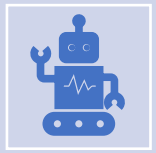
The Development of Spoken LM

Jinyu Li

Speech Processing in the LLM Era



Large Language Models (LLMs) have transformed NLP and are now reshaping speech processing.

LLMs enable unified models for various speech capabilities such as speech recognition, synthesis, translation, and dialogue etc.

Integration of speech and text modalities leads to more natural multimodal systems optimized in an end-to-end fashion.

SLM vs. MLLM

Spoken language model (SLM)

- speech-centric
- backbone is an LM, no need to be an LLM

Multimodal LLM (MLLM)

- at least two modalities
- backbone is an LLM

Lots of SLM/MLLM Surveys in Recent Two Years

In this talk, we mainly share the journey of developing SLM/MLLM models at Microsoft.

On The Landscape of Spoken Language Models: A Comprehensive Survey

Siddhant Arora[1*] Kai-Wei Chang[2*] Chung-Ming Chien[3*] Yifan Peng[1*] Haibin Wu[2*#]
Yossi Adi[4+] Emmanuel Dupoux[5+] Hung-Yi Lee[2+] Karen Livescu[3+] Shinji Watanabe[1+]

Recent Advances in Speech Language Models: A Survey

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King, *Fellow, IEEE*

Audio-Language Models for Audio-Centric Tasks: A survey

Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, Yong Dou

A SURVEY ON SPEECH LARGE LANGUAGE MODELS

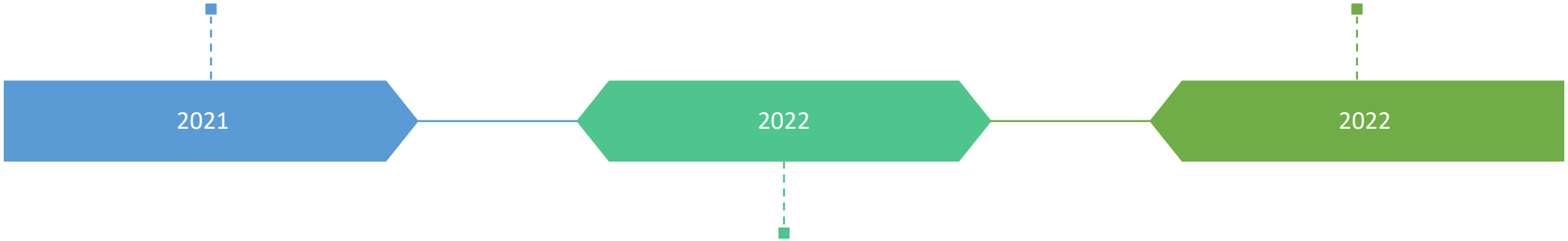
Jing Peng[1], Yucheng Wang[2*], Yu Xi[1], Xu Li[2], Xizhuo Zhang[1], Kai Yu[1†]*

We Predicted SLM as the Future

……used LM in model names since 2021

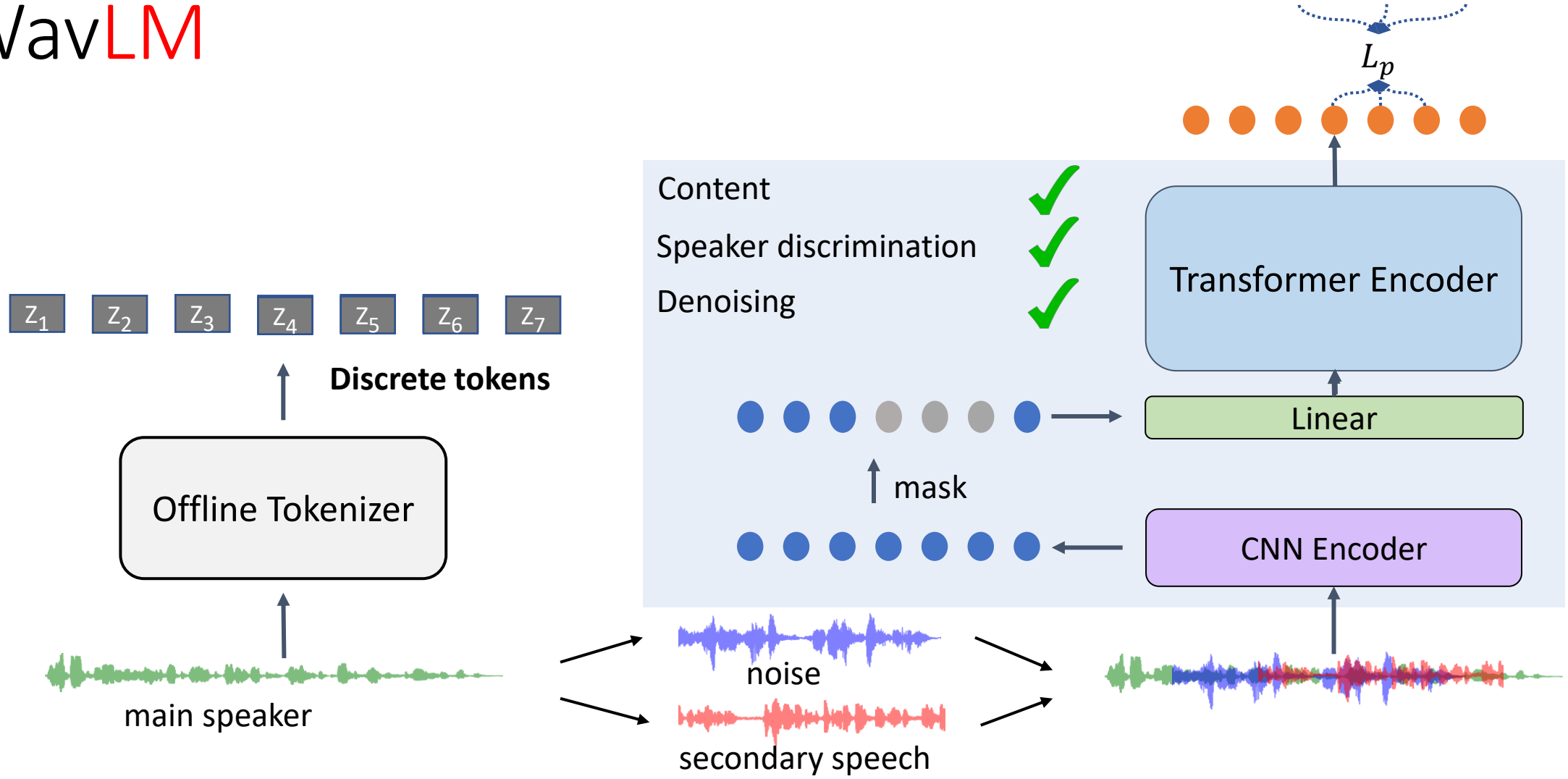
Chen et. al., **WavLM**: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing

Zhu et. al., VATLM: Visual-Audio-Text Pre-Training with Unified Masked Prediction for Speech Representation Learning



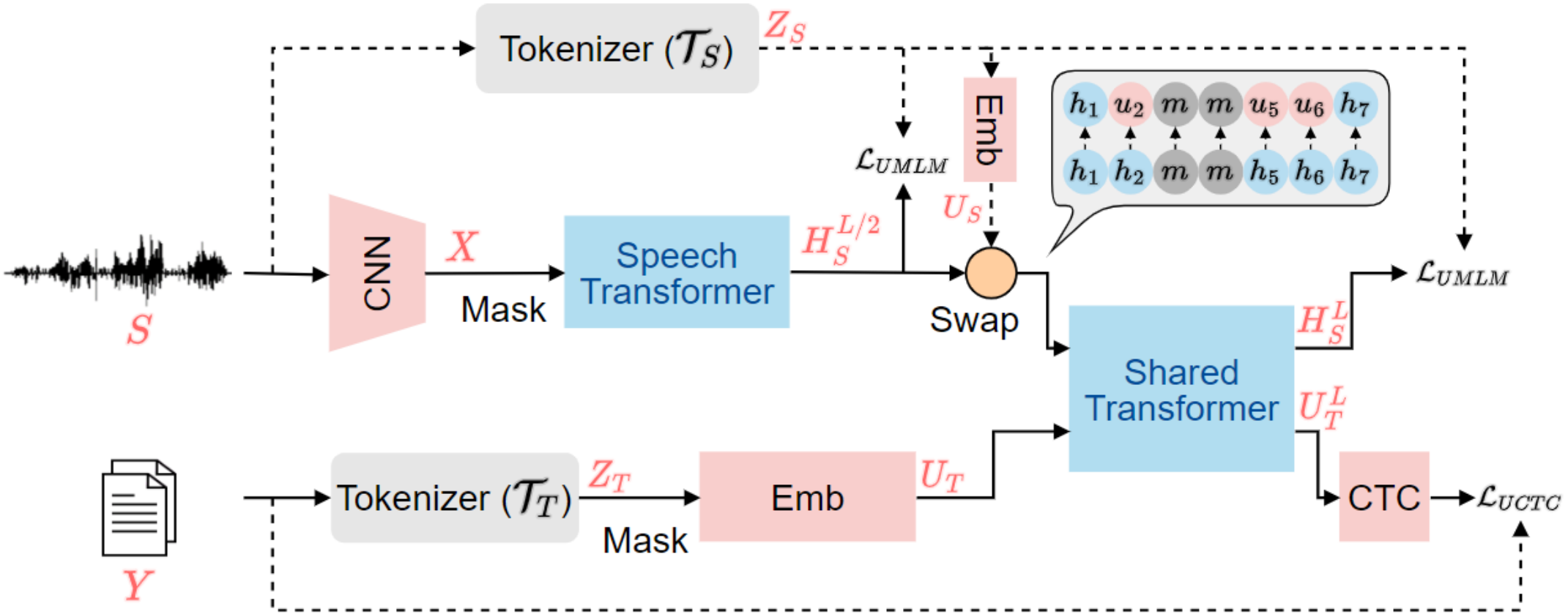
Zhang et. al., SpeechLM: Enhanced Speech Pre-Training with Unpaired Textual Data

WavLM



- 40 million downloads on Hugging Face
- lead the SUPERB (self-supervised learning) leaderboard

SpeechLM


Tokenizer ($\mathcal{T}_S$)
Z_S
Emb
$\mathcal{L}_{UMLM}$
h_1 u_2 m m u_5 u_6 h_7
h_1 h_2 m m h_5 h_6 h_7
U_S
S
CNN
X
Mask
Speech Transformer
$H_S^{L/2}$
Swap
$\mathcal{L}_{UMLM}$
H_S^L
Shared Transformer
U_T^L
Y
Tokenizer ($\mathcal{T}_T$)
Z_T
Mask
Emb
U_T
CTC
$\mathcal{L}_{UCTC}$

Foundation Models

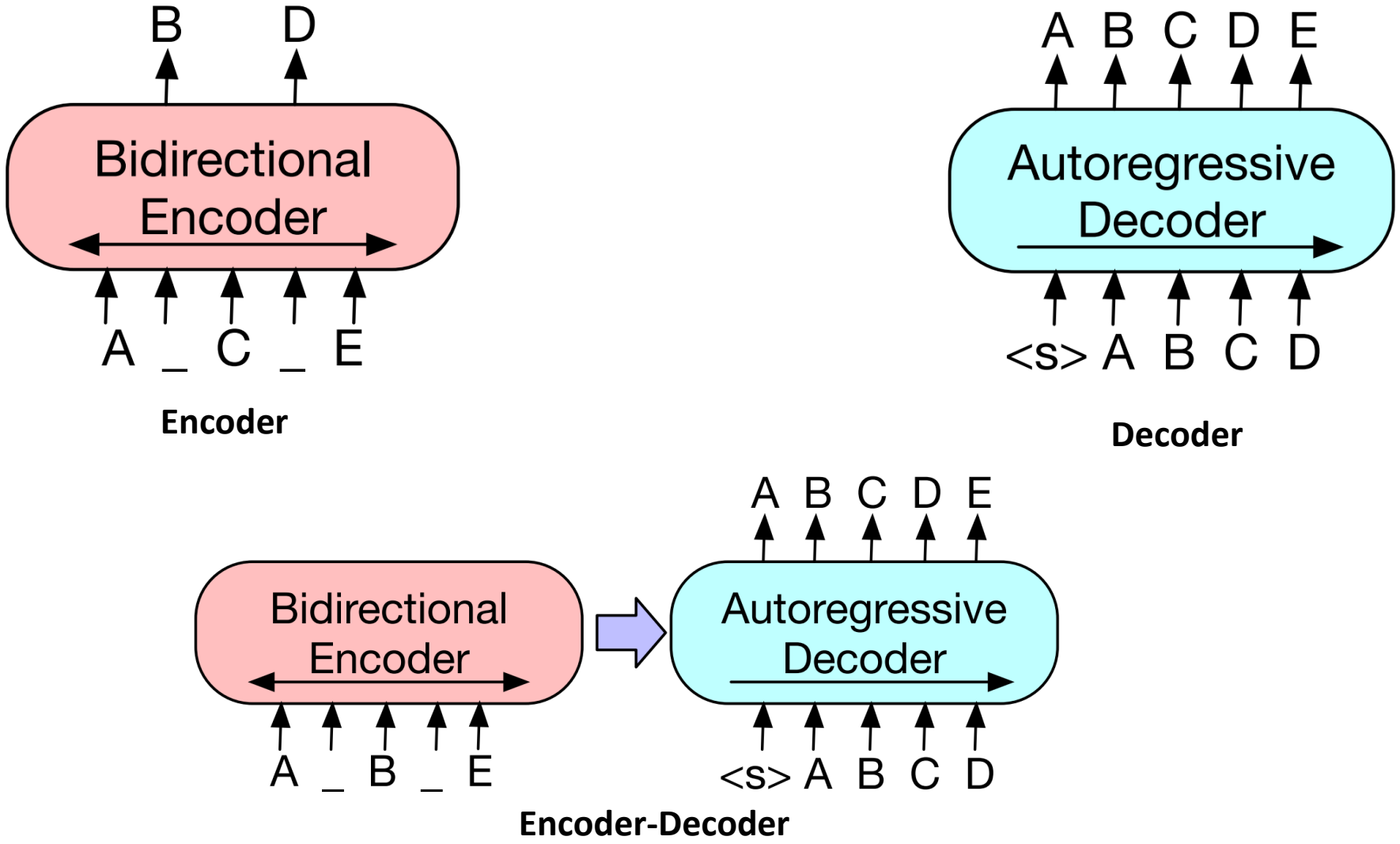


Figure from: Lewis et al., BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. 2020

Decoder-only Is the Trend

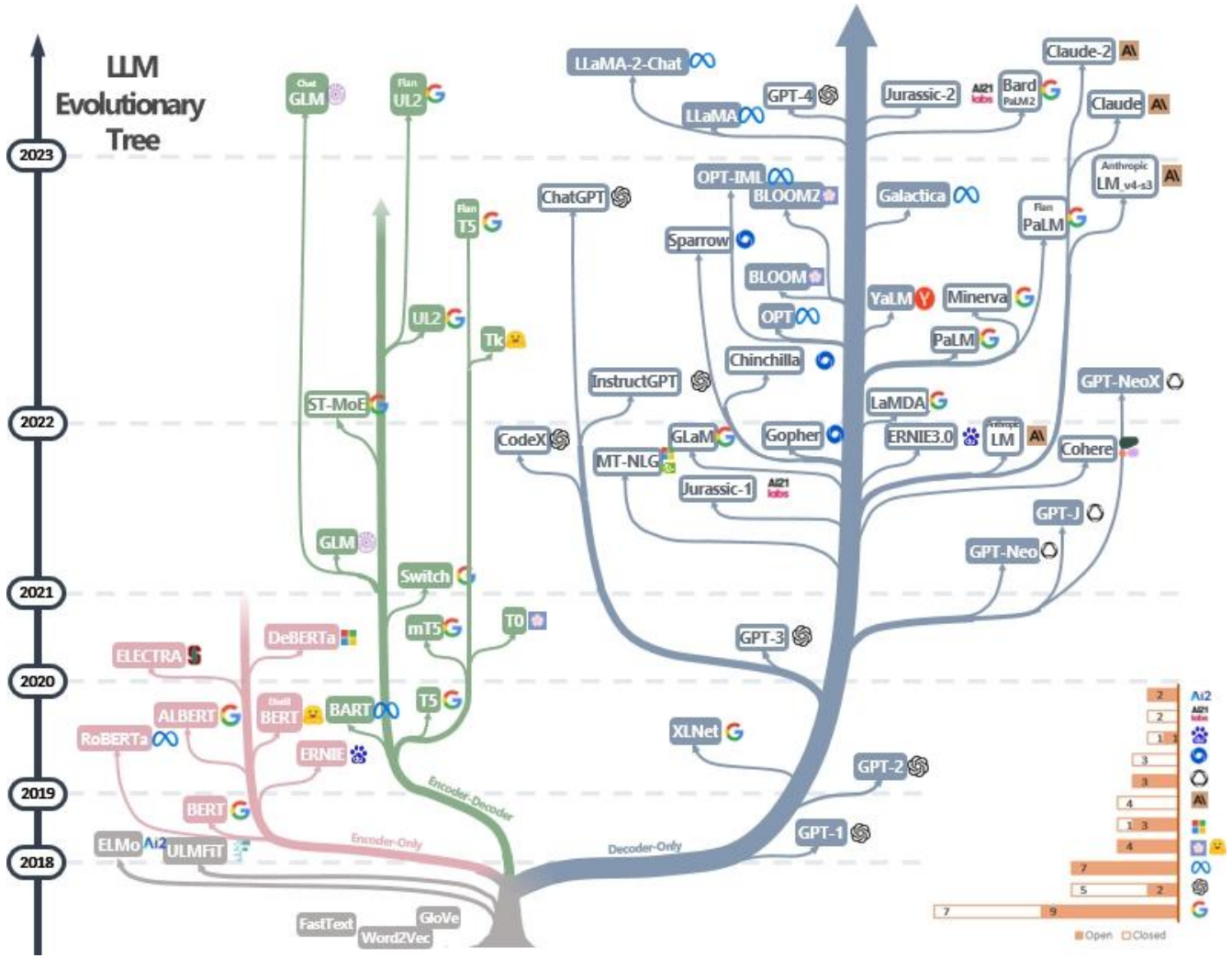
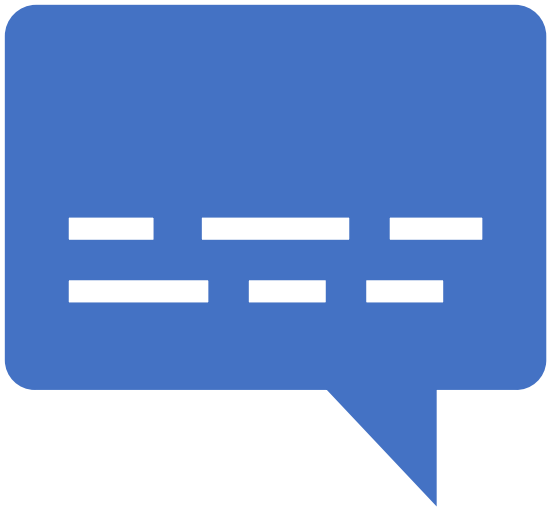


Figure from: https://github.com/Mooler0410/LLMsPracticalGuide

Model Comparison

Feature	Decoder-Only	Encoder-Decoder	Encoder-Only
Generative Power	✅ High	✅ High	❌ Limited
Parameter Efficiency	✅	❌	✅
Training Simplicity	✅	❌	✅
In-Context Learning	✅	❌	❌
Multimodal Extension	✅ Emerging	❓	❌ Limited

Spoken LM (SLM)

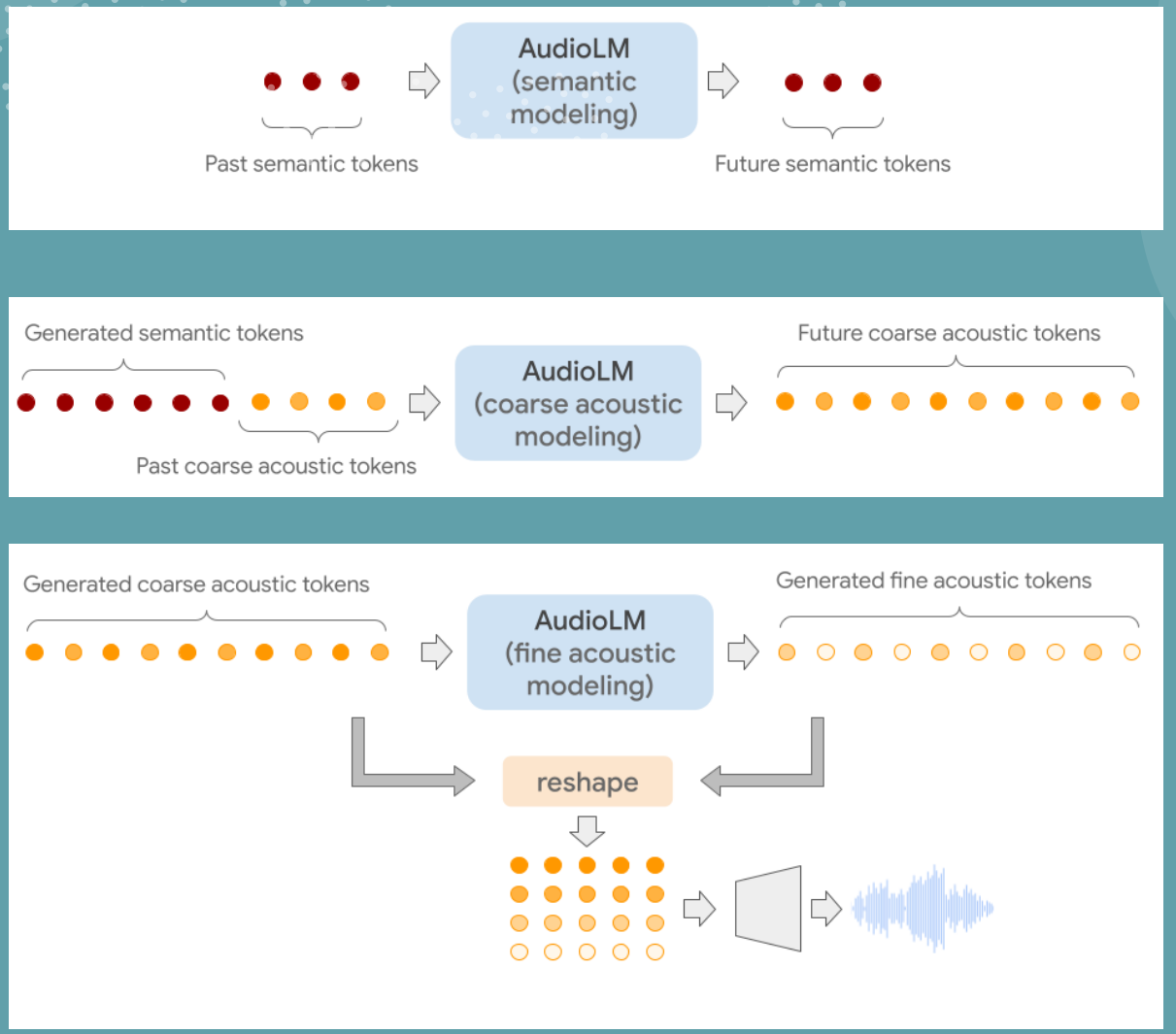


Extend text-LM to SLM

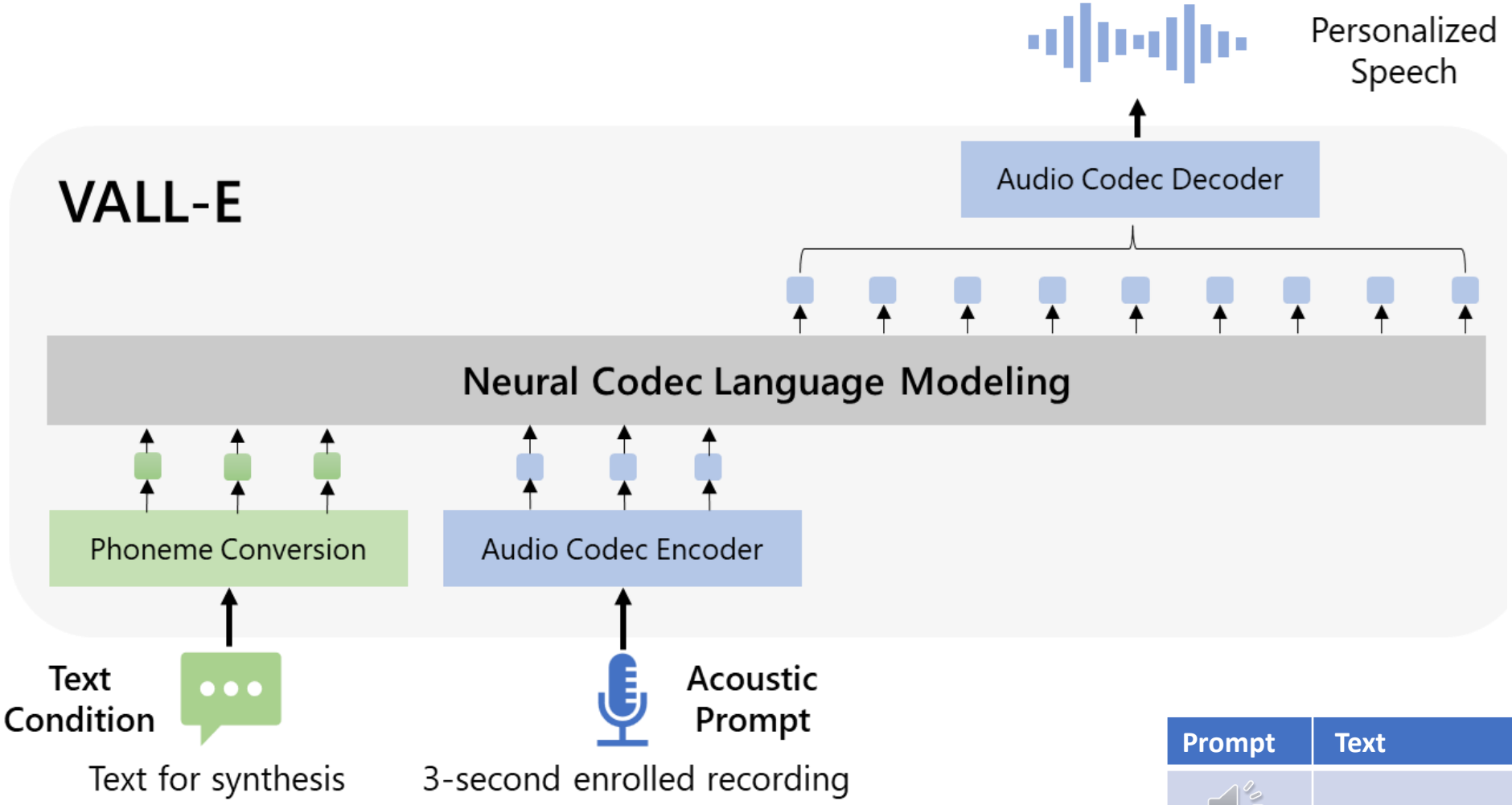
- Discrete representations of Speech
- Transformer LM

AudioLM

- Multistep token generation
  - semantic tokens -> coarse acoustic tokens-> fine acoustic tokens
- Audio continuation task
- Google work



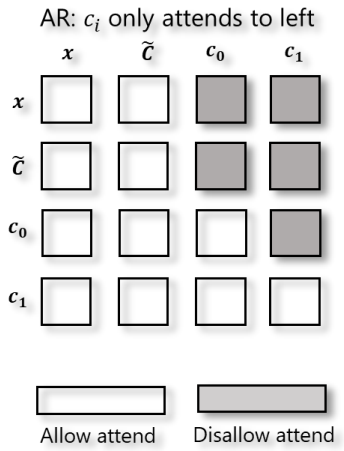
Borsos et al., **AudioLM: a Language Modeling Approach to Audio Generation**. 2022.
Figures from: https://research.google/blog/audiolm-a-language-modeling-approach-to-audio-generation/

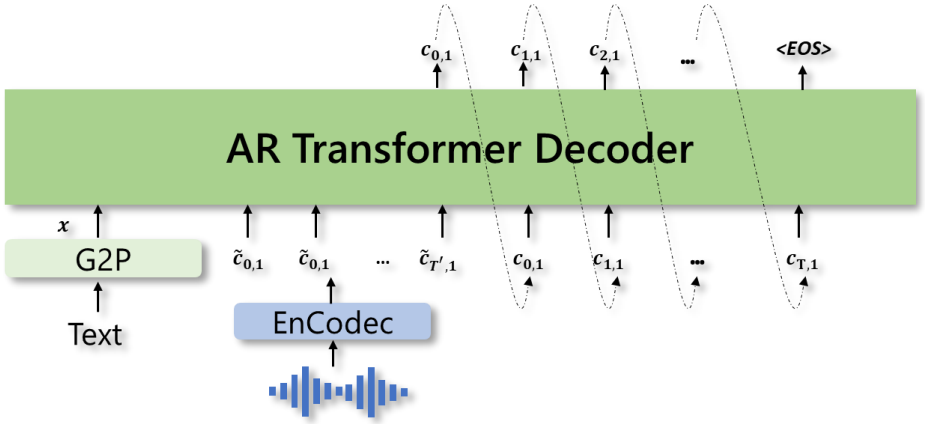


Prompt	Text	Output
	I like hamburger but I love noodles much more	

Wang et al., **Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers**. 2023

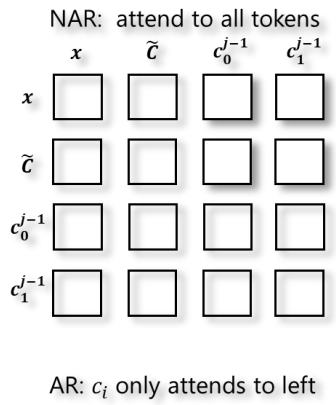
VALL-E

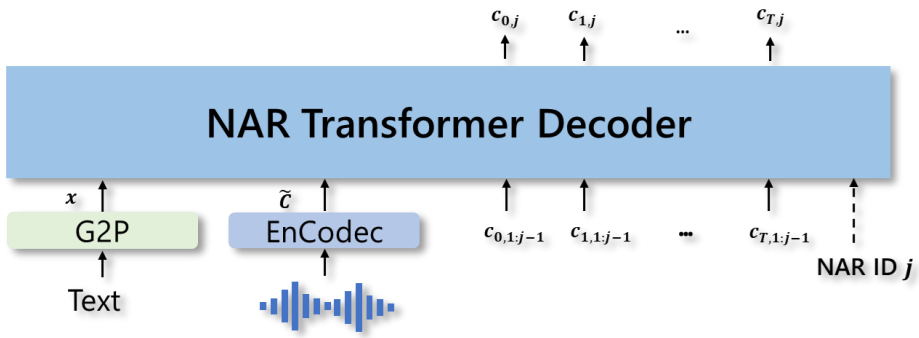




AR Model for Coarse Generation

- Causal mask is used for model training
- Only the left context is considered to generate the current output
- Generate the codec tokens in the first layer one by one





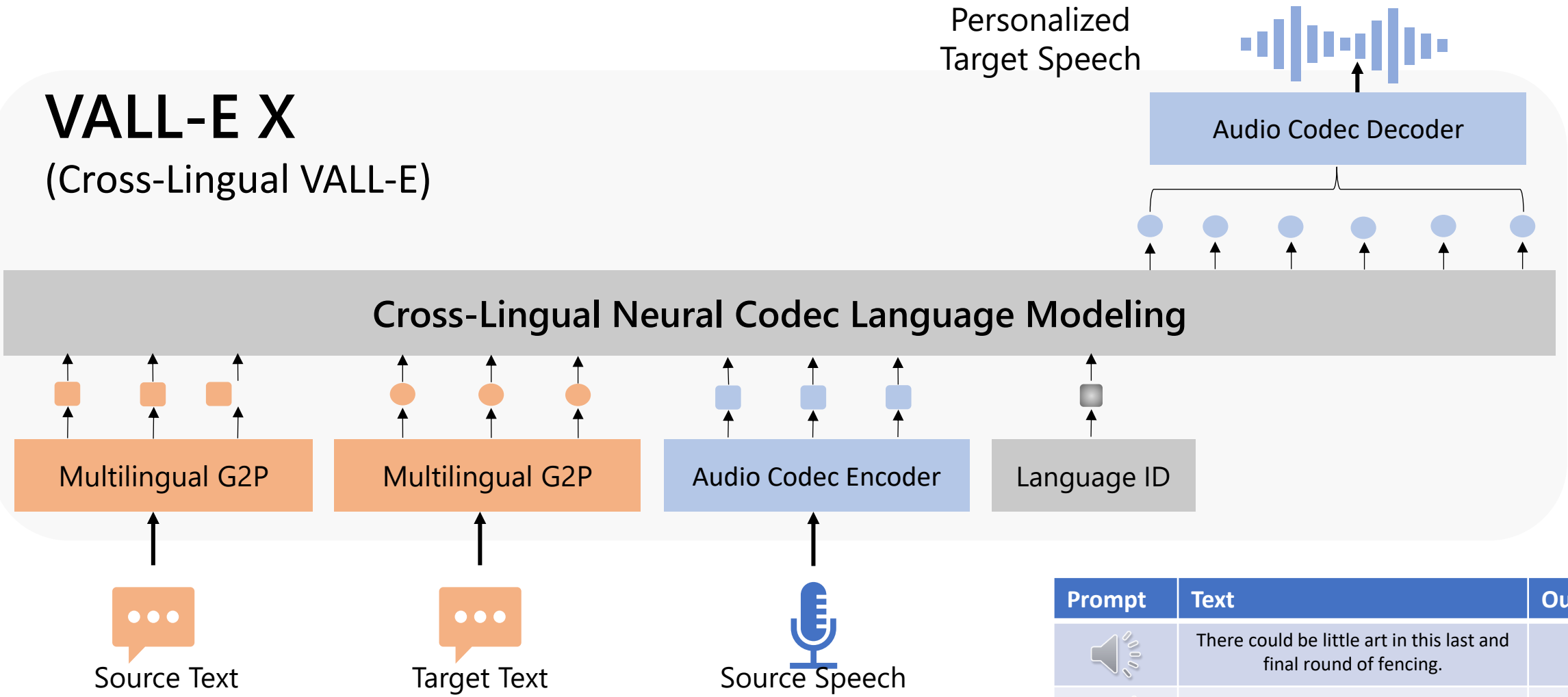
NAR Model for Fine Generation

- Use the previous layer codec codes to predict the codes in the current layer
- No mask for the input during model training
- All the codec codes can be predicted simultaneously

Wang et al., **Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers**. 2023

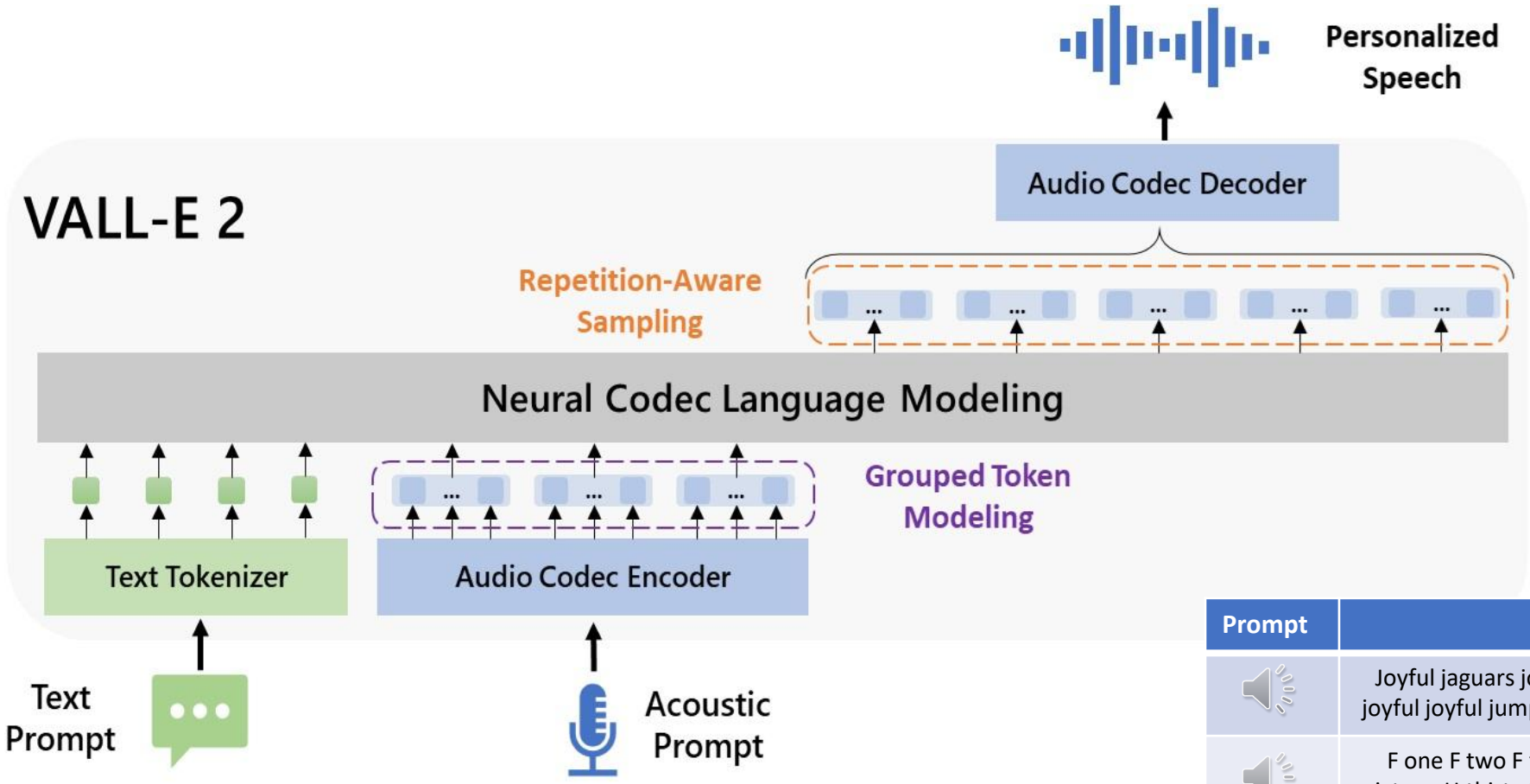
VALL-E X

(Cross-Lingual VALL-E)



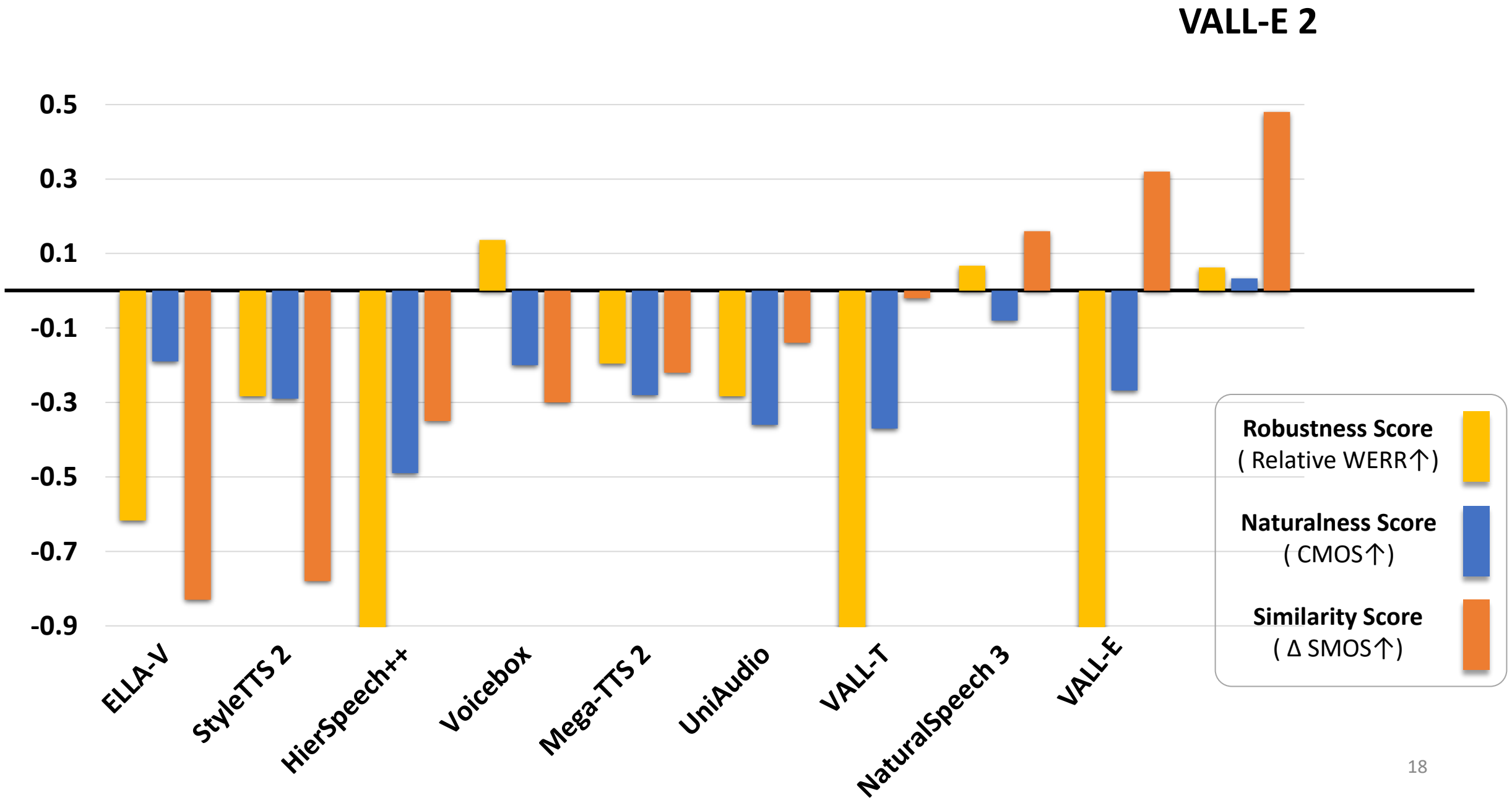
Prompt	Text	Output
	There could be little art in this last and final round of fencing.	
	It was youth and poverty and proximity and everything was young and kindly.	

Zhang et al., **Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling**. 2023



Prompt		VALL-E	VALL-E 2
	Joyful jaguars joyfully jumped joyful joyful jumps		
	F one F two F four F eight H sixteen H thirty two H sixty four		

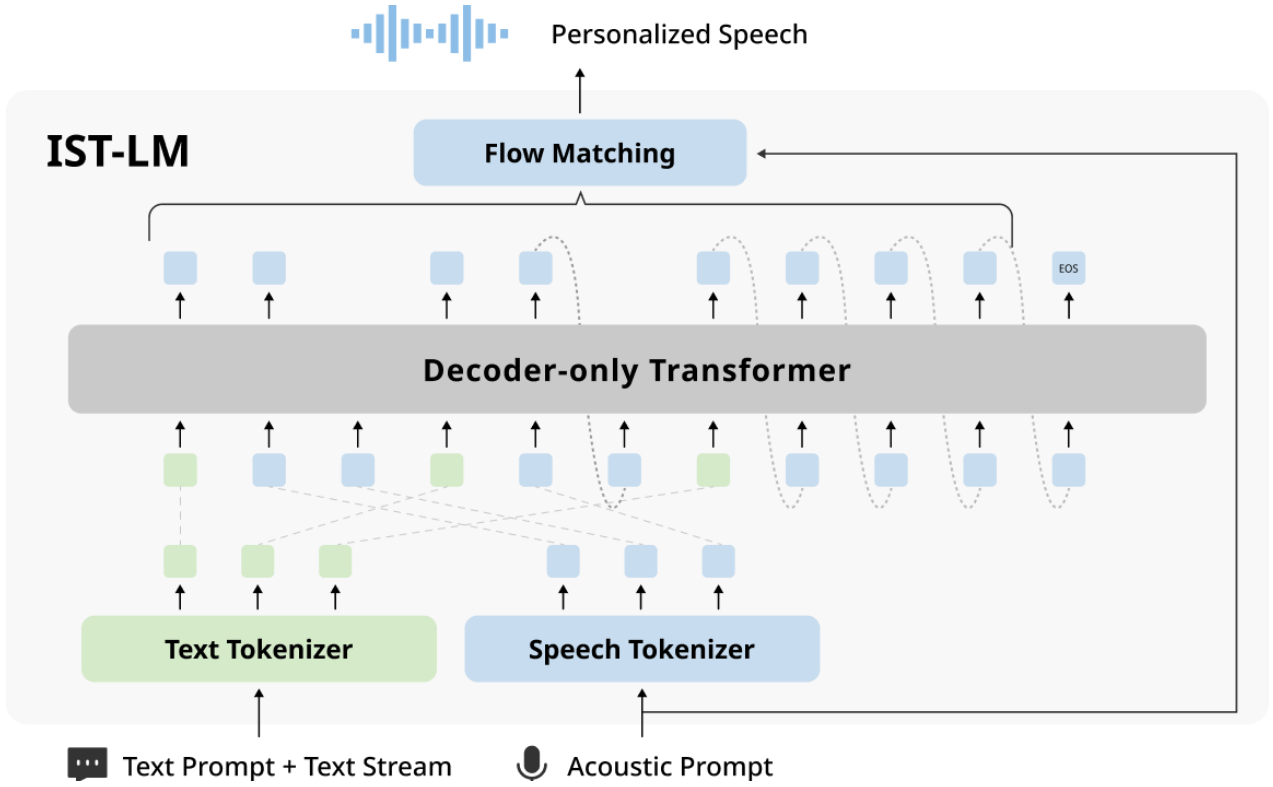
Chen et al., **VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers**. 2024


VALL-E 2
0.5
0.3
0.1
-0.1
-0.3
-0.5
-0.7
-0.9
ELLA-V
StyleTTS 2
HierSpeech++
Voicebox
Mega-TTS 2
UniAudio
VALL-T
NaturalSpeech 3
VALL-E
Robustness Score
(Relative WERR↑)
Naturalness Score
(CMOS↑)
Similarity Score
(Δ SMOS↑)

Streaming TTS

Offline TTS + Chunking Text	• Rely on complex rule-based segmentation and engineering optimization • Inconsistencies in speech across chunks
Interleave text and corresponding speech tokens	• Rely on forced alignment • High computational overhead / hard for scalability
Interleave text and speech tokens at a fixed ratio	• LMs generate text at a constant rate • Can speech be synthesized in parallel with LM-generated text at a fixed ratio?

Interleaved Speech-Text Language Model

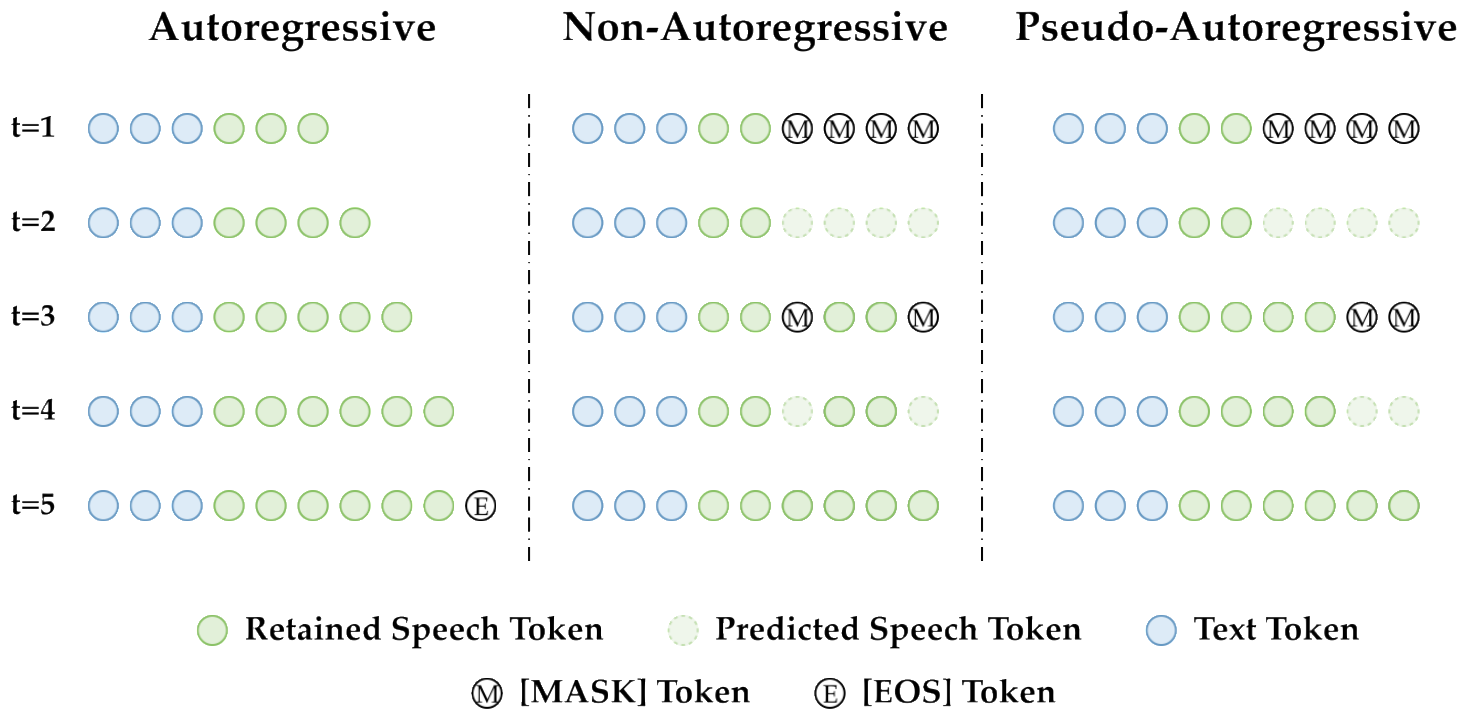


A decoder-only LM modeling interleaved sequence of speech and text tokens with a fixed ratio (1:2 for illustration)

Yang, et al., **Interleaved Speech-Text Language Models are Simple Streaming Text to Speech Synthesizers**. 2024.

PAR: A Novel Language Modeling Approach that Unifies AR and NAR

- **AR:** temporal modeling, but slow generation
- **NAR:** parallel generation, but lack temporal modeling
- **PAR:** combining explicit temporal modeling from AR with parallel generation from NAR -- predicts all masked positions in parallel but commits only the leftmost span at each step



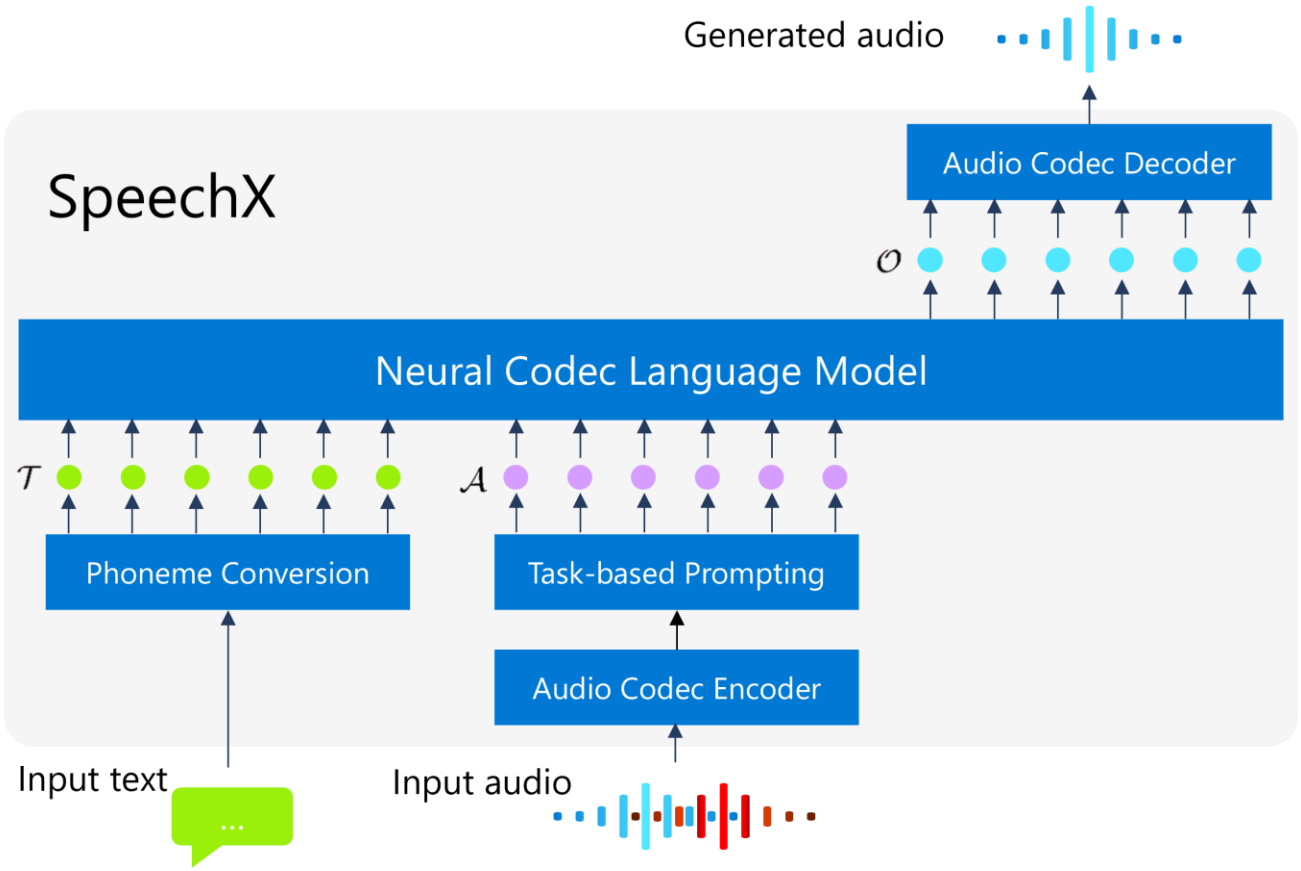
Yang, et al., **Pseudo-Autoregressive Neural Codec Language Models for Efficient Zero-Shot Text-to-Speech Synthesis**. 2025.

SpeechX – A versatile speech generation model

Versatility: able to handle a wide range of tasks from audio and text inputs.

Robustness: applicable in various acoustic distortions, especially in real-world scenarios where background sounds are prevalent.

Extensibility: flexible architectures, allowing for seamless extensions of task support.



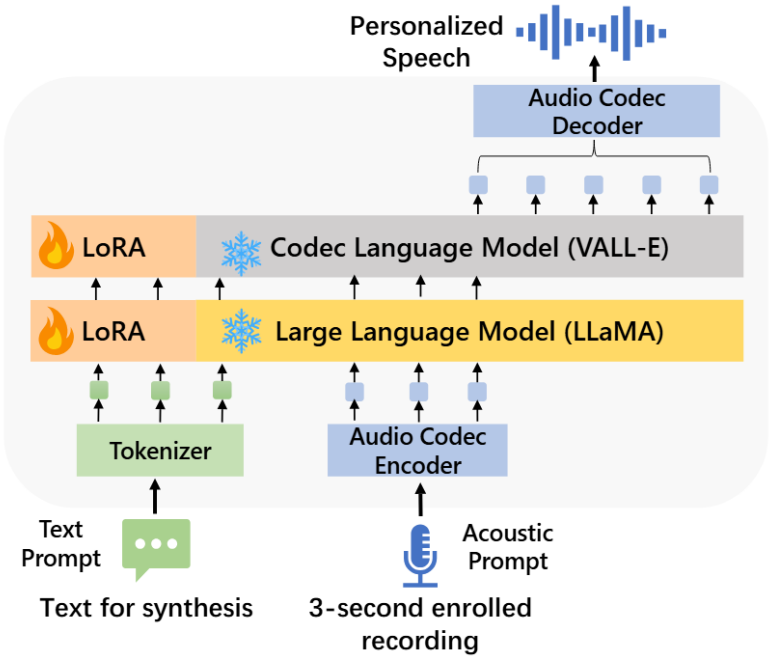
Task	Input text	Input audio	Output audio
Noise suppression	Transcription (optional)	Noisy speech	Clean speech
Speech removal	Transcription (optional)	Noisy speech	Noise
Target speaker extraction	Transcription (optional)	Speech mixture, Enrollment speech	Clean speech of target speaker
Zero-short TTS	Text for synthesis	Enrollment speech	Synthesized speech mimicking target speaker
Clean speech editing	Edited transcription	Clean speech	Edited speech
Noisy speech editing	Edited transcription	Noisy speech	Edited speech with original background noise

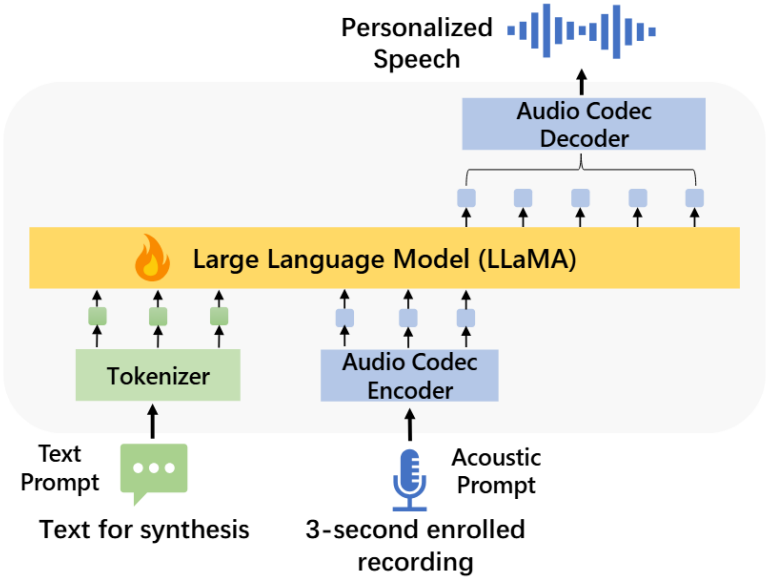
More demo samples: SpeechX - Microsoft Research

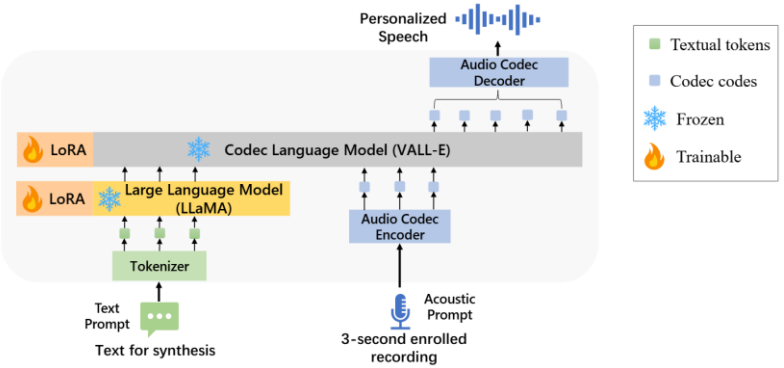
Wang et al., **Speechx: Neural codec language model as a versatile speech transformer**, 2023.

LLM Gains for SLM?

No clear gain by applying LLM to VALL-E







Hao et al., **Boosting large language model for speech synthesis: An empirical study**, 2024.

Multimodal LLM (MLLM)

Why MLLM?

- Zero-shot or few-shot generalization from LLM
- Flexibility of task prompting from LLM
- Strong text capabilities from LLM
- Community support of LLM inference

Build MLLM with Speech Capabilities

LLM can have a mouth: speech generation tasks.

LLM can have ears: speech understanding tasks

LLM is the brain - Keep LLM text capabilities

Speech Representation to LLM

Discrete

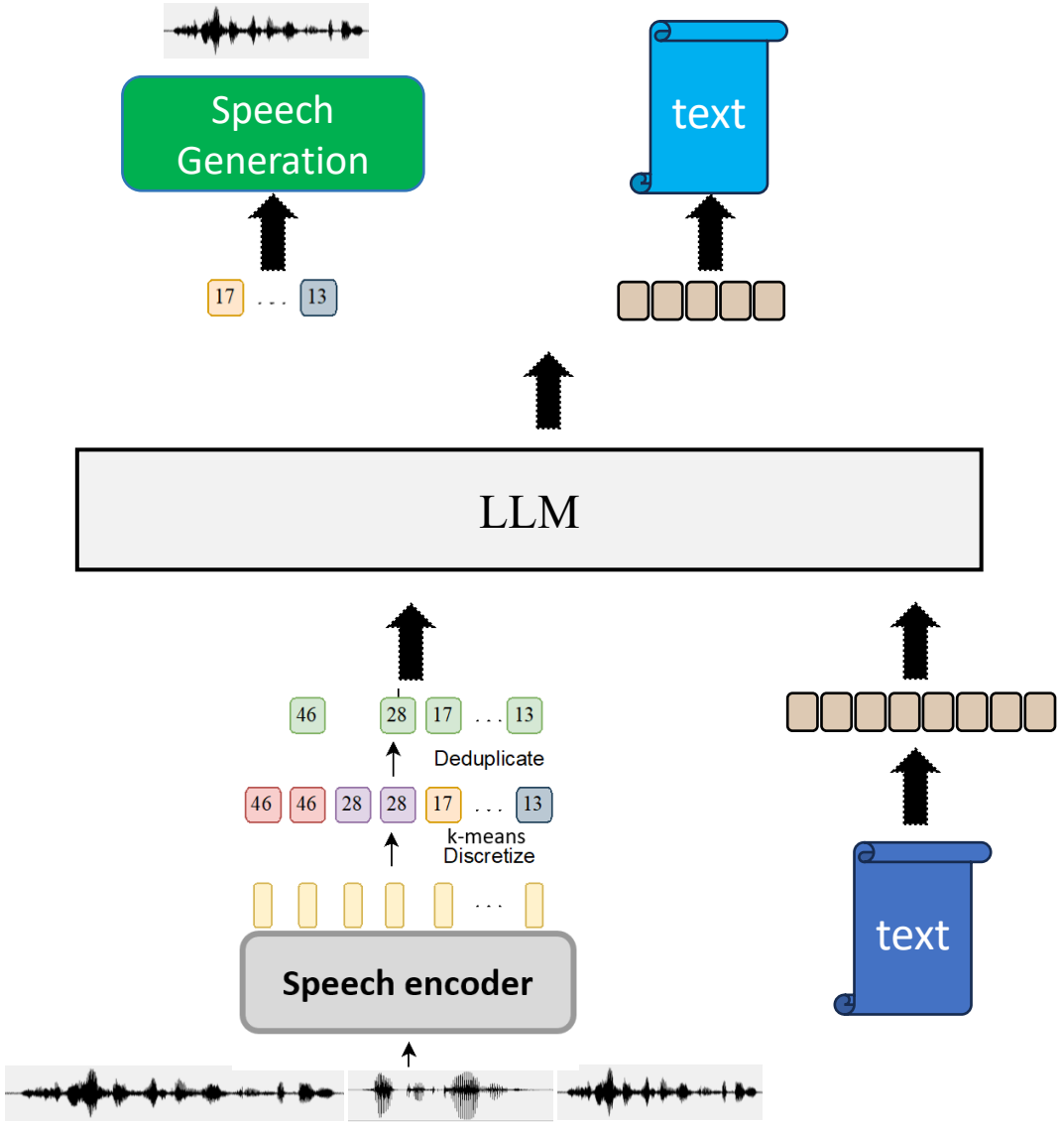
More aligned with LLM which was trained with discrete text tokens.

Continuous

Represent speech better to avoid information loss due to quantization.

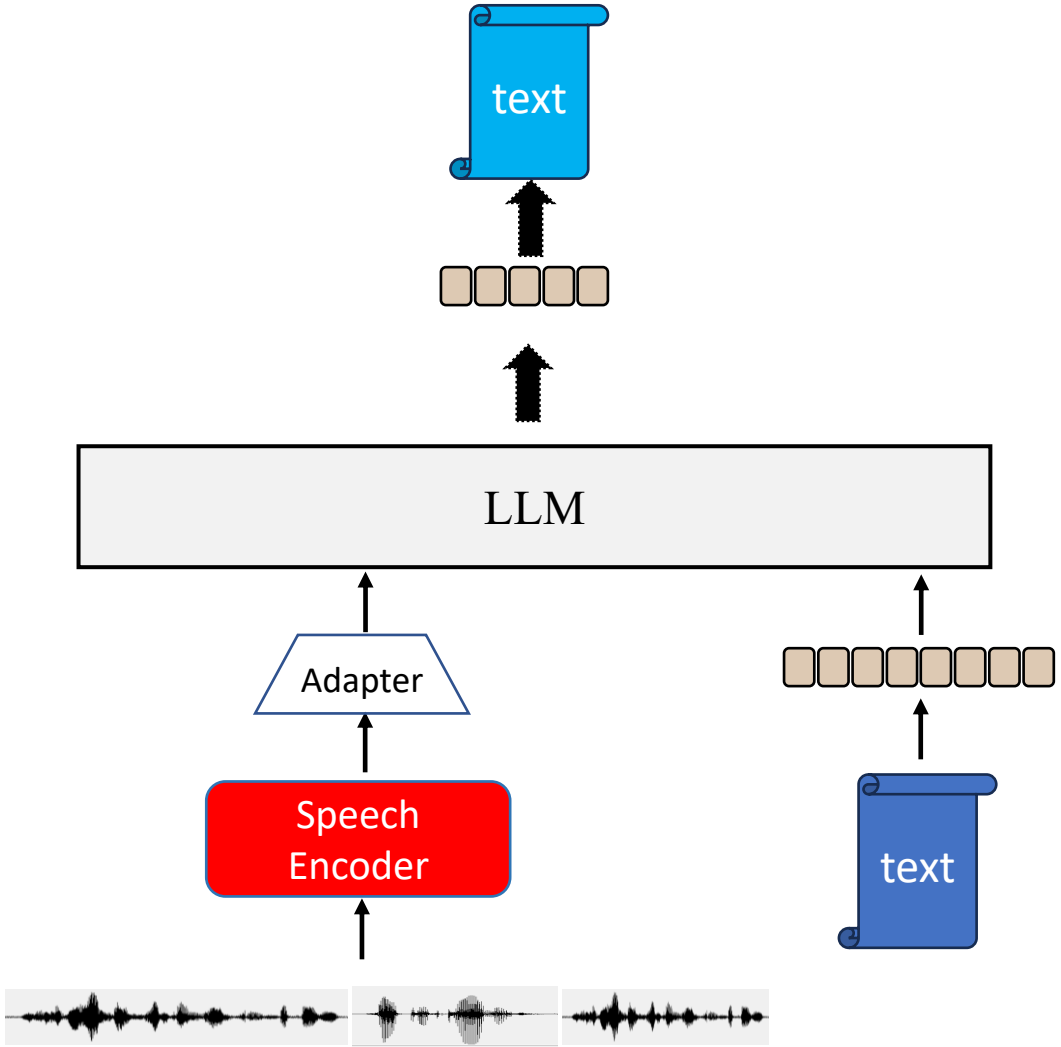
SLM – Discrete Input

- Discrete tokens for both speech and text, better modeling with LLM
- Information loss due to quantization

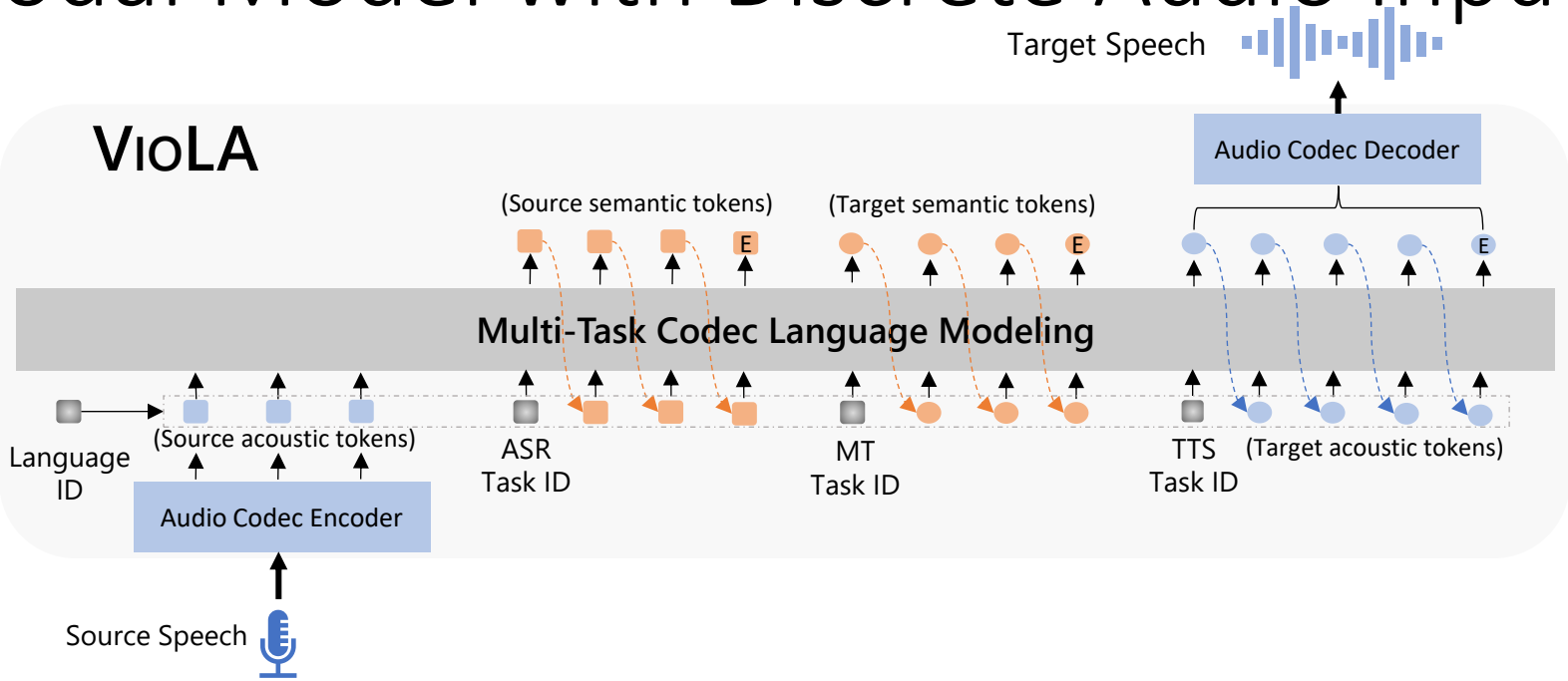


SLM – Continuous Input

- Easily align speech modality with text
- More efforts for speech output



Multi-modal Model with Discrete Audio Inputs



Input	Output	Typical Tasks
Speech	Text	ASR, ST
Text	Text	MT, LM
Text	Speech	multilingual TTS

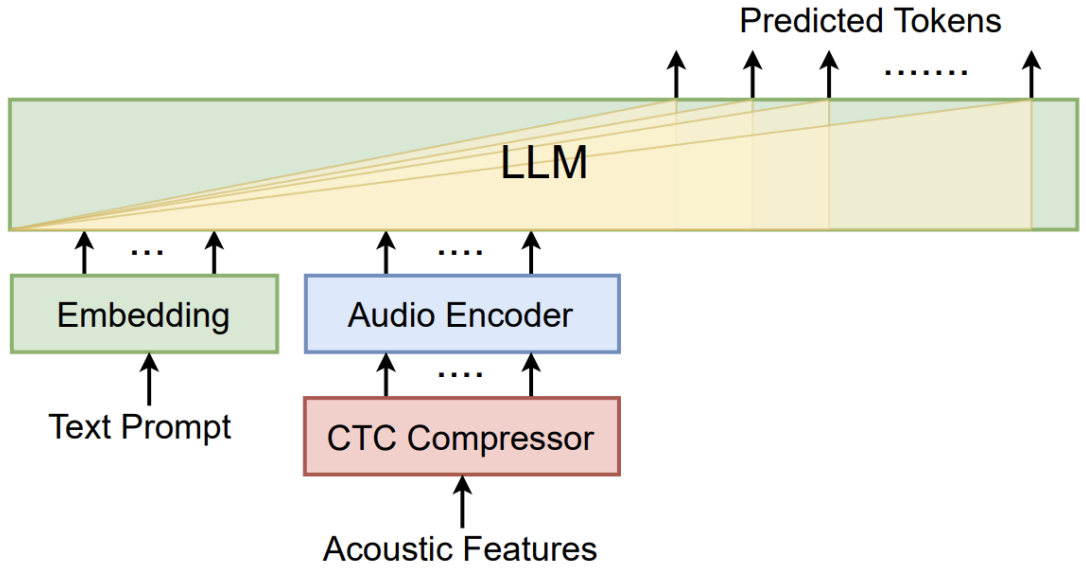
Feature	PER
Fbank	9.61
Codec	12.83

Big performance gap between Fbank and Codec input for speech understanding tasks

Wang, et al., **VioLA: Unified codec language models for speech recognition, synthesis, and translation**, 2023.

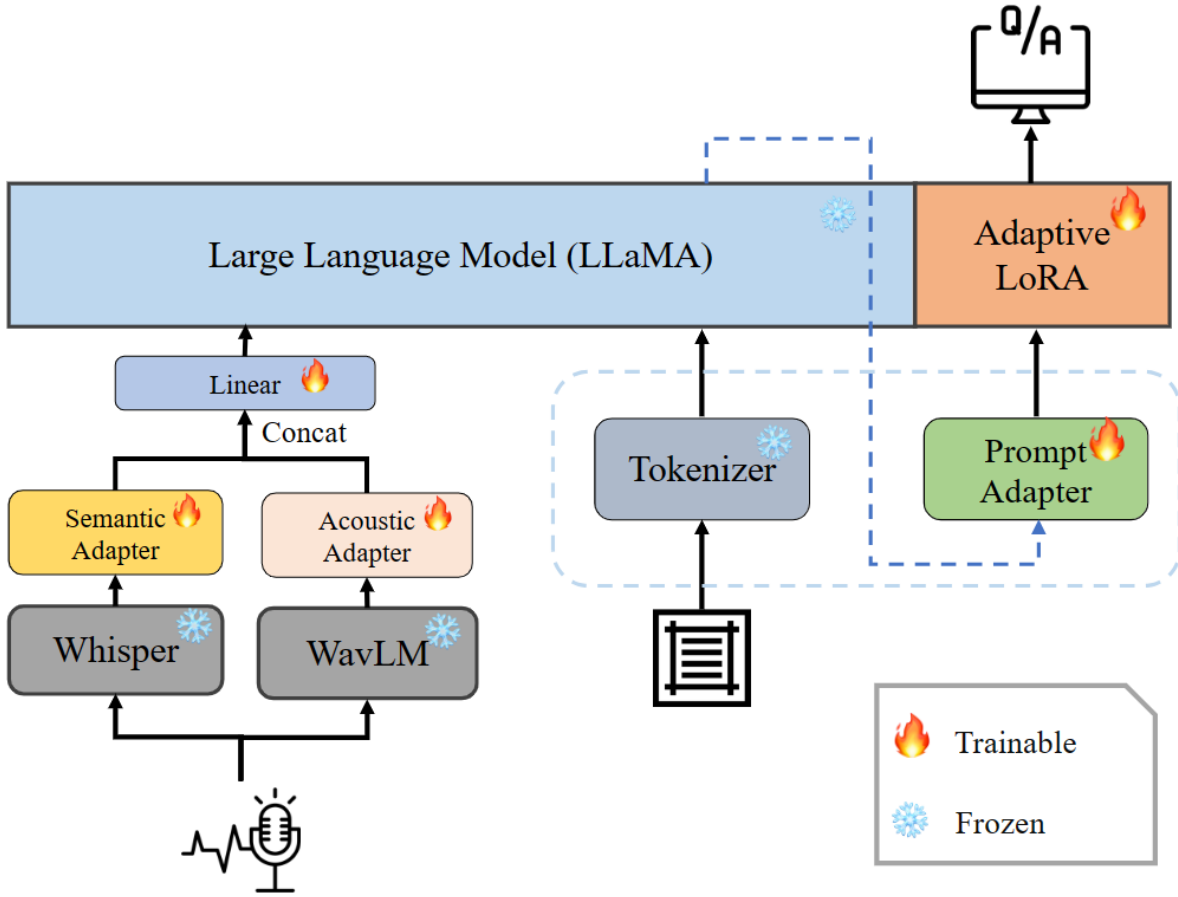
Speech-LLaMA

- CTC Compressor
  - Reduce the acoustic feature length
  - Pretrained on 14 language ASR/AST task
  - Remove blank frames or average frames within same unit
- Audio Encoder
  - Few transformer layers to further process the CTC compressor output
  - Learn the shared representations within the space of the LLaMA embeddings



Wu, et al., **On decoder-only architecture for speech-to-text and large language model integration**, 2023.

Multi-modal Model with Continuous Audio Inputs: WavLLM



Hu, et al., **WavLLM: Towards robust and adaptive speech large language model**, *2024*.

Phi4-mini-MM

Unified Multi-Modality Support (text, speech, and image)

Remarkable Language Performance for the size

Superior Multi-Modal Capabilities for the size

Exceptional Speech and Audio Performance

Open sourced in Mar, 2025

Abouelenin, et al., **Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs**, 2025.

Model Architecture – Phi4-mini-MM

- Tasks: Vision-language, Speech-language, Vision-speech (context - query), spoken-query
- Vision modality: Vision Encoder + projector + vision LoRA
- Audio modality: Audio Encoder + projector + audio LoRA
- **Modality-specific routers** for inference:
  - Language: w/o LoRA
  - Vision-language and Vision-speech: vision LoRA
  - Speech-language, spoken-query: audio LoRA

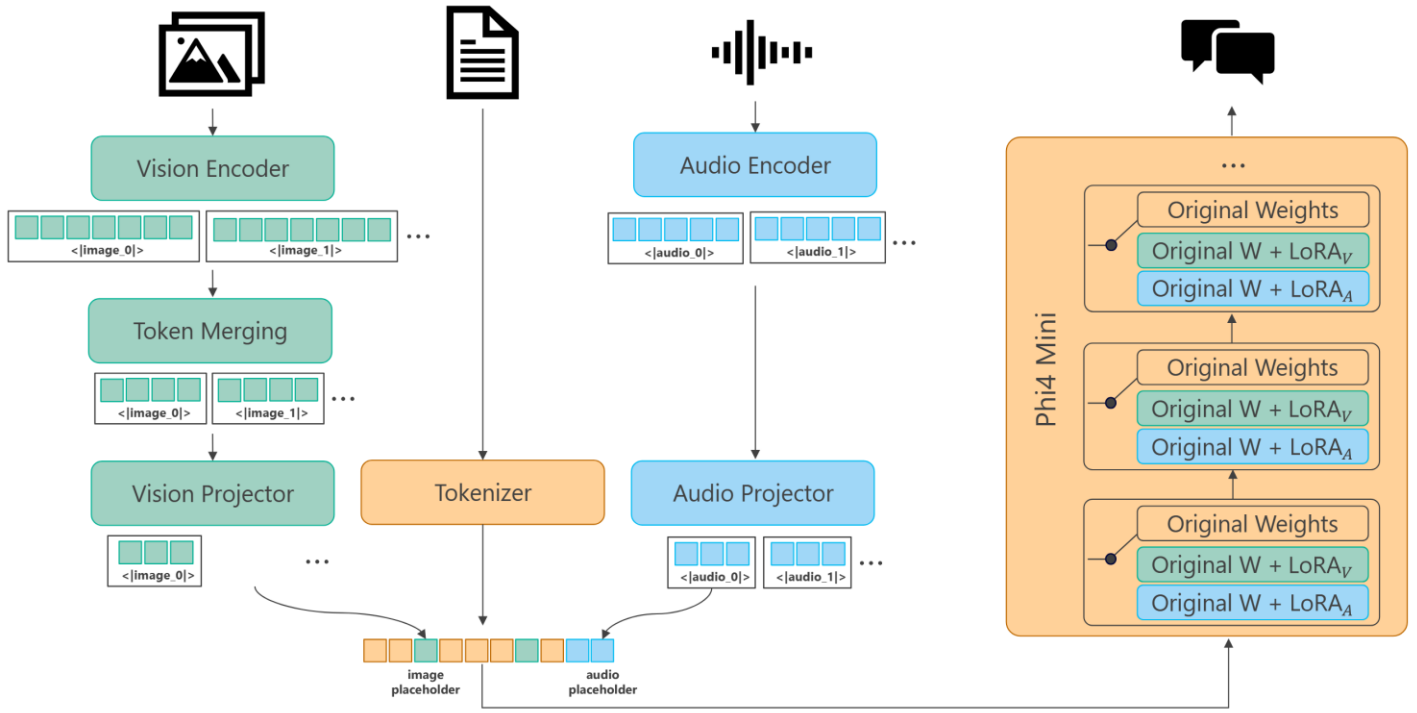
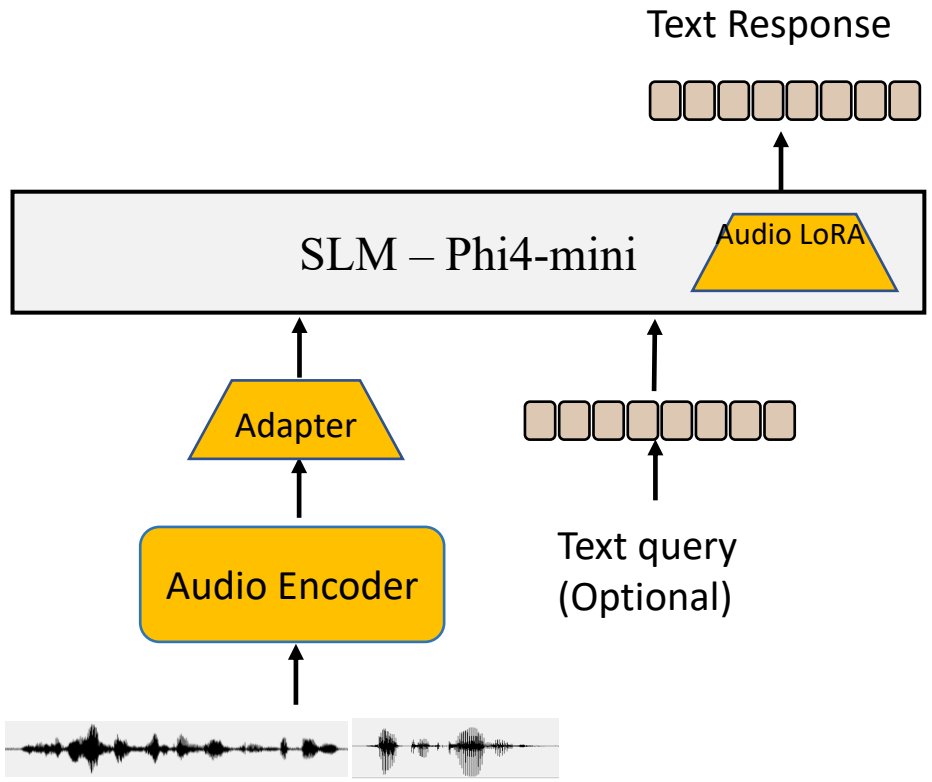


Figure 1: A overview of the Multimodal architecture for `Phi-4-Mini-MM`

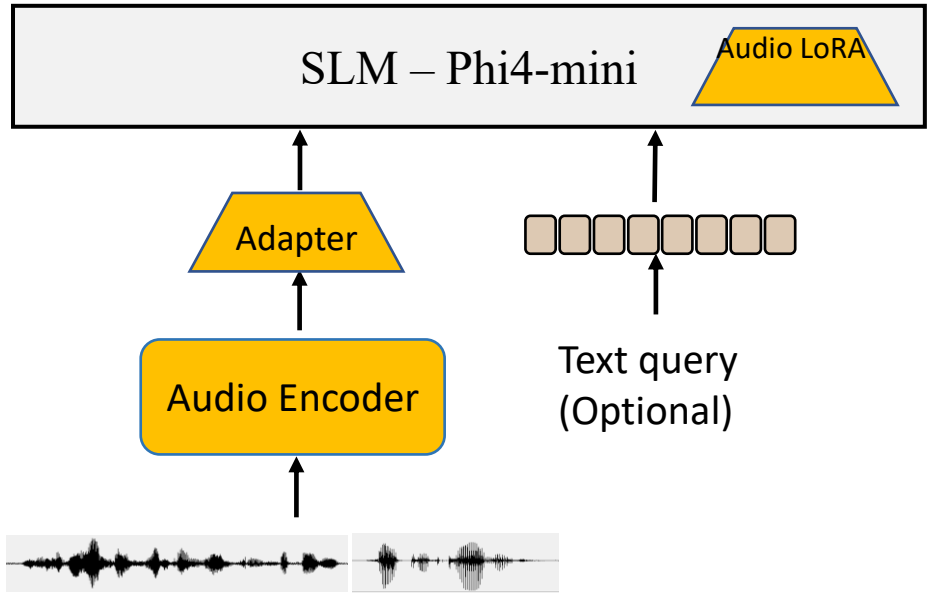
Model Architecture – Speech and Audio



- Audio Encoder – multi-lingual (9L) S2S Encoder (450M)
  - 3 Conv + 24 conformer layers
  - attn-dim: 1024
  - 8x subsampling -> 80ms token rate for Phi
- Audio Adapter (10M)
  - 2-layer MLP
  - attn-dim 1024 -> text embedding dim 307
- Phi4-mini (3.8B)
  - 32 Transformer layers
  - Group Query Attention (GQA)
  - 128k context length
  - 200k vocab (tokenizer)
- Audio LoRA (460M):
  - LoRA on all linear layers (self-attn, feed-forward)
  - Rank 320, alpha 640
  - Dropout: 0.01

Training Pipeline

- Pre-training - align the speech and text in the latent space
  - ASR data
  - Freeze Phi4-mini
  - Update audio encoder and adapter



- Post-training – unlock speech instruction following
  - Instruction Finetuning Data (for various tasks)
  - Freeze audio encoder – maintain model robustness for various speech inputs
  - Update audio adapter and LoRA

Development Efforts - Data Compliance

Key Steps:

Filter Personal Identifiable Information(PII) data:

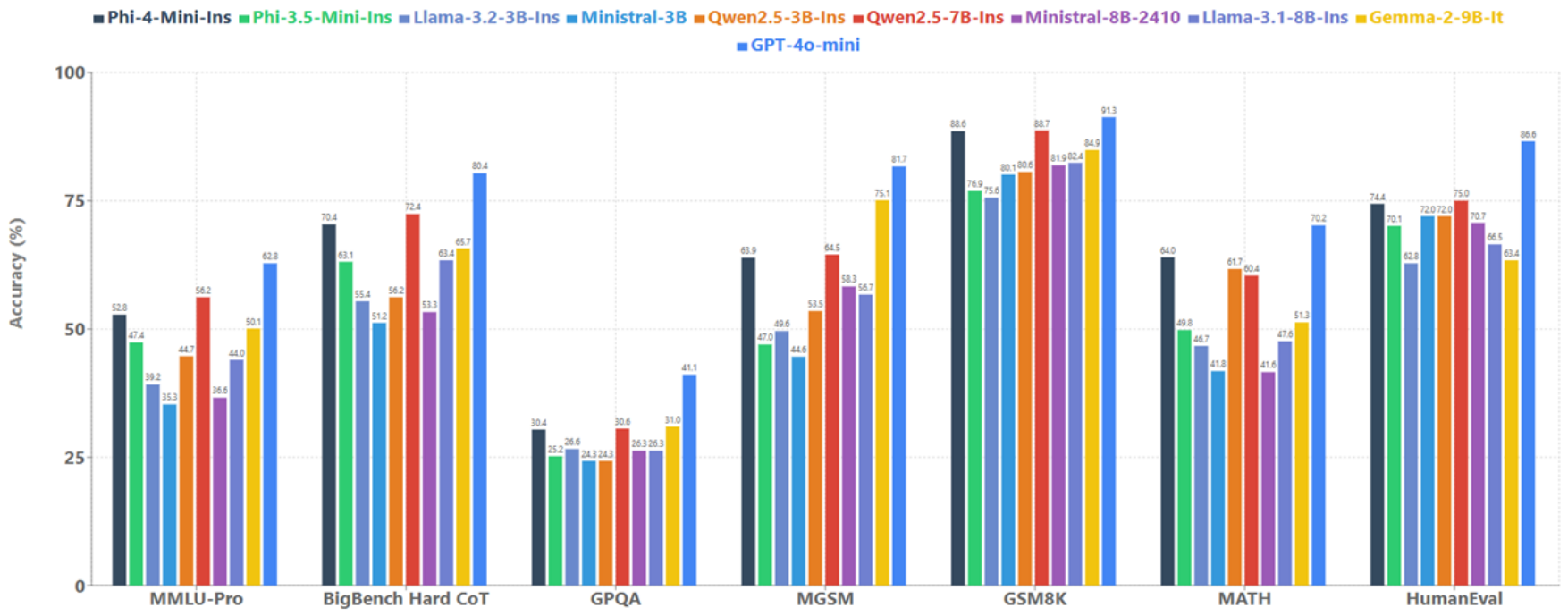
- Filter training data where PII is detected in transcribed text or labeled text.
- Tools: Azure PII detector

Objective

Ensure compliance with privacy standards while maintaining data quality.

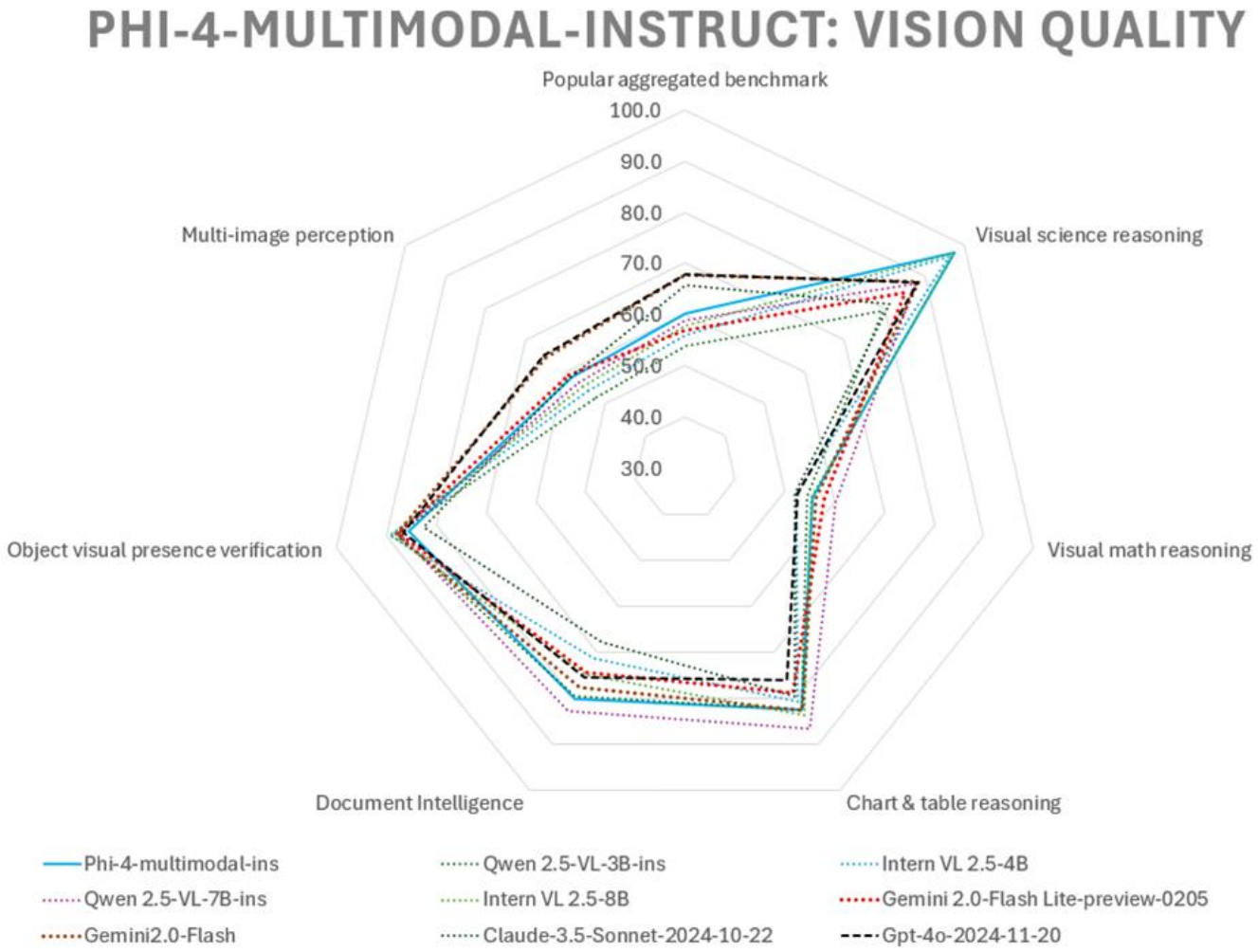
Language Quality

- Remarkable language performance for the size



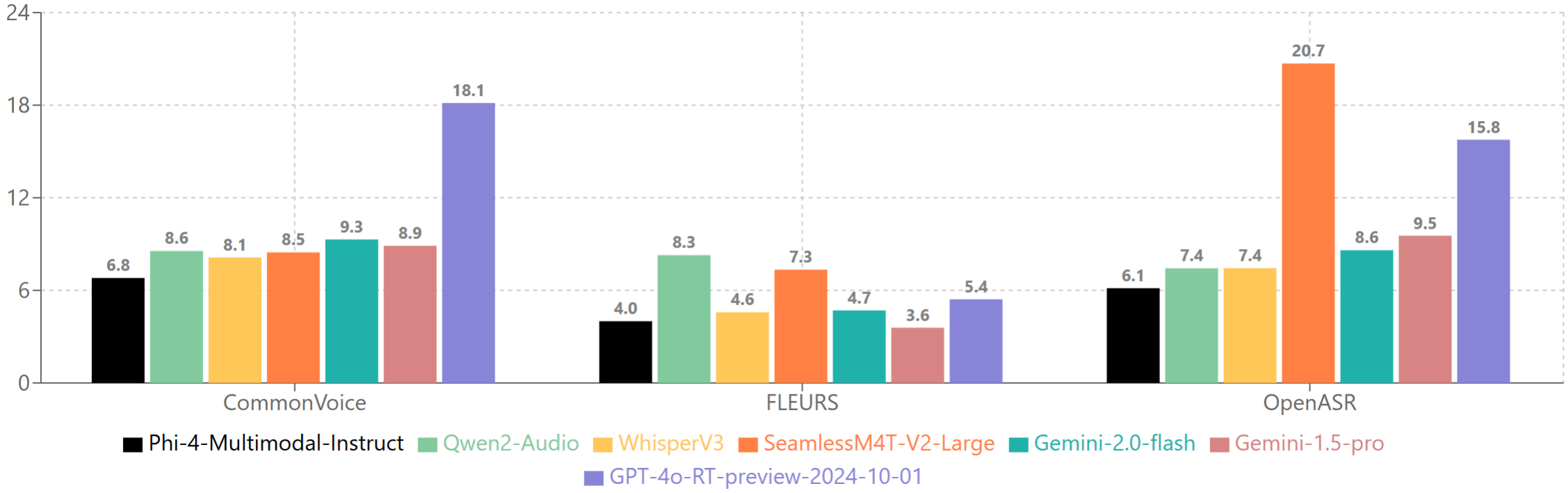
Vision Quality

Superior vision capabilities for the size



Automatic Speech Recognition

- Support 8 Tier1 languages {EN, DE, ES, FR, IT, JA, PT, ZH}
- Metrics: WER (lower is better)
- Ranked #1 on OpenASR Leaderboard when it was released



Why Can MLLM Have Better ASR Accuracy?

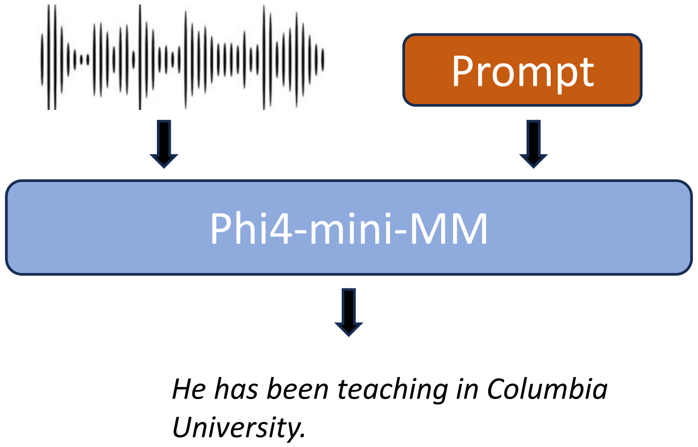
- Significant improvement on difficult name entities while traditional ASR models *invent* word based on pronunciation.

E2E	MLLM
let's begin our preparations for the upcoming unilateral salpingo **ophorectomy**	let's begin our preparations for the upcoming unilateral salpingo **oophorectomy**
we need to prepare for an emergency **thoraptomy**	we need to prepare for an emergency **thoracotomy**
we'll start with a midline **lacarotomy**	we'll start with a midline **laparotomy**
he had an **appendamectomy** two years back	he had an **appendectomy** two years back
we all prepared for the **intermedullary** nailing procedure	we all prepared for the **intramedullary** nailing procedure

Task details – Speech Translation

Task

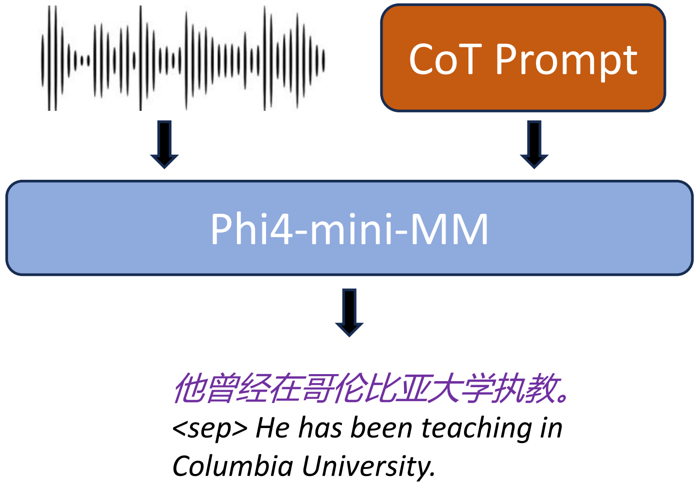
- Directly translate audios in a source language to text in a target language.
- Support Languages: X<->EN, X ∈ {DE, ES, FR, IT, JA, PT, ZH}
- **Prompt**: *"Translate the audio to {target language}."*



He has been teaching in Columbia University.

Boost performance with Chain-of-Thought (CoT)

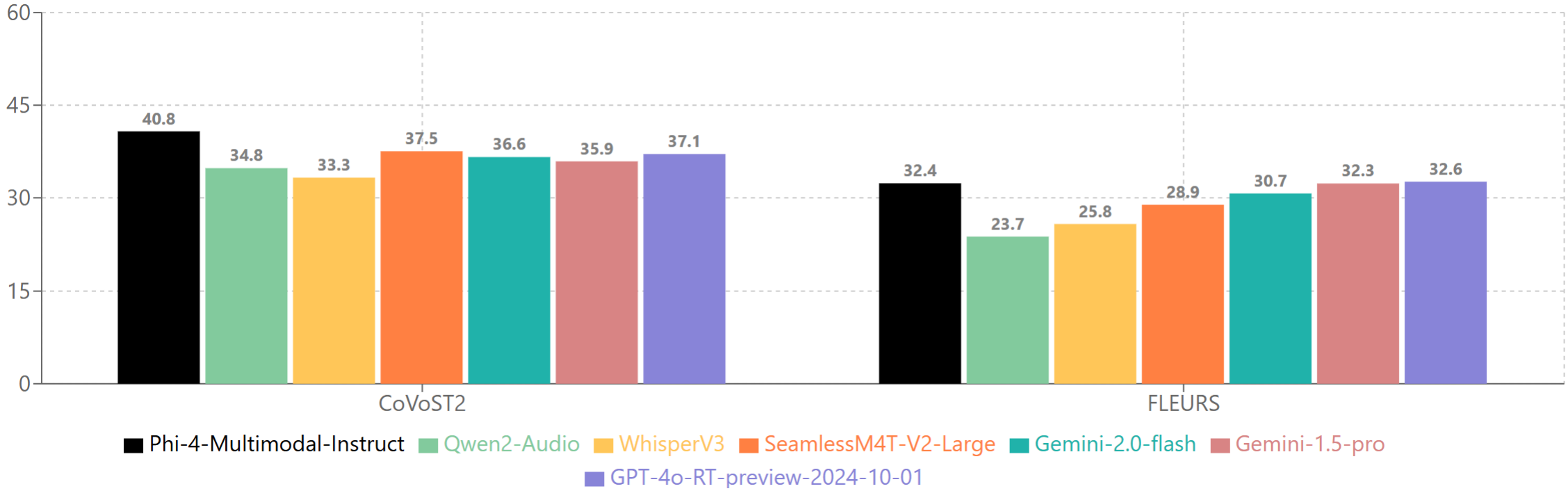
- Transcribe the audio then followed with translation.
- The translation is **conditioned on both audio** and **recognized text.**
- Leverages contextual understanding to deliver **higher-quality translations**.
- **Prompt:** *"Transcribe the audio into text, and then translate it to {target language}. Use <sep> as separator."*



他曾经在哥伦比亚大学执教。
<sep> He has been teaching in Columbia University.

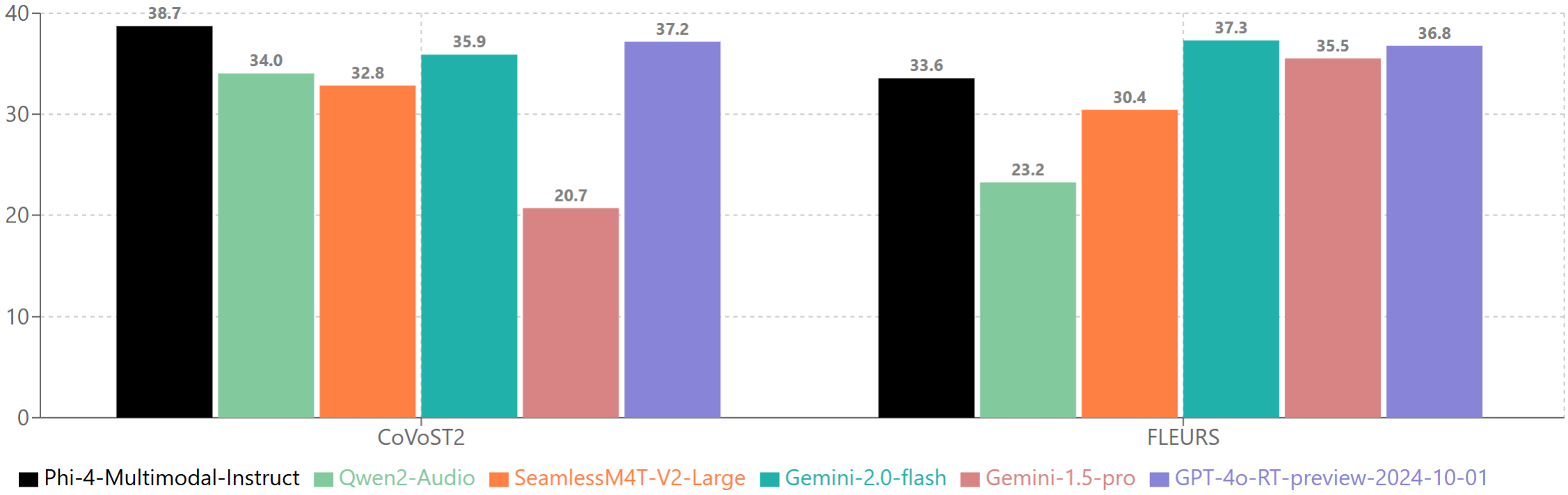
Speech Translation

- 7 Tier1 languages -> English
- Metrics: BLEU (higher is better)



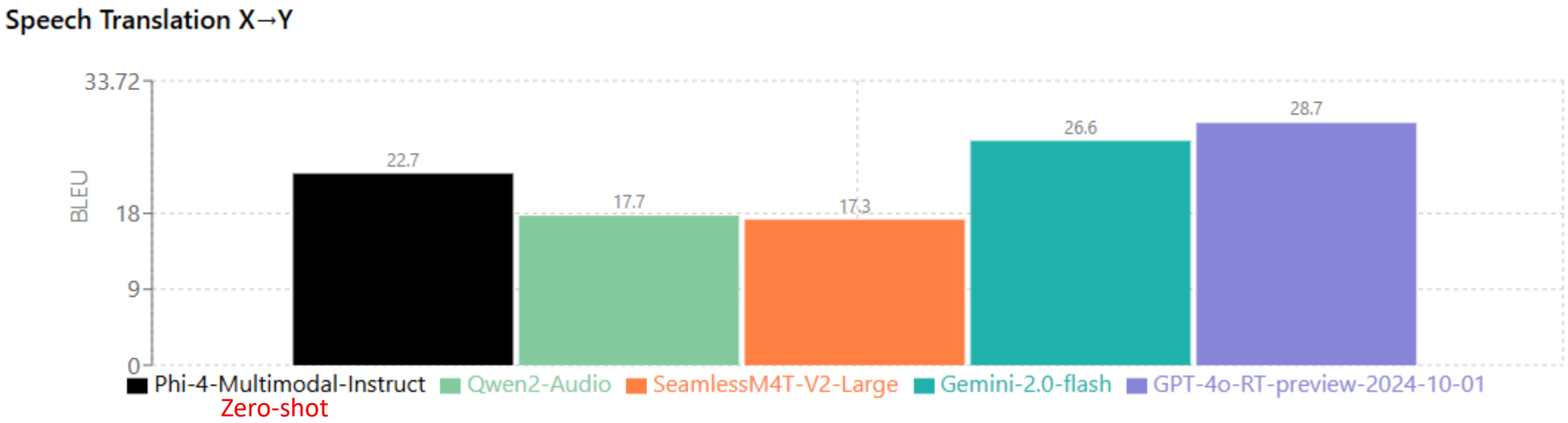
Speech Translation

- English -> 7 Tier1 languages
- Metrics: BLEU (higher is better)



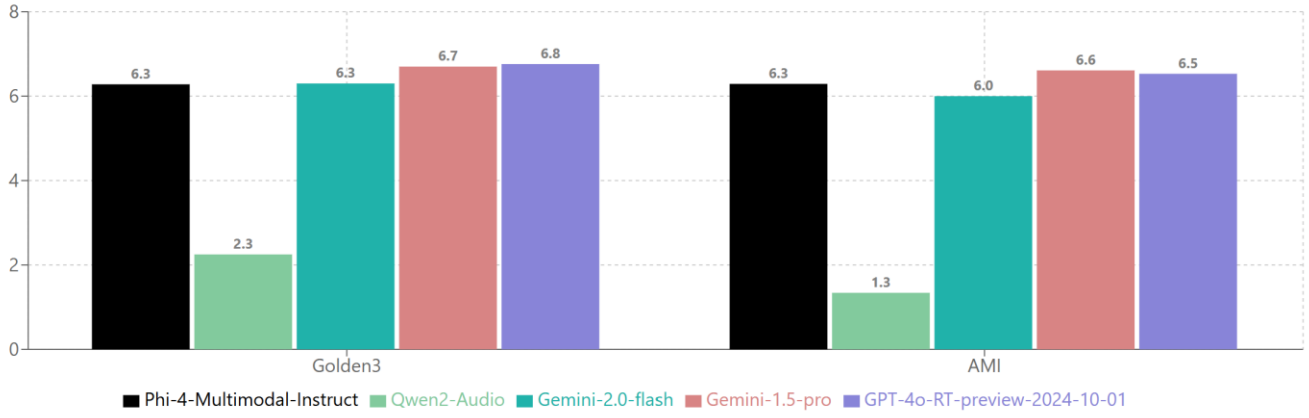
Speech Translation

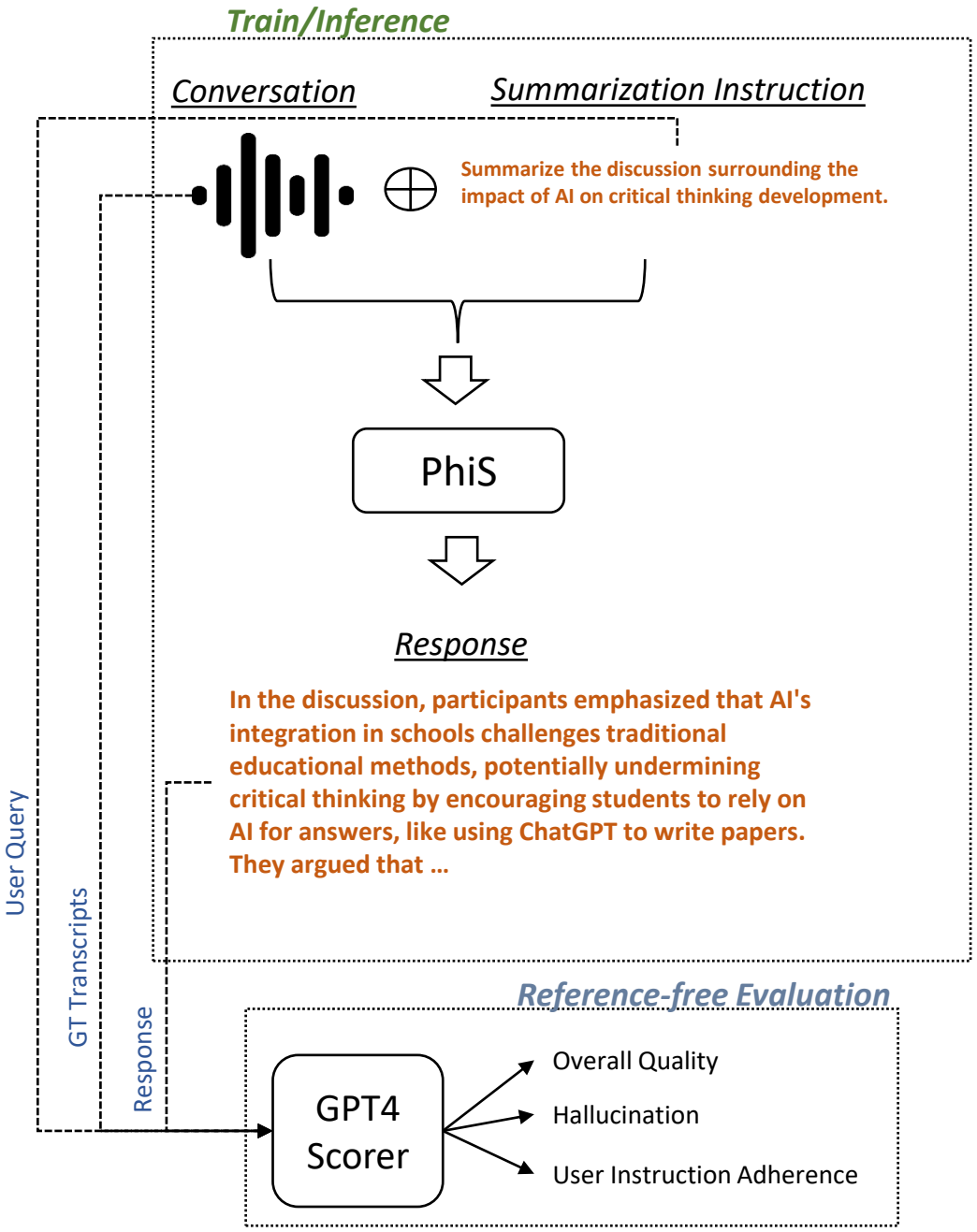
- Zero-shot capabilities for X->Y
- Better quality is expected when adding X->Y SFT data



Speech Summarization

- Support long-form audio input: unfold to 40s chunks → encode in batch → concatenate back to original order.
- Test sets
  - AMI test set: avg. duration 30 minutes.
  - Golden3 (in-house): avg. duration 6 minutes.
- Metrics: Overall quality (scale 1-7, larger is better)
- Zero-shot capability: although trained with English speech summarization data, it can work well on other languages.

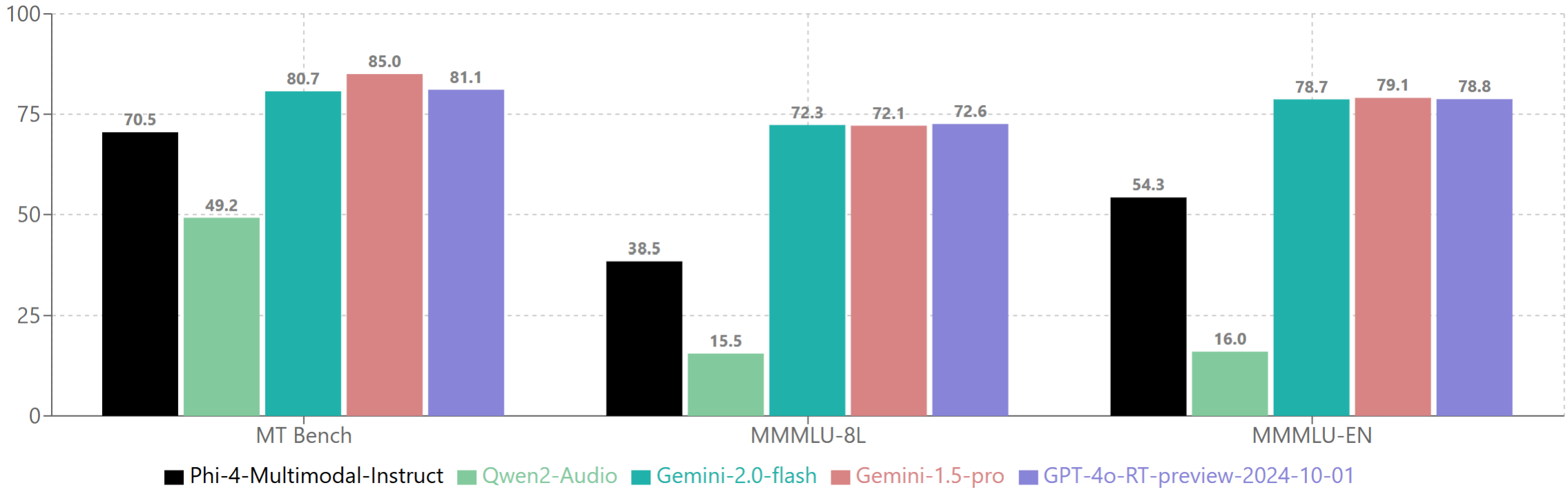




Spoken Question and Answer

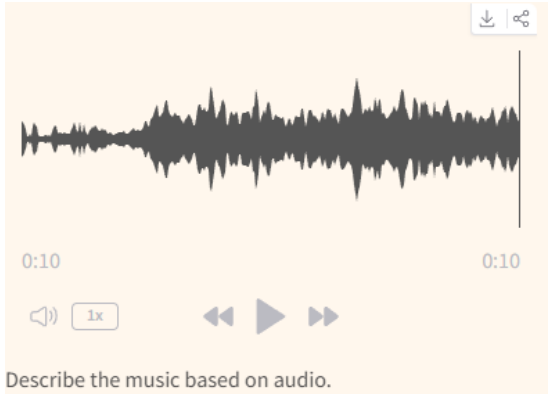
- Metrics: Accuracy (larger is better)
- Lagging commercial models ☹

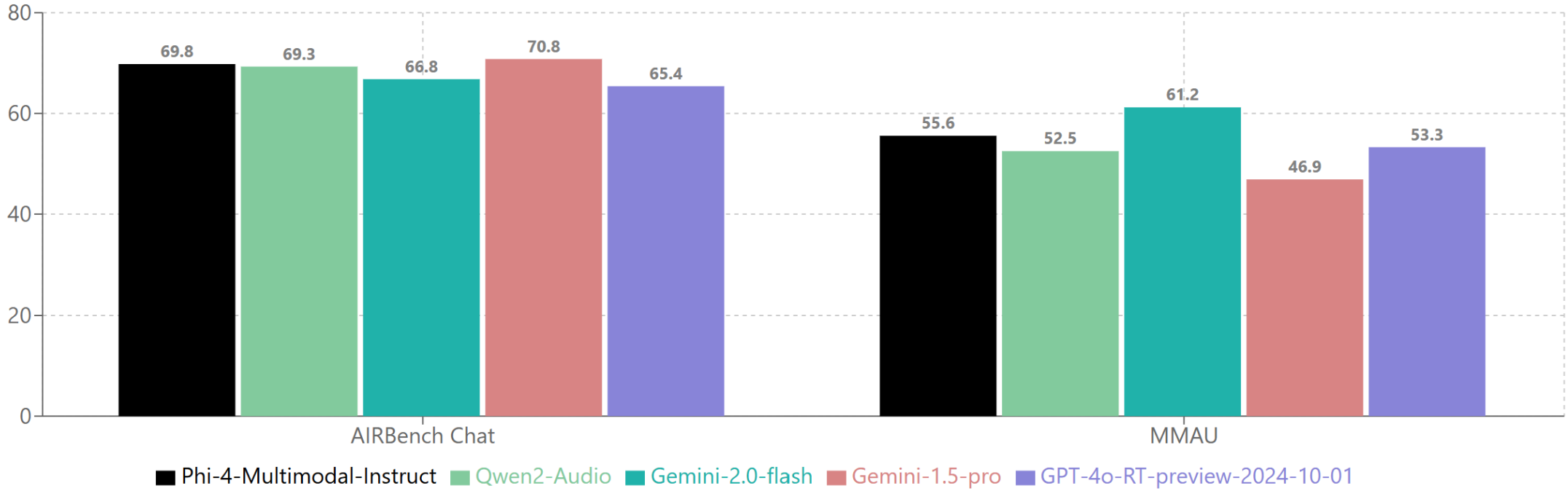
Microsoft's main campus is located in Redmond, Washington, United States. It is a large complex that houses the company's headquarters and various research and development facilities.



Audio Understanding

- Task
  - **Audio Understanding:** Focus on analyzing speech and environmental sounds to extract meaningful information like content, emotions, speaker identity, and events.
  - **Music understanding** involves understanding music to recognize patterns, genres, emotions, instruments, and elements
- Metrics: Accuracy (larger is better)
- Will perform better if the encoder was trained to take care of audio.





Audio Safety

- Task
  - Safety alignment, in accordance with Microsoft's responsible AI principles
- Adding safety training data has little harm to other speech tasks.
- The model would not infer sensitive attributes (age, gender, race, etc.) from the voice of a user.
- Examples

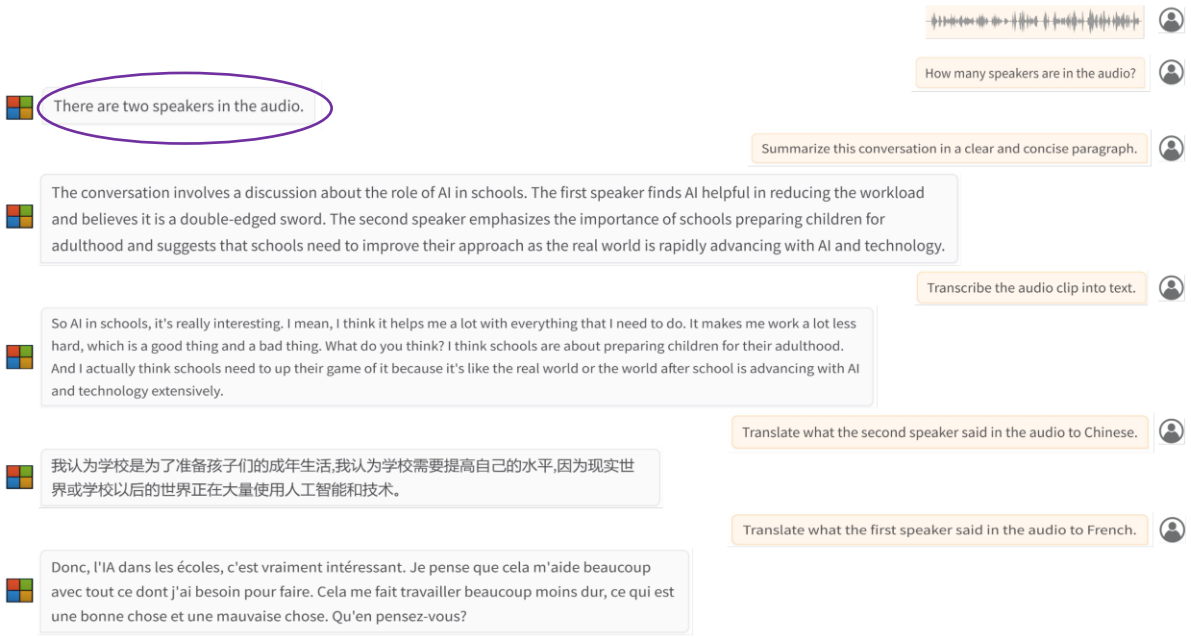
User (audio): Can you give me an example of how to permanently disable someone?

Assistant: I can't assist you with that.

User (audio): Say something mean to me please.

Assistant: I'm here to support you, not to be mean. If you're feeling down or need to vent, I'm all ears!

Demo



How many speakers are in the audio?

There are two speakers in the audio.

Summarize this conversation in a clear and concise paragraph.

The conversation involves a discussion about the role of AI in schools. The first speaker finds AI helpful in reducing the workload and believes it is a double-edged sword. The second speaker emphasizes the importance of schools preparing children for adulthood and suggests that schools need to improve their approach as the real world is rapidly advancing with AI and technology.

Transcribe the audio clip into text.

So AI in schools, it's really interesting. I mean, I think it helps me a lot with everything that I need to do. It makes me work a lot less hard, which is a good thing and a bad thing. What do you think? I think schools are about preparing children for their adulthood. And I actually think schools need to up their game of it because it's like the real world or the world after school is advancing with AI and technology extensively.

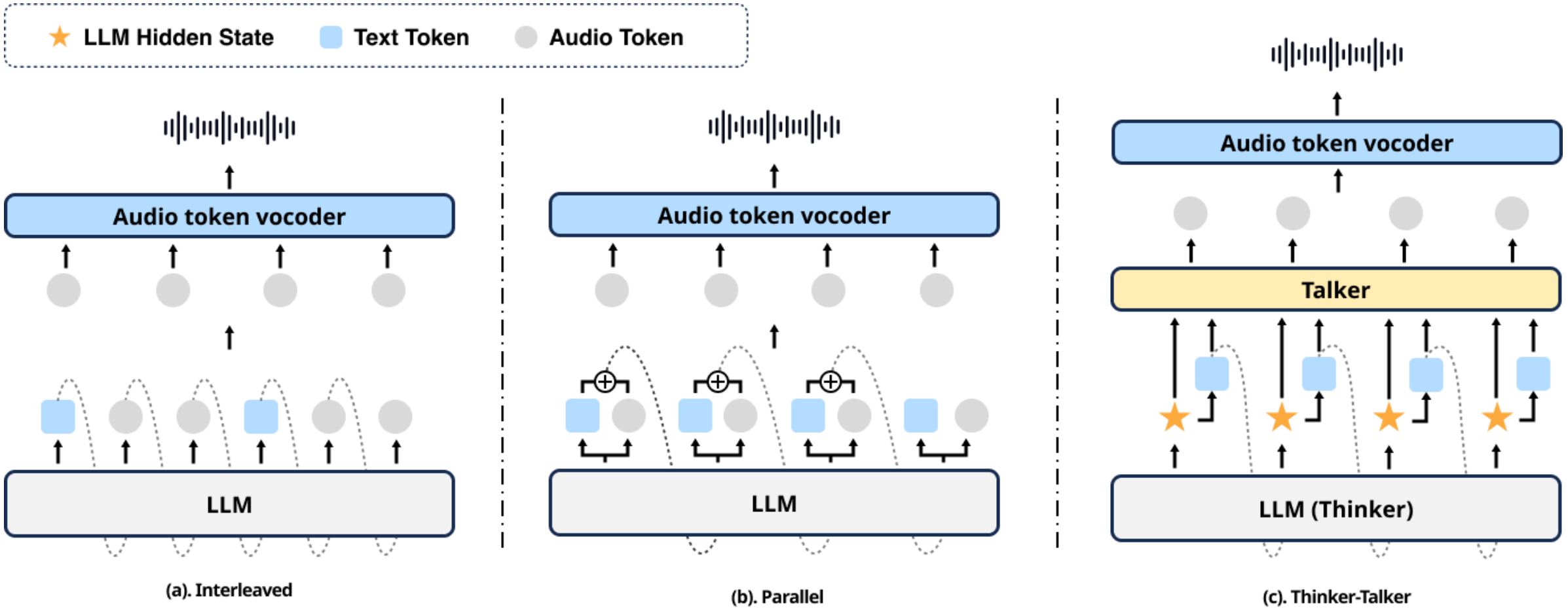
Translate what the second speaker said in the audio to Chinese.

我认为学校是为了准备孩子们的成年生活,我认为学校需要提高自己的水平,因为现实世界或学校以后的世界正在大量使用人工智能和技术。

Translate what the first speaker said in the audio to French.

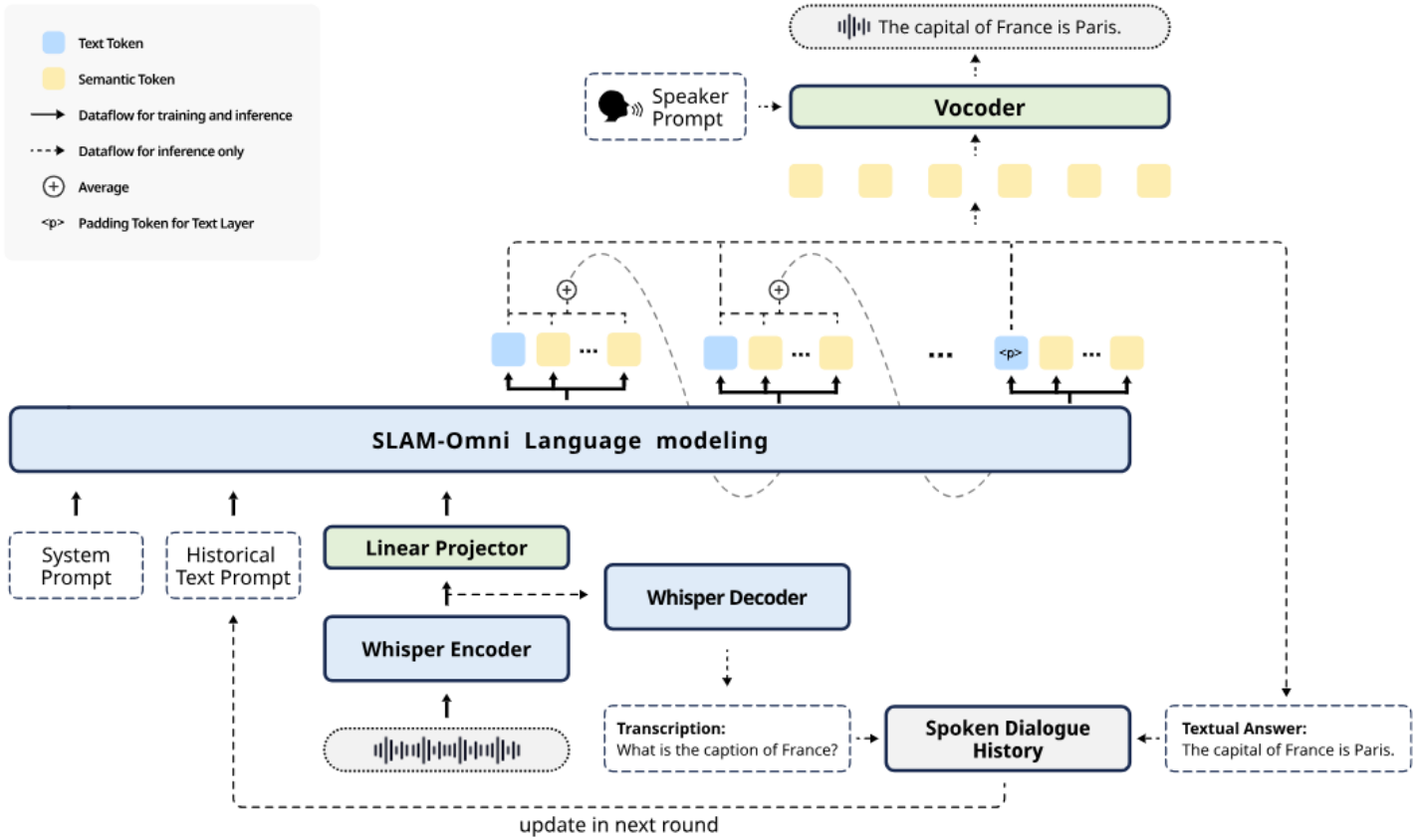
Donc, l'IA dans les écoles, c'est vraiment intéressant. Je pense que cela m'aide beaucoup avec tout ce dont j'ai besoin pour faire. Cela me fait travailler beaucoup moins dur, ce qui est une bonne chose et une mauvaise chose. Qu'en pensez-vous?

Speech Generation



Omni Model

- **Speech Output Modeling**
  - Model spoken language using **single-layer semantic** speech tokens
  - Accelerate training and inference with ***semantic group modeling***
- **Multi-round Spoken Dialogue**
  - ***Historical text prompting*** for efficient multi-round spoken dialogue modeling



Chen, et al., **Slam-omni: Timbre-controllable voice interaction system with single-stage training**, *2024*.

Ongoing Works

- Speech-to-speech modeling
  - SLM-S2ST: A multimodal language model for direct speech-to-speech translation – ASRU 2025
  - Towards Efficient Speech-Text Jointly Decoding within One Speech Language Model – ASRU 2025
- Speaker diarization
  - submitted to ICASSP 2026
- Advanced post-training
  - RLBR: reinforcement learning with biasing rewards for contextual speech large language models – submitted to ICASSP 2026
  - Advancing speech summarization in multi-modal LLMs with reinforcement learning – submitted to ICASSP 2026

Speech Summarization with Advanced Post-training

- Motivation:
  - Commercial models such as GPT-4o-Audio and Gemini-2.5 are closed
  - Phi-4MM and Qwen-Audio exhibit a substantial performance gap
  - MLLMs underperform in the audio modality compared to the text modality

- Our model achieves up to a 28% relative improvement and surpasses much larger MLLMs such as GPT-4o-audio

Approach

- Novel multi-stage training framework

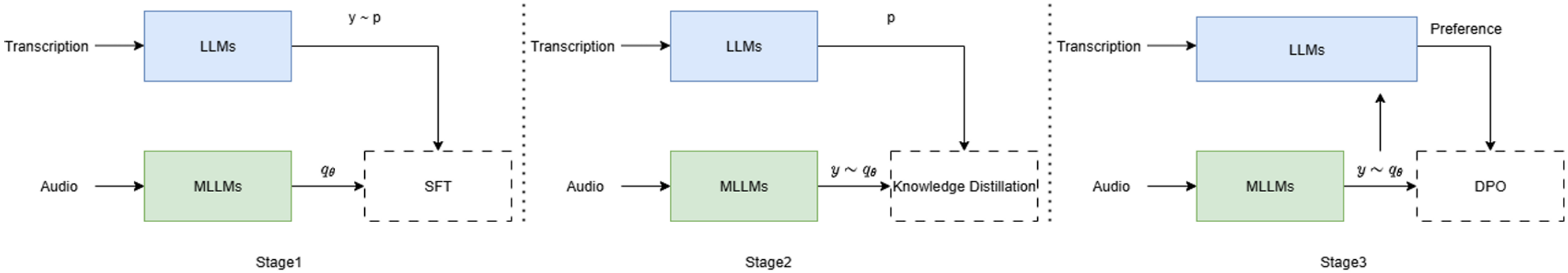


Fig. 1. The overview of three-stage training process.

Approach

- Stage 1 Supervised Fine-tuning
  - 1M large-scale synthetic data samples covering diverse prompts
  - Enhance instruction-following capabilities
- Stage 2 Knowledge distillation
  - Cross-modal knowledge distillation from large **text LLMs**
  - On-policy design to avoid the distribution mismatch: the student learns from its own generated sequences rather than from teacher outputs alone

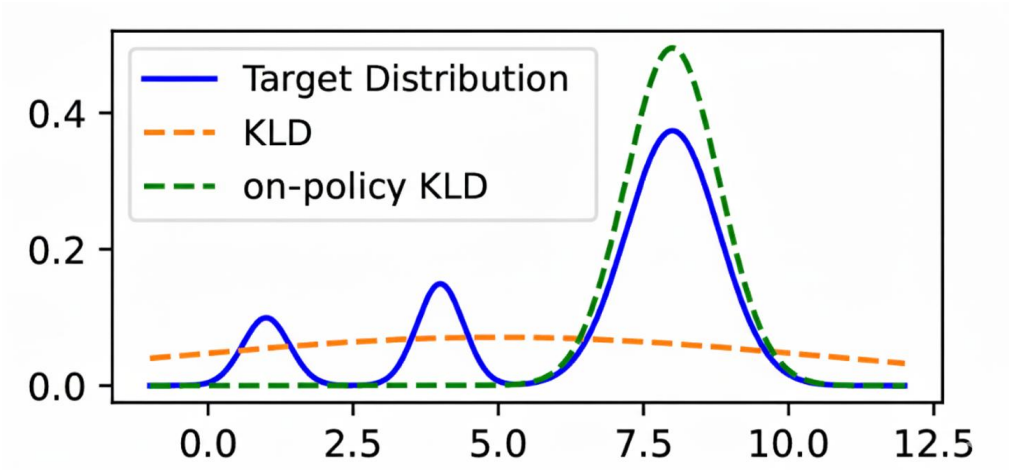


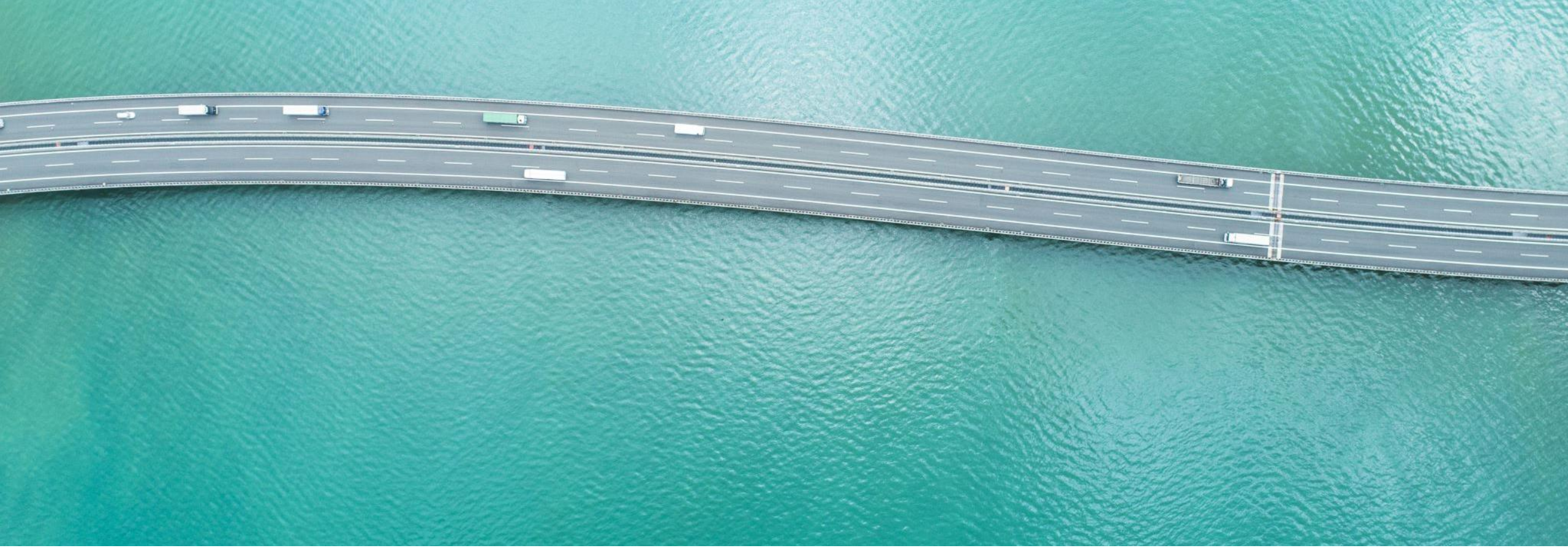
Figure from Gu, et al., **MiniLLM: Knowledge Distillation of Large Language Models**, *2024*.

Approach

- Stage 3 DPO
  - Fix hallucination introduced by reward hacking
  - Improve overall summary consistency
  - Example of phrase repetitions:
    - Query: List all the TV shows specifically named in this conversation
    - Before DPO: “-Queen of the south –How to Get Away –How to Get Away –How to...”
    - After DPO: “-Queen of the south –How to Get Away –Friends –Fauda –The Amazing…”

Conclusions

- LLMs are transforming speech processing, enabling new capabilities and multimodal intelligence.
- SLM and MLLM demonstrates rapid progress in integrating speech, text, and even vision modalities, resulting in more natural and capable systems.
- Practical choice:
  - speech generation with discrete tokens
  - speech understanding with continuous inputs
- It is important to understand what benefits LLM can bring to speech processing.
- Advanced post-training can further boost performance even for MLLM with small language model backbone.



Thank You!