (ASR + LLM)								
Fanar (Aura) + ALLaM-7B	0.053	0.332	0.305	0.812	0.024	0.310	0.291	0.801
Fanar (Aura) + Fanar-C-1-8.7B	0.393	0.664	0.576	0.878	0.388	0.674	0.564	0.883
Fanar (Aura) + Fanar-C-2-27B	0.408	0.707	0.607	0.900	0.415	0.694	0.594	0.892

Ali et al. MENASpeechBank: A Reference Voice Bank with Persona-Conditioned Multi-Turn Conversations for AudioLLMs, 2026

Speech Instruction Data

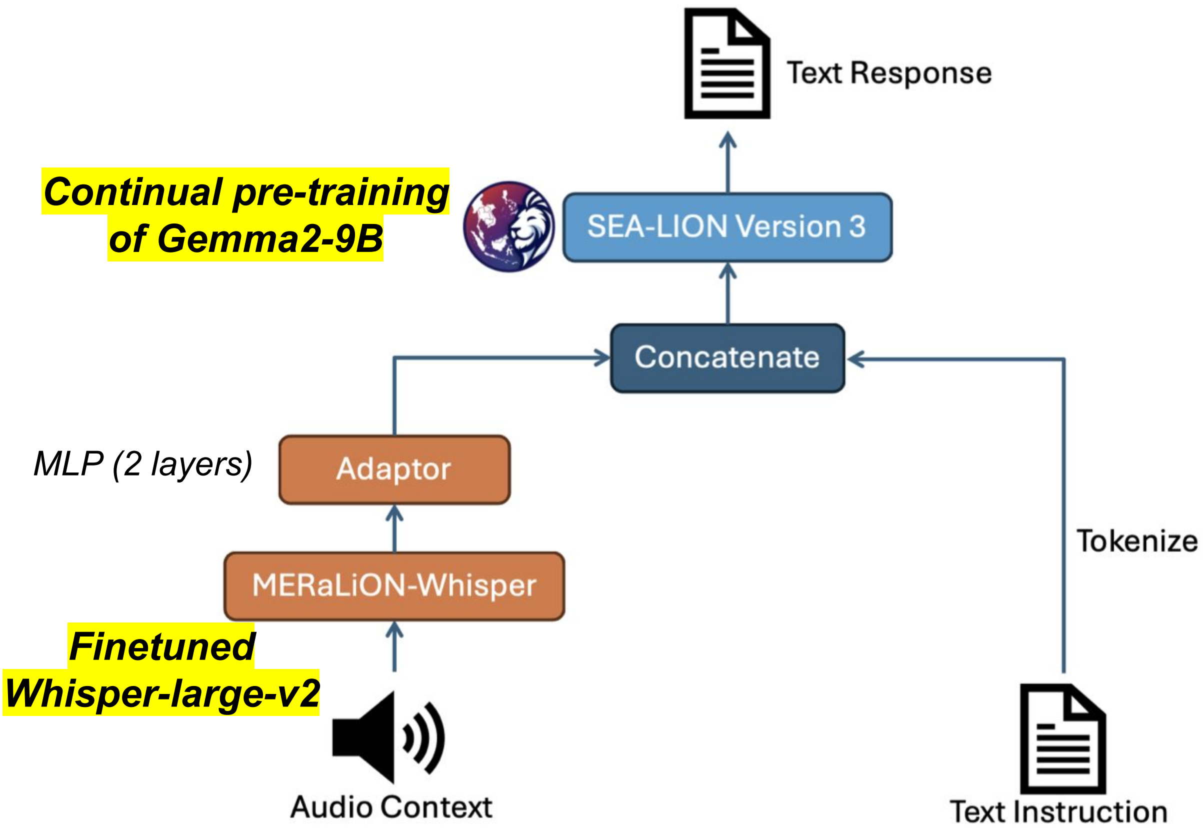
- TTS is a luxury for low-resource.
- **Core Idea is to simulate speech-representation instead of having the speech instruction itself !**



Speechless: Speech Instruction Training Without Speech for Low Resource Languages. 2025

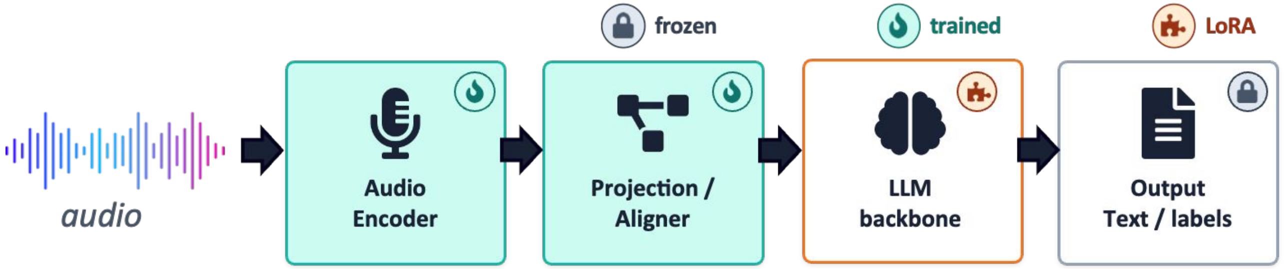
MERaLiON-AudioLLM

Languages: English • Chinese • Malay • Tamil • Vietnamese • Indonesian • Thai • Filipino • Khmer • Lao • Burmese • Javanese • Sundanese



260K hrs Speech-text instruction pairs
Public + synthesised + curated data

1. **Local speech encoder adaptation**
2. **Multimodal Instruction Fine-tuning**



MERaLiON Instruction Data Overview

260K hrs
Speech-text instruction pairs
Public + synthesised + curated data

13+ langs
Coverage

6 Tasks: ASR, ST, SQA, SDS, Para (Gender/Emotion/..)
Instruction Types

National Speech Corpus (NSC) — 10,600 hrs

Part 1-2 6,500 hrs	Prompted readings — people, food, brands
Part 3 900 hrs	Conversational data — daily life, games
Part 4 900 hrs	Code-switching — Singlish + Mother Tongue
Part 5 1,000 hrs	Scripted recordings across domains
Part 6 1,300 hrs	Simulated phone calls — hotel, bank, govt

Speaker metadata (gender, age, ethnicity, L1) used for paralinguistics tasks. Verified + filtered for label accuracy.

Public Datasets

GigaSpeech	10K hrs English multi-domain ASR
GigaSpeech 2	SEA languages — Thai, Indonesian, Vietnamese
Common Voice	Massively multilingual crowd-sourced speech
AudioCaps	Audio event captioning with descriptions
VoiceAssistant-400K	400K instruction-following pairs
YODAS2	Large-scale diverse audio data

Synthesised + Augmented

SDS	Spoken Dialogue Summarization
SQA	Speech Question Answering
GR	Gender Recognition from speech

MERaLiON-AudioLLM

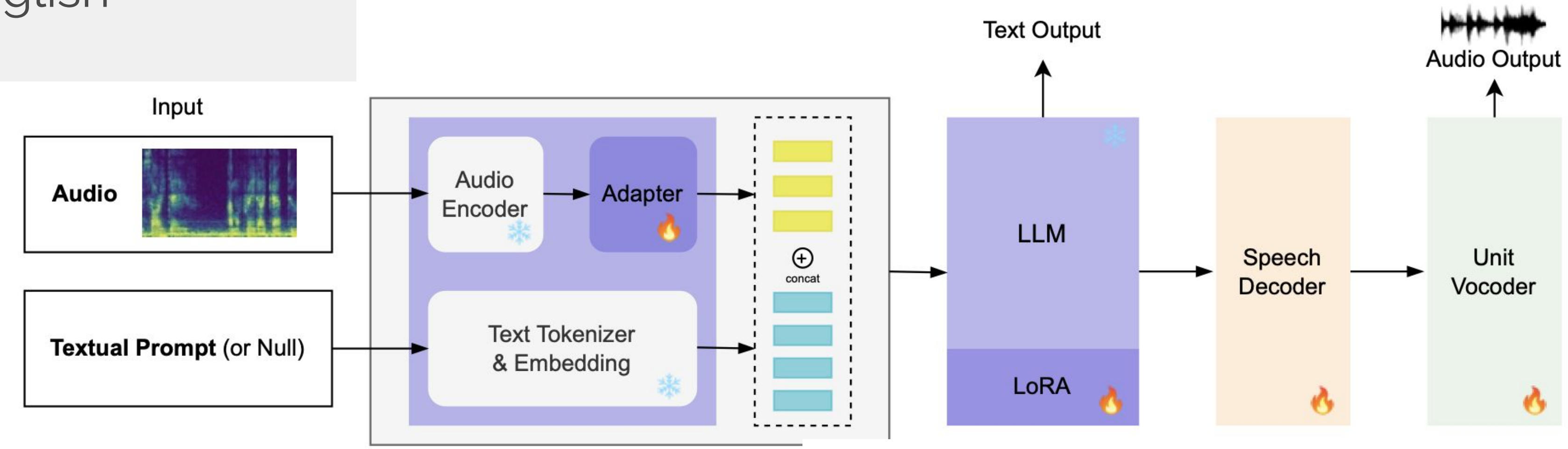
- **Local speech adaptation plus multitask instruction tuning (even synthetic) improves speech-language understanding for Singapore-centric audio tasks.**

- **Key lesson: regional AudioLLMs need local speech, accents, and task mixtures, not only generic multilingual models**

Typhoon2-Audio

2.46M total training examples (1.82M pre-train + 640K SFT)

Languages: Thai (primary) + English (bilingual)



Audio Understanding:

1. Pre-training: Trains Q-Former + MLP adapter for audio-to-LLM alignment (ASR, AC)
2. SFT: Adapter + LoRA

Component	Initialization
Whisper Encoder	Thonburian Whisper
BEATs	BEATs pre-trained weights
Q-Former	Randomly Initialized
Linear Layers	Randomly Initialized
LLM	Typhoon2-8B (Llama-3.1-based)
Speech Decoder	Randomly Initialized
Vocoder	Randomly Initialized

Speech decoder training:

1. Train decoder to predict discrete speech units from LLM hidden states.
   - Use CTC loss.
2. Vocoder training: Train HiFi-GAN-style unit vocoder.

Experiment	#Ex	ASR*↓	Th2En↑	SpokenQA↑	SpeechIF↑	ComplexIF†↑
Pre-trained	-	**13.52**	0.00	28.33	1.12	1.41
100% English SFT	600K	80.86	6.01	36.88	1.48	6.35
Our SFT recipe	640K	16.89	**24.14**	**64.60**	**6.11**	**7.54**

Typhoon2-Audio

Pre-training Data (1.82M examples)

ASR-EN	LibriSpeech + GigaSpeech (English read/web speech)
ASR-TH	CommonVoice 17 + GigaSpeech 2 (Thai spontaneous)
AC-EN	AudioCaps + Clotho (English audio descriptions)
AC-TH	AudioCaps + Clotho translated to Thai
NOTE	Adapter-only training; encoder frozen throughout

No Thai audio captioning exists natively; all translated via Typhoon LLM.

SFT Data (640K examples)

ASR	English + Thai transcription with diverse prompts
ST	Speech translation (TH→EN, EN→TH, CoVoST2)
QA	Audio document QA from SALMONN + LTU data
AC	Audio captioning with GPT-4o generated prompts
SIF	Speech instruction following (TTS-synthesised Thai)
GC	Gender classification from FLEURS (EN + TH)

Strategy

TTS	TTS-synthesised Thai speech instructions (no native data exists)
MIX	90/10 EN/TH ratio in SFT; 10% prompt translation to Thai

Typhoon2-Audio

- **Thai-specific adaptation (adapter + LoRA-LLM) improves Thai audio performance while preserving broader English/multimodal capabilities.**

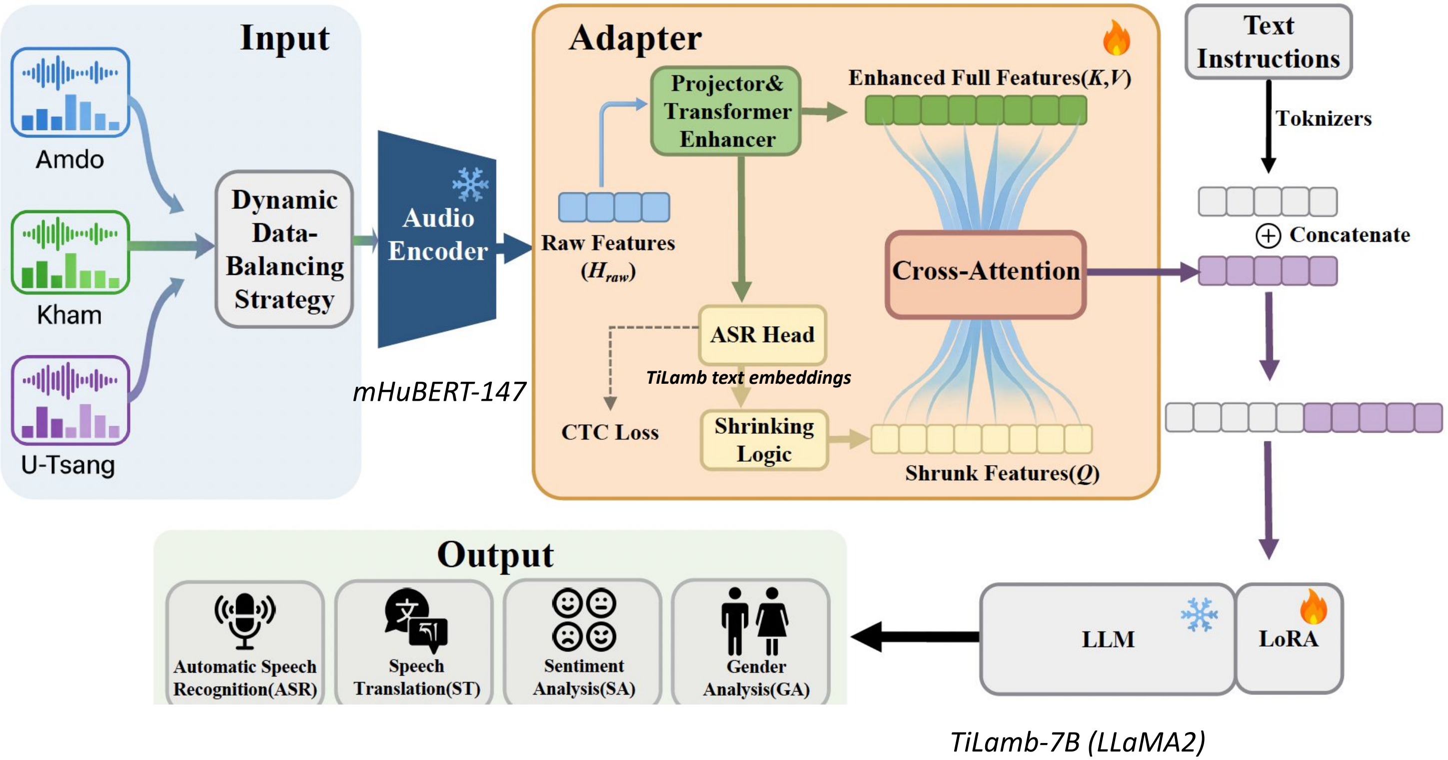
- **Key lesson:** low-resource AudioLLMs benefit from language-specialized encoders, adapter training, and efficient LoRA-based SFT. Synthesized data is also a key!

Ti-Audio

499 hrs Total audio from 6 Tibetan speech datasets, 319K+ samples

Dialects: Amdo (26.9%) • Ü-Tsang (46.9%) • Kham (24.9%)

1. **Computes sampling probability using temperature-aware sampling**
   - Uses dialect-task pair size and a temperature hyperparameter
   - Upsamples low-resource combinations
   - Prevents overfitting to mainstream dialects
2. **Train Dynamic Q-Former Adapter for speech–LLM alignment.**
   - Use CTC loss (peaks) to guide informative frame selection.
3. **Apply LoRA to adapt the Tibetan LLM.**
   - Jointly optimize generation loss + CTC loss.



Key Results

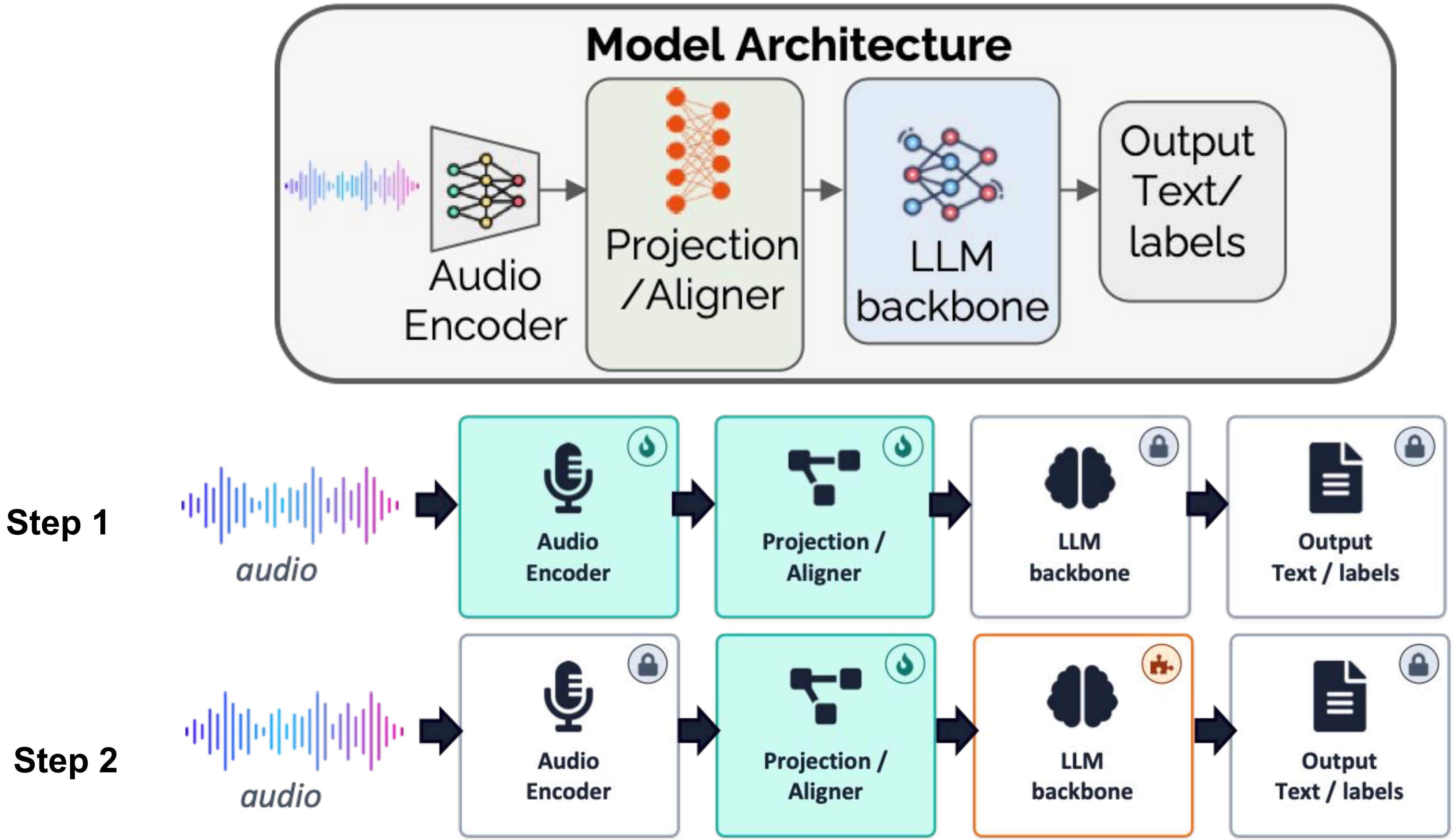
E2E	Outperforms cascaded mHuBERT+LLM and Meta Omnilingual (WER 73%)
DIAL	Cross-dialect training improves all 3 dialects consistently
12%	WER reduction over baselines via Dynamic Q-Former Adapter

Ti-Audio

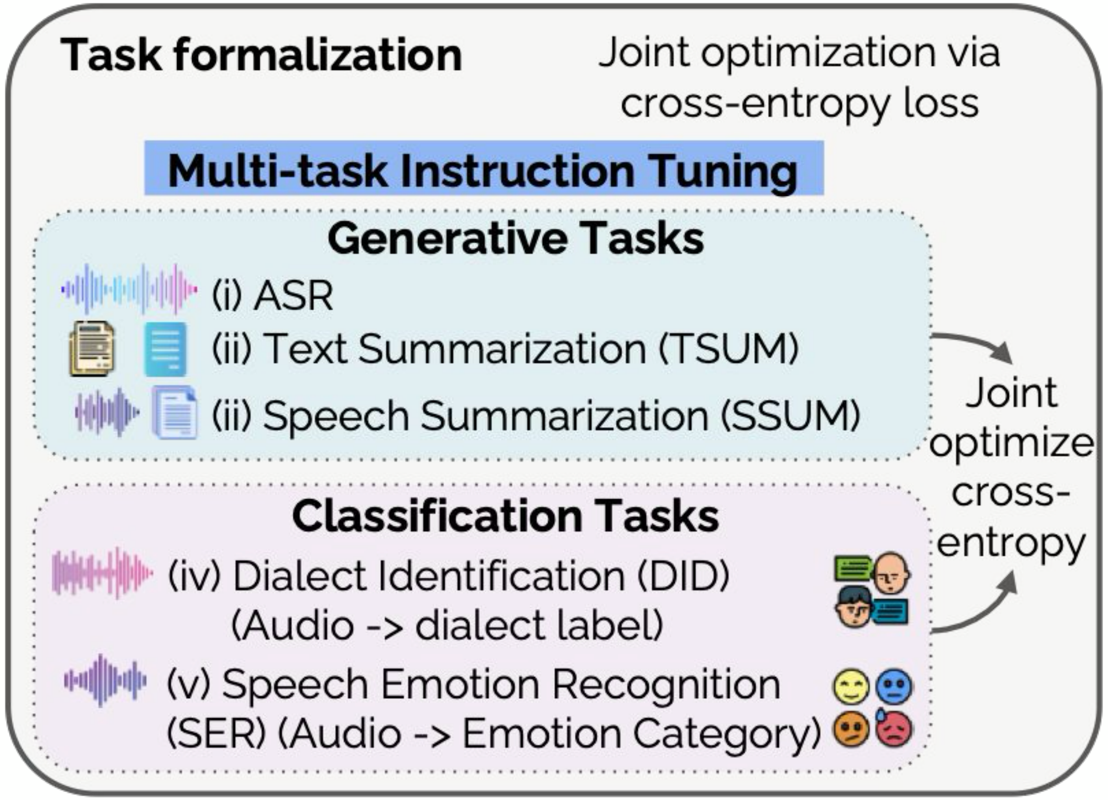
- **Cross-dialect joint training outperforms isolated dialect training and achieves state-of-the-art results on Tibetan ASR and speech translation benchmarks.**

- **Key lesson:** Related dialects can help each other when sampling is balanced and alignment is data-efficient.

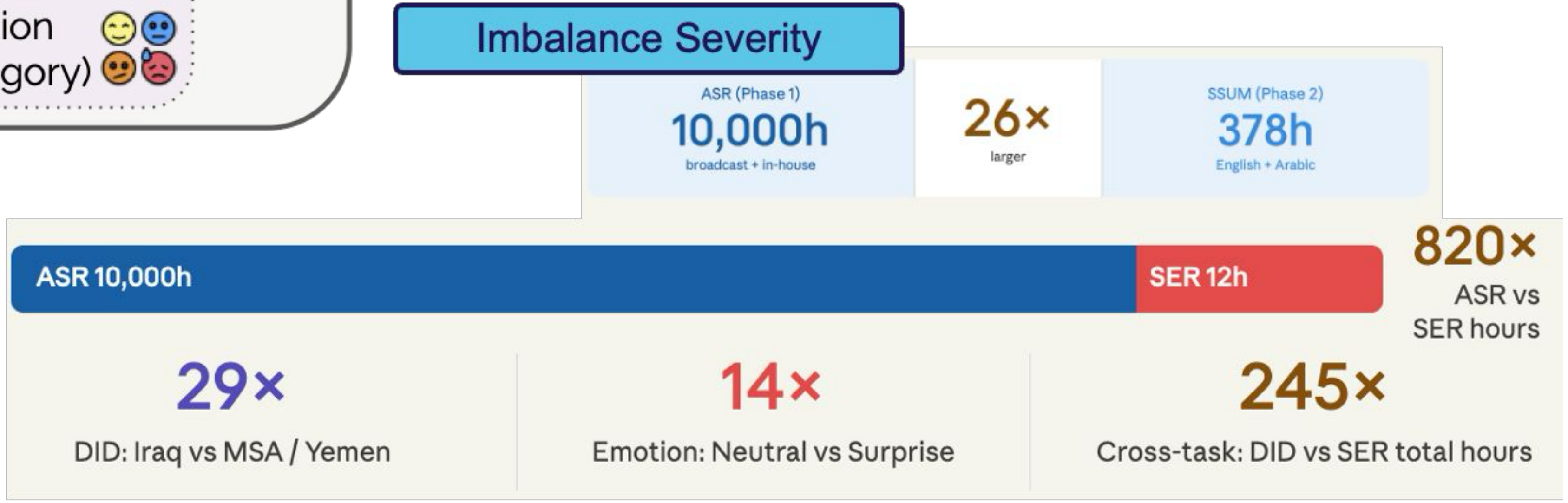
Arabic AudioLLM



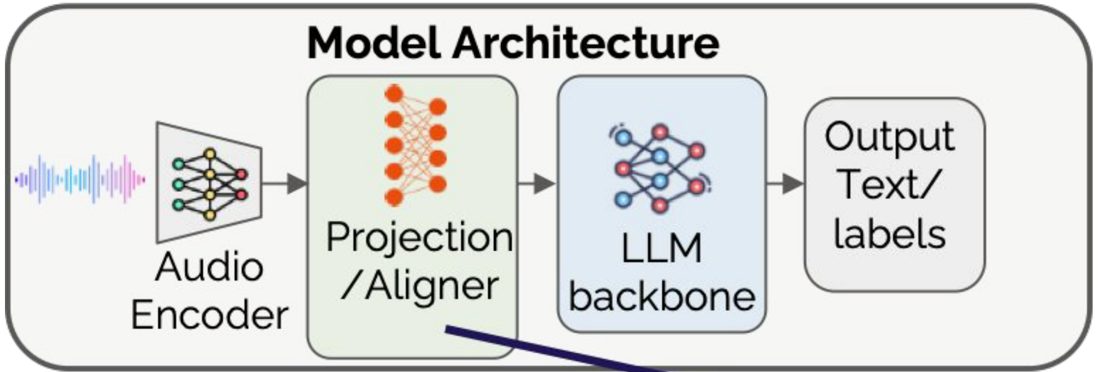
Arabic AudioLLM

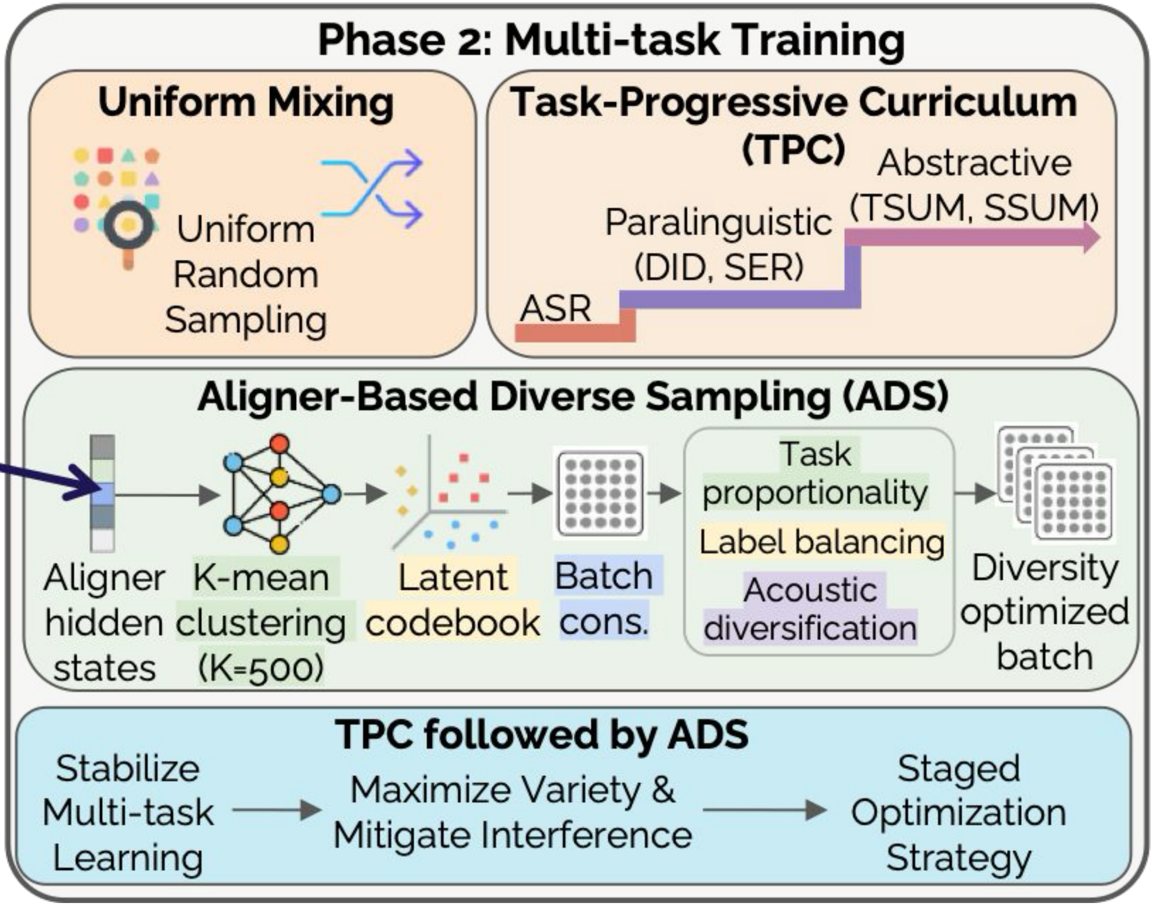


Phase 2: Multi-task Training			
Gen.	ASR	20% subset of Phase 1	≈1,740h
Gen.	SSUM	MegaSSUM (Matsuura et al., 2024) + AraMega-SSum (ours)	378h
Gen.	TSUM	MegaSSUM (Matsuura et al., 2024) + AraMega-SSum (ours)	101,242#
Disc.	DID	ADI-17 (Shon et al., 2020) + ADI-5 (MSA only) (Ali et al., 2017b)	≈2,983h
Disc.	Emo.	ANAD (Klaylat et al., 2018) + EAED (Safwat et al., 2023) + KEDAS (Belhadj et al., 2022) + KSU (Meftah et al., 2021) + BAVED (Aouf, 2020) + YSED (Derhem et al., 2025)	12.17h
Total			**5,113h + 101,242#**



Arabic AudioLLM





Findings

01 — Curriculum ordering improves ASR, weakens paralinguistics

02 — ADS accelerates paralinguistic learning, trades off generative quality

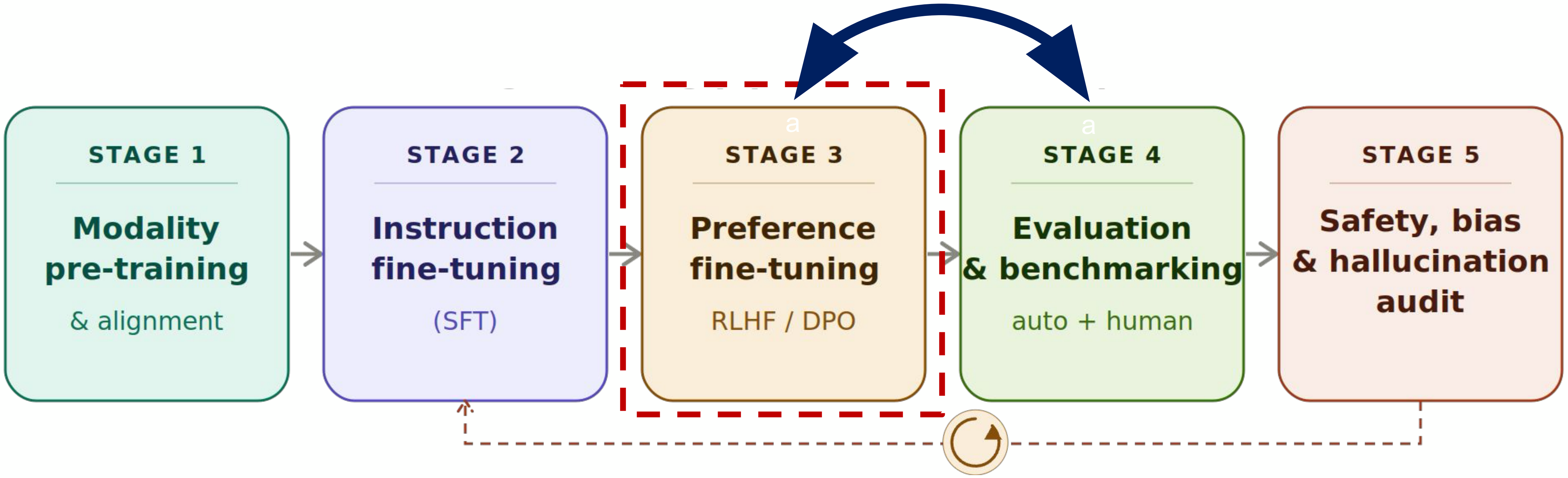
03 — TPC→ADS hybrid gives the most balanced performance

Bhatti et al., Multi-Task Instruction Tuning via Data Scheduling for Low-Resource Arabic AudioLLMs, 2026

Instruction Datasets · Low-Resource & Dialectal

Dataset	Languages	Hours	# Examples	Use
MENASpeechBank	Arabic (MSA · Egyptian · Maghrebi · Gulf · Levantine) + EN	26.4 h ref + synthetic	18k utt · 124 spk · ≈417k dialogues	Persona-conditioned MENA dialog SFT
MERaLiON-AudioLLM	Singlish · EN · Malay · Tamil · Mandarin	≈260k h	62M multimodal instructions	Dialect-aware Southeast Asia
Typhoon-Audio SFT	Thai + English	re-uses public	1.82M PT · 640k SFT	Low-resource Thai instruction
BhasaAnuvaad	13 Indian langs (Hi · Ta · Bn · Te · Mr · Gu · Ka · Ml · Pa · Or · As · Ne · Ur)	≈44,400 h	17M text seg · S2T	Indic AST training
IndicVoices-R	22 Indic langs (multi-dialect, multi-domain)	≈1,600+ h	speaker- and dialect-balanced	Open Indic multilingual SFT
IndicST	8 Indian langs ↔ EN	smaller, matched	translation pairs	Indian Speech LLM AST
Indic-TEDST	9 Indian langs	≈203 h	EN⟶Indic TED talks	Indic ST eval/train
Inkuba Instruct (text)	5 African langs (Hausa · Swahili · Zulu · Yoruba · Xhosa)	— (text only)	MT · NER · QA · classification	Text instruction ⟶ audio adapt
African ASR pseudo-labels	Kinyarwanda · Luganda · Swahili · Amharic · Yoruba	≈3,800 + 1,700 + 1,100 h	ASR corpora used for SFT bootstrap	African low-resource AudioLLM
Ti-Audio Tibetan dialect mix	Tibetan dialects (Lhasa · Kham · Amdo)	small — pooled across dialects	dialect-balanced QA + ASR	Multi-dialect Tibetan SpeechLLM

Recipes For Low-Resource



Most pre-2024 AudioLLMs skip Preference training entirely.
Voice-dialog systems and reasoning-heavy models add it on top of SFT.

Preference Data

Work	Task	Preference data	Learning method	Main finding
SpeechAlign	TTS / codec LM	Gold codec tokens preferred over synthetic tokens; human verification	DPO / PPO iterative alignment	Bridges train–inference distribution gap; improves speech generation
Speechworthy ITLM	Text response for speech delivery	20K response pairs from Dolly prompts with listened judgments	Reward model + preference tuning	Models learn more speakable, concise responses
IPO	Audio captioning + speech translation	Few hundred human referee labels over beam hypotheses	Imbalanced Preference Optimization	Weak and strong models improve from limited feedback
BPO-AVASR	Audiovisual ASR	Simulated audio/vision corruptions and transcript rewrites	Bifocal Preference Optimization	Improves real-world video ASR by emphasizing correct vs. error pairs
Spoken dialogue	Full-duplex S2S dialogue	>150K preference pairs from raw multi-turn conversations with AI feedback	Offline alignment	Covers linguistic content plus timing, interruption, and turn context

Mainly focus on the Generation Part!

Research Trends 2025–2026

Real-time omni	**GPT-4o · Qwen2.5-Omni · Step-Audio 2** — streaming, low-latency conversation.
Audio reasoning	**Audio-Reasoner · Step-Audio-R1 · ART** — beyond transcription, into multi-hop reasoning.
Preference alignment	**FPO · INTP · DPO-SE** — DPO and reward modelling for TTS and audio enhancement.
Low-resource LLMs	**Speechless · Ti-Audio · MERaLiON** — cross-lingual adaptation and in-the-wild benchmarking.
Long-context audio	Lectures · meetings · field recordings — **New benchmarks** emerging for long-form understanding.

QA

Evaluation, Benchmarking Resources, Error analysis

Firoj Alam
Senior Scientist, QCRI

Evaluation

Evaluation Metrics

Classification & Speech

- Accuracy, Precision, Recall, F1
- WER, CER

Surface-level Similarity

Exact Match, BLEU, Rouge, METEOR

Semantic Similarity

BERTScore with language specific BERT model

Neural Learned Metrics

BLEURT, COMET

Evaluation

Evaluation Metrics (Challenge)

Diverse LMM Tasks

Modern LMM tasks (visual QA, document QA, OCR, etc.) are increasingly diverse and complex, requiring more than simple matching.

Complex Answers

A correct answer may be longer, paraphrased, multilingual, or culturally grounded, making surface overlap unreliable.

Incomplete References

Reference answers are often missing or incomplete; many questions naturally have multiple valid answers.

Metric Limitations

BLEU/ROUGE fail to directly measure visual grounding, factuality, reasoning, safety, or helpfulness.

Evaluation

Evaluation Metrics (Challenge)

*"Evaluation is moving from reference-overlap to **judgment-based**, **rubric-based**, and task-aware scoring."*

LLM-as-a-judge

Strong LLMs are increasingly used to grade/score open-ended responses when reference answers are insufficient.

Rubric-based scoring

Instead of one number, evaluators score multiple dimensions:

Correctness Grounding Completeness Reasoning Fluency Safety

Evaluation

In the wild testing

Chatbot Arena

- Rank LLMs by real-user preference, not static benchmarks

- Comparisons over open-ended prompts users actually care about

- Every pairwise vote updates an Elo-style rating per model

(e.g., Zheng et al. 2024)

Users preference

- Asking users/testers to interact with the LLMs and provide their preferences (likes/dislikes) and why they disliked
  (e.g., Fanar Team et al., 2025)

Rubric Based Evaluation

What is a rubric?

- An explicit, multi-criterion scoring guide that asks an evaluator (human or LLM) to score a response **criterion by criterion** rather than producing a single overall preference.
- Recent work (*Prometheus, HealthBench, BiGGen-Bench, Autorubric, Rulers*) emphasises instance-specific rubrics and evidence-based scoring.

Prometheus (Kim et al. ICLR 2024, arXiv:2310.08491); HealthBench (Arora et al. arXiv:2505.08775); LLM-Rubric (Hashemi et al., arXiv:2501.00274); Autorubric (Rao et al., arXiv:2603.00077); Rulers (Hong et al., arXiv:2601.08654)

Rubric Based Evaluation

An example for Spoken Query Response:

Intent precision: does the response address the user's intended task without drifting to irrelevant or mismatched content?

Context awareness: does it correctly use information referred to in the query and stay consistent with the dialogue state?

Specificity: does it provide concrete, actionable details rather than vague generalities?

Depth and thoroughness: does it cover the key aspects needed to satisfy the request, including conditions or steps when applicable?

Grounding and honesty: does it avoid unsupported claims and communicate uncertainty or limitations rather than hallucinating?

Format and language compliance: does it follow requested formatting constraints and use the requested language and register consistently?

Coherence: is the response logically organised, internally consistent, and easy to follow?

Ali et al., MENASpeechBank: A Reference Voice Bank with Persona-Conditioned Multi-Turn Conversations for AudioLLMs, 2026

Rubric Based Evaluation

Average Pass Rate (APR): define a binary pass indicator p_i for query i as the product of its individual rubric checks. **APR** is then the percentage of responses that pass completely across the whole data set.

$$p_i = \prod_{j=1}^{N_i} \mathbb{1}_{r_{ij}} \qquad \text{APR} = \frac{1}{N}\sum_{i=1}^{N} p_i$$

Average Rubric Score (ARS): computed as the mean pass rate across all individual rubric checks. For each query i, it computes the fraction of satisfied criteria. The overall ARS is the average of these per-query pass rates across the entire test set

$$s_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbb{1}_{r_{ij}} \qquad \text{ARS} = \frac{1}{N}\sum_{i=1}^{N} s_i$$

Benchmarking Tools

- **LLMeBench** (Dalvi et al., 2023)
- **lm-evaluation-harness** (Gao et al., 2023)
- **LMMs-Eval: Probing Intelligence in the Real World** (Li et al., 2024)

LLMeBench

Make it super-simple and quick to **start experimenting** with LLMs, and **easily transfer that effort** to large scale evaluation

Exploration	Model comparison	Benchmarking suite	Many more...
Try a model with different prompts over the same dataset	Run the same prompt with multiple models	Create a suite of tasks and datasets and track a model's progress across all	Framework is flexible and extensible for new tasks, datasets, and models

http://llmebench.qcri.org/

LLMeBench: A Flexible Framework for Accelerating LLMs Benchmarking, (Dalvi et al. EACL, 2024)

Why LLMeBench?

Current LLM usage and benchmarking process

1. Read the data
2. Figure out how to access an LLM (e.g. GPT4)
3. Understand and write code to read the response
4. Explore with different prompts
5. Write some sort of loop over the data and prompts to see model responses on all samples
   a. Realize the request fails for many reasons ⇒ Write some code to retry failed requests
   b. Realize every time you run your code, you get different results ⇒ Modify code to set all appropriate model parameters for reproducible results
   c. Have an idea for a new prompt, figure out changing existing code to only run for new prompt while keeping results from older prompts
6. Process results
7. Rinse and Repeat for a new problem/dataset/task

Why LLMeBench?

1. Find your task, dataset and model in LLMeBench
   ⇒ Task/Data/Model not found?
   a. Edit existing task/data/model script for your needs
2. Run experiment!

Add a layer of abstraction so that you as a user can focus solely on getting the best performance out of the LLM

LLMeBench

Once an *asset* is written, LLMeBench takes care of everything else!

```
python -m llmebench assets/ results/
```


```
{
  "num_processed": 872,
  "num_failed": 0,
  "evaluation_scores": {
    "Macro F1": 0.858,
    "Micro F1": 0.861,
    "Acc": 0.861,
    "Weighted Precision": 0.882,
    "Weighted Recall": 0.861,
    "Weighted F1": 0.858
  }
}
```

LLMeBench Features

- **~800 assets** across **16 languages**
- Extensive support for **reading datasets**
  - HuggingFace datasets + generic data loaders (csv, tsv, json)
  - Over 50 dataset-specific loaders
  - Automatic downloading of data (when allowed)
- Supports popular **task types** (Classification, regression etc.)
- Supports popular **model providers** (OpenAI, FastChat, Petals, HuggingFace Inference API)
- Extensive caching
- **Extensible and Plug-and-play!**
  - Easily add new datasets, tasks, evaluation metrics and model providers

Language Model Evaluation Harness

A framework to evaluate LLMs on a large number of tasks and datasets

Latest News 📣

- [2025/12] **CLI refactored** with subcommands (`run`, `ls`, `validate`) and YAML config file support via `--config`. See the CLI Reference and Configuration Guide.
- [2025/12] **Lighter install:** Base package no longer includes `transformers` / `torch`. Install model backends separately: `pip install lm_eval[hf]`, `lm_eval[vllm]`, etc.
- [2025/07] Added `think_end_token` arg to `hf` (token/str), `vllm` and `sglang` (str) for stripping CoT reasoning traces from models that support it.
- [2025/03] Added support for steering HF models!
- [2025/02] Added SGLang support!
- [2024/09] We are prototyping allowing users of LM Evaluation Harness to create and evaluate on text+image multimodal input, text output tasks, and have just added the `hf-multimodal` and `vllm-vlm` model types and `mmmu` task as a prototype feature. We welcome users to try out this in-progress feature and stress-test it for themselves, and suggest they check out `lmms-eval`, a wonderful project originally forking off of the lm-evaluation-harness, for a broader range of multimodal tasks, models, and features.
- [2024/07] API model support has been updated and refactored, introducing support for batched and async requests, and making it significantly easier to customize and use for your own purposes. **To run Llama 405B, we recommend using VLLM's OpenAI-compliant API to host the model, and use the `local-completions` model type to evaluate the model.**
- [2024/07] New Open LLM Leaderboard tasks have been added ! You can find them under the leaderboard task group.

https://github.com/EleutherAI/lm-evaluation-harness

A framework for few-shot language model evaluation (Gao et al., 2023)

Language Model Evaluation Harness

Pros

- Does not require explicit prompting
- Evaluation is based on log-likelihood
- Good for fast evaluation of LLMs

Cons

- Evaluation is not based on token(s) to represent candidate answer
- Lack of chat-templates

https://github.com/EleutherAI/lm-evaluation-harness

Gao et al., A framework for few-shot language model evaluation, 2023

LMMs-Eval

LMMs-Eval: Probing Intelligence in the Real World

pypi v0.7.1 | downloads rate limited by upstream service | contributors 184 | closed issues 496 | open issues 26

> We are building the unified evaluation toolkit for frontier models and probing the abilities in real world, shape what we build next.

▶ 🌐 Available in 17 languages

📚 Documentation | 📖 100+ Tasks | 🌟 30+ Models | ⚡ Quickstart

🏠 Homepage | 💬 Discord | 🤝 Contributing

Why `lmms-eval`?

Benchmarks decide what gets built next. A model team that trusts its eval numbers can focus on real improvements instead of chasing noise. But the multimodal evaluation ecosystem is fragmented - scattered datasets, inconsistent post-processing, and single-number accuracy scores that hide whether a gain is real or random. Two teams evaluating the same model on the same benchmark routinely report different results.

We believe better evals lead to better models. Good evaluation maps the border of what models can do and shapes what we build next.

We are building `lmms-eval` and focusing on three core principles:

- **Reproducible** - One pipeline, deterministic results. Same model, same benchmark, same numbers, every time.
- **Efficient** - Evaluation should not be the bottleneck, even at large scale. Async serving, adaptive batching, and video I/O optimizations keep your GPUs saturated end to end.
- **Trustworthy** - Not just accuracy. Confidence intervals, clustered standard errors, paired comparisons, and ongoing research into evaluation methodology. Results you can trust enough to act on.

For how the pipeline works and the concrete mechanisms behind these principles, see How the Evaluation Pipeline Works and Why it's Efficient and Trustworthy.

https://github.com/evolvinglmms-lab/lmms-eval

Zhang et al., LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models, 2024

Rubric-based

omniscore 0.1.0

```
pip install omniscore
```

✓ Latest version

Released: Mar 23, 2026

A Hugging Face-native text scoring package for multi-dimensional quality evaluation.

Navigation

- Project description
- Release history
- Download files

Verified details ✓

These details have been verified by PyPI

Project description

🎯 omniscore

python 3.8+ | 🤗 Hugging Face Models

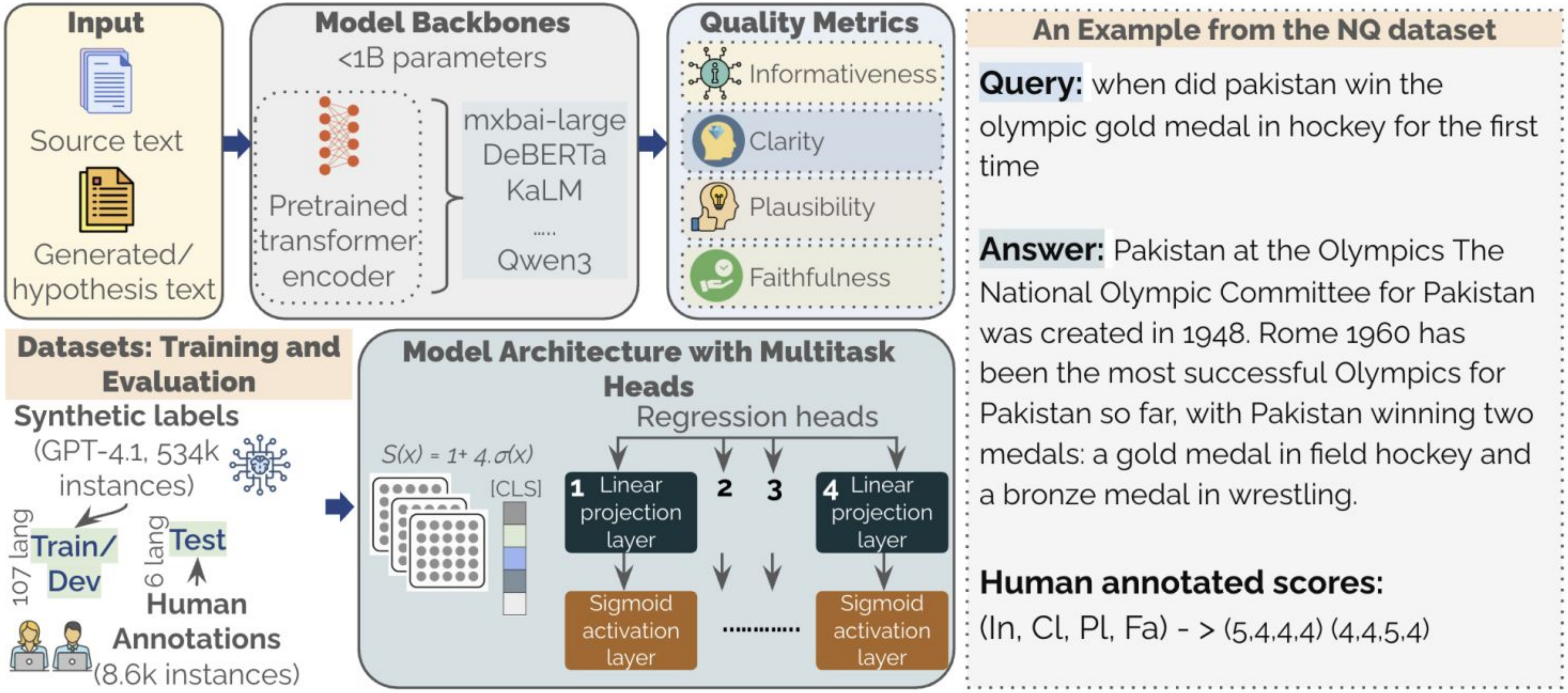
`omniscore` is a lightweight Python package for evaluating the quality of Natural Language Generation (NLG) and generated text. Whether you are evaluating Question Answering (QA), text summarization, explanations, or LLM chat interactions, `omniscore` is designed to integrate seamlessly with OmniScore models.

https://pypi.org/project/omniscore/
https://huggingface.co/datasets/QCRI/OmniScore-Data
https://huggingface.co/collections/QCRI/omniscore

Alam et al., Beyond LLM-as-a-Judge: Deterministic Metrics for Multilingual Generative Text Evaluation, 2026

Rubric-based

Four quality dimensions: informativeness, clarity, plausibility, and faithfulness



Alam et al., Beyond LLM-as-a-Judge: Deterministic Metrics for Multilingual Generative Text Evaluation, 2026

Error/Diagnostic Analysis

Probing what and why an LLM fails - beyond aggregate accuracy

Why diagnostics?

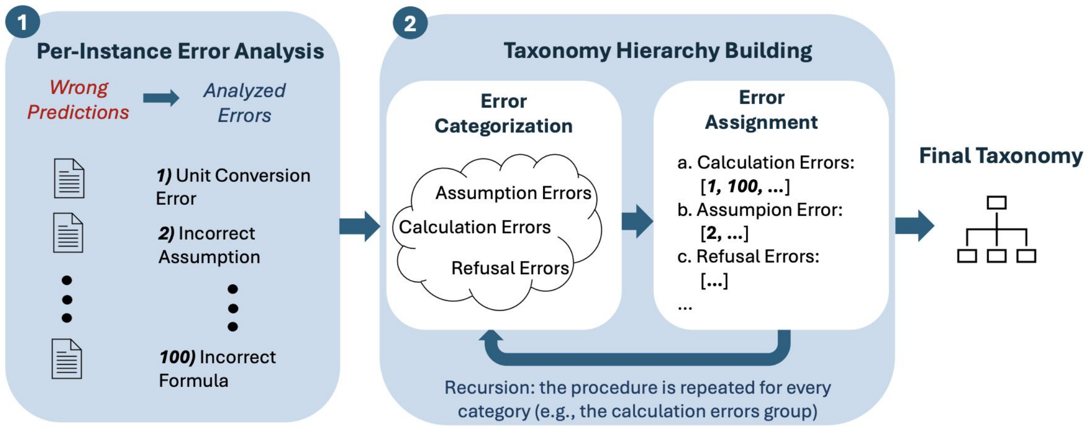
- Static benchmarks report a single score;
- ***diagnostic analysis*** asks
  - *where* the model breaks,
  - *which abilities* the score actually probes,
  - and *which classes of inputs* trigger systematic failure - using error taxonomies, ablation, stress tests and red-teaming.

Error/Diagnostic Analysis

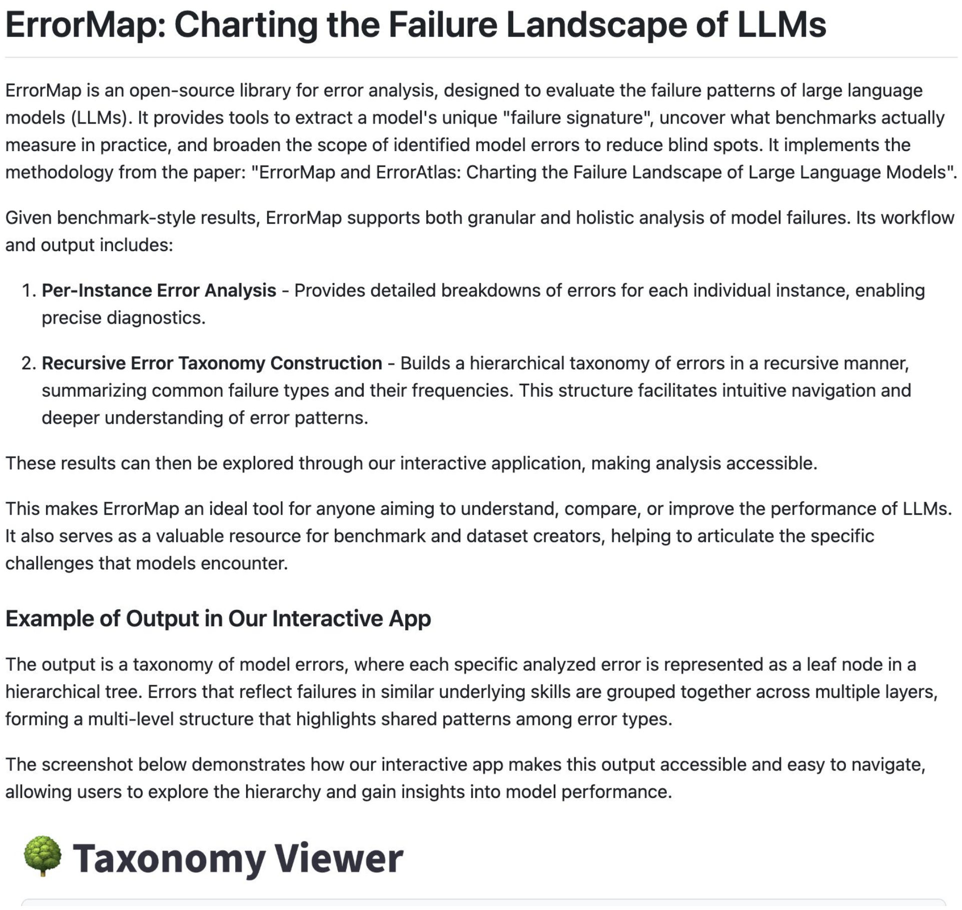
ErrorMap / ErrorAtlas

Charting the failure landscape of LLMs

- Hierarchical taxonomy of 17 top-level error categories.
- Pipeline extracts, labels, and clusters errors into reusable failure signatures - moving beyond aggregate metrics.
- Applied ErrorMap to 83 models and 35 datasets



https://github.com/IBM/ErrorMap



ErrorMap: Charting the Failure Landscape of LLMs

ErrorMap is an open-source library for error analysis, designed to evaluate the failure patterns of large language models (LLMs). It provides tools to extract a model's unique "failure signature", uncover what benchmarks actually measure in practice, and broaden the scope of identified model errors to reduce blind spots. It implements the methodology from the paper: "ErrorMap and ErrorAtlas: Charting the Failure Landscape of Large Language Models".

Given benchmark-style results, ErrorMap supports both granular and holistic analysis of model failures. Its workflow and output includes:

1. **Per-Instance Error Analysis** - Provides detailed breakdowns of errors for each individual instance, enabling precise diagnostics.
2. **Recursive Error Taxonomy Construction** - Builds a hierarchical taxonomy of errors in a recursive manner, summarizing common failure types and their frequencies. This structure facilitates intuitive navigation and deeper understanding of error patterns.

These results can then be explored through our interactive application, making analysis accessible.

This makes ErrorMap an ideal tool for anyone aiming to understand, compare, or improve the performance of LLMs. It also serves as a valuable resource for benchmark and dataset creators, helping to articulate the specific challenges that models encounter.

Example of Output in Our Interactive App

The output is a taxonomy of model errors, where each specific analyzed error is represented as a leaf node in a hierarchical tree. Errors that reflect failures in similar underlying skills are grouped together across multiple layers, forming a multi-level structure that highlights shared patterns among error types.

The screenshot below demonstrates how our interactive app makes this output accessible and easy to navigate, allowing users to explore the hierarchy and gain insights into model performance.

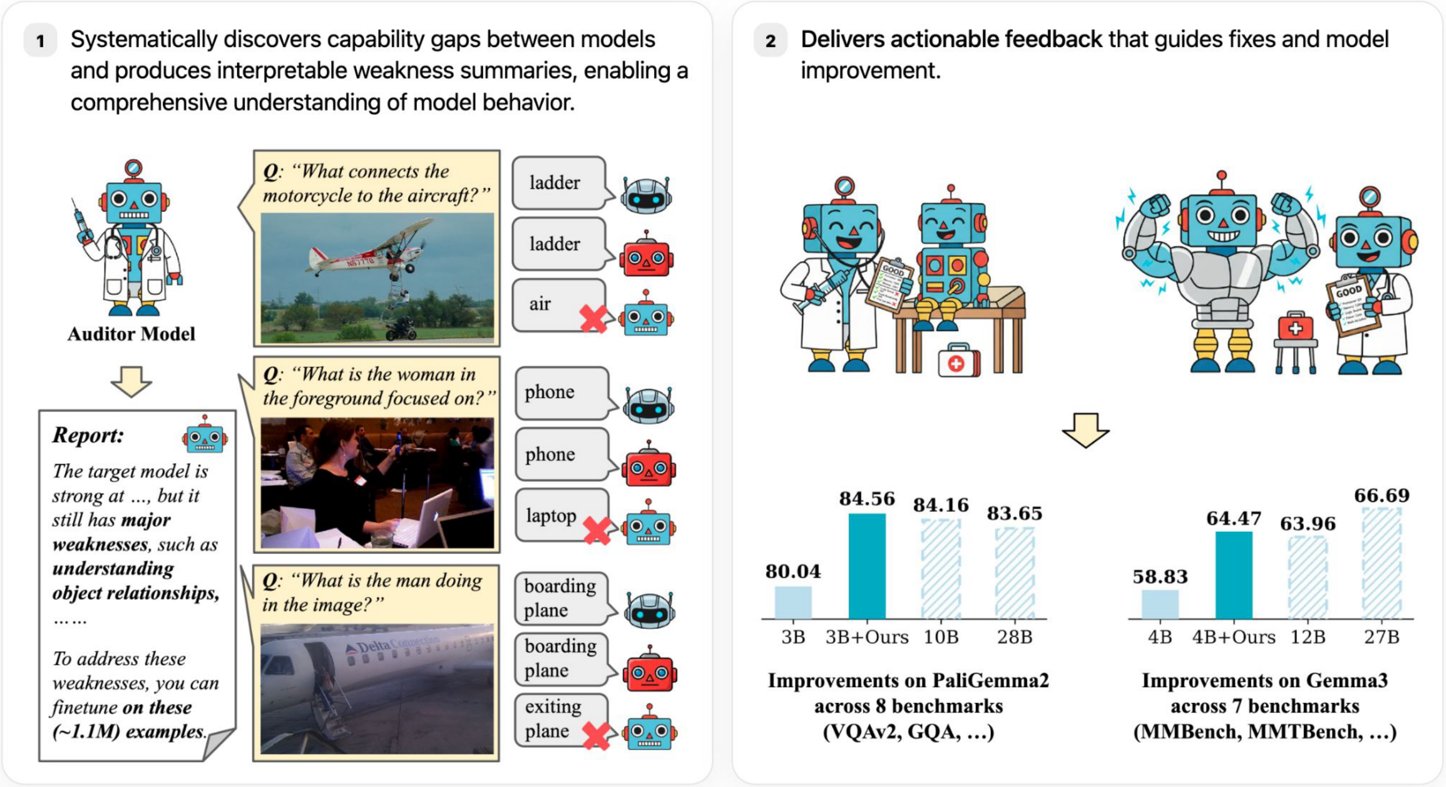
🌳 Taxonomy Viewer

Ashury-Tahan et al., ErrorMap and ErrorAtlas: Charting the Failure Landscape of Large Language Models, 2026

Error/Diagnostic Analysis

AuditDM (Multimodal Auditor)

- Diagnostic framework for MLLM capability gaps
- Fine-tunes an MLLM as an auditor that generates probing questions and counterfactual images on which the target model fails - surfacing capability gaps in a single inference pass.



Liu et al., Differences That Matter: Auditing Models for Capability Gap Discovery and Rectification, 2026

Error/Diagnostic Analysis

Benchmark Profiling

- Mechanistic diagnosis of LLM benchmarks
- Decomposes 10 operationalised abilities (deductive reasoning, contextual recall, ...) and measures benchmark dependence on each via ability-specific parameter ablation.

(Kim et al., 2025, arXiv:2510.01232)

Testing Model Specs

- Character differences across LLMs
- 300K+ value-tradeoff scenarios; 70K with substantial cross-model divergence. Spec violations are 5-13× more frequent in high-disagreement scenarios.

(Zhang et al., 2025, arXiv:2510.07686)

Advertisement: Become Part of Our Team!

Internships, Visiting Faculty, Senior, Principal Scientist and Research Collaboration

Arabic Language Technologies

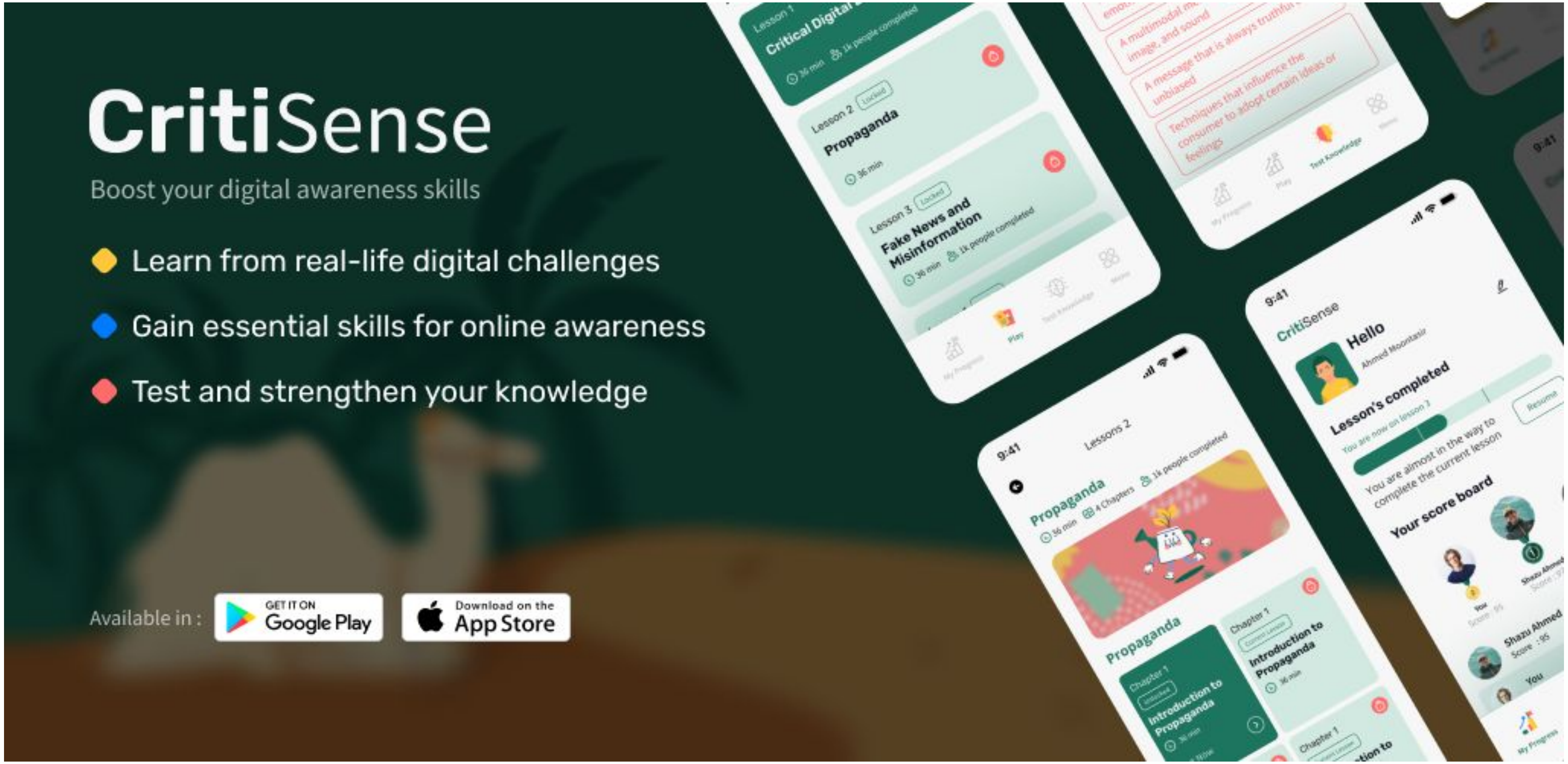
ALT

تقنيات اللغة العربية

https://alt.qcri.org/

Advertisement

CritiSense is mobile app designed to build resilience against fake news and propaganda.



https://critisense-web.digitqr.net

CritiSense

Apple store

Play Store

Alam et al., CritiSense: Critical Digital Literacy and Resilience Against Misinformation, accepted at ACL-demo, 2026

Thank You

Slides and other resources will be available at:

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

QA