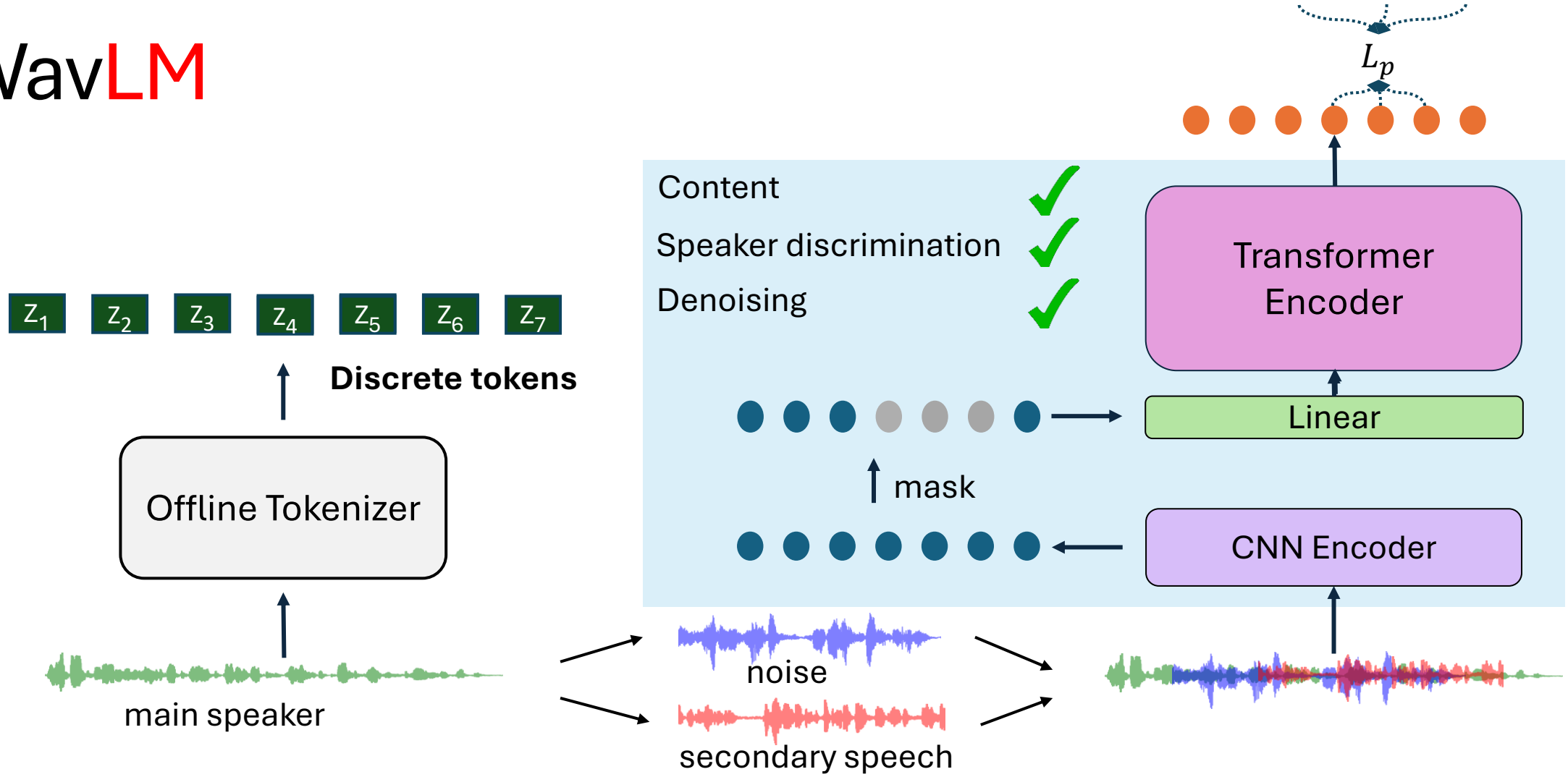
# Speech LM Journey at Microsoft

Jinyu Li

close collaboration with MSRA multi-media group

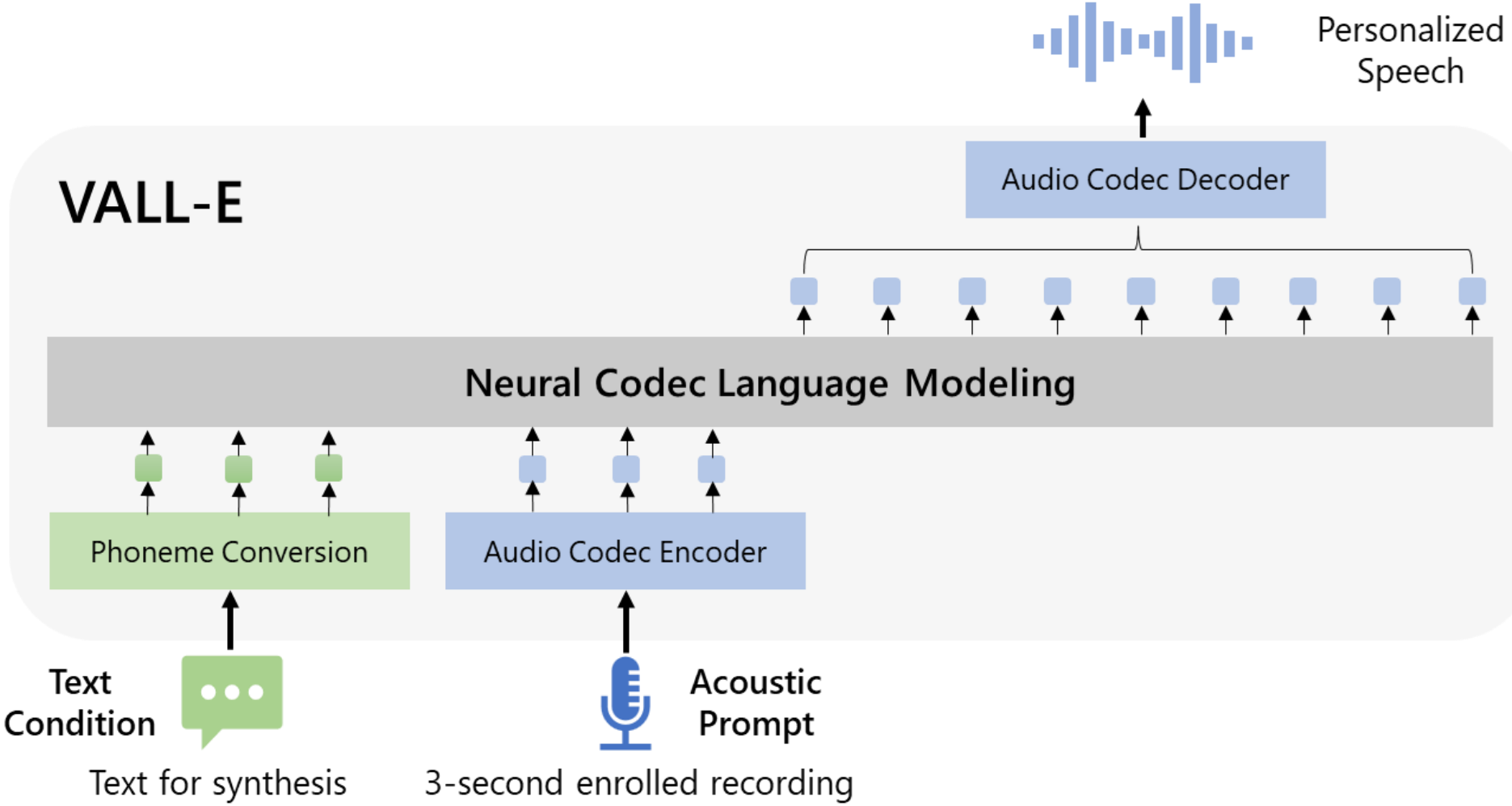
# WavLM





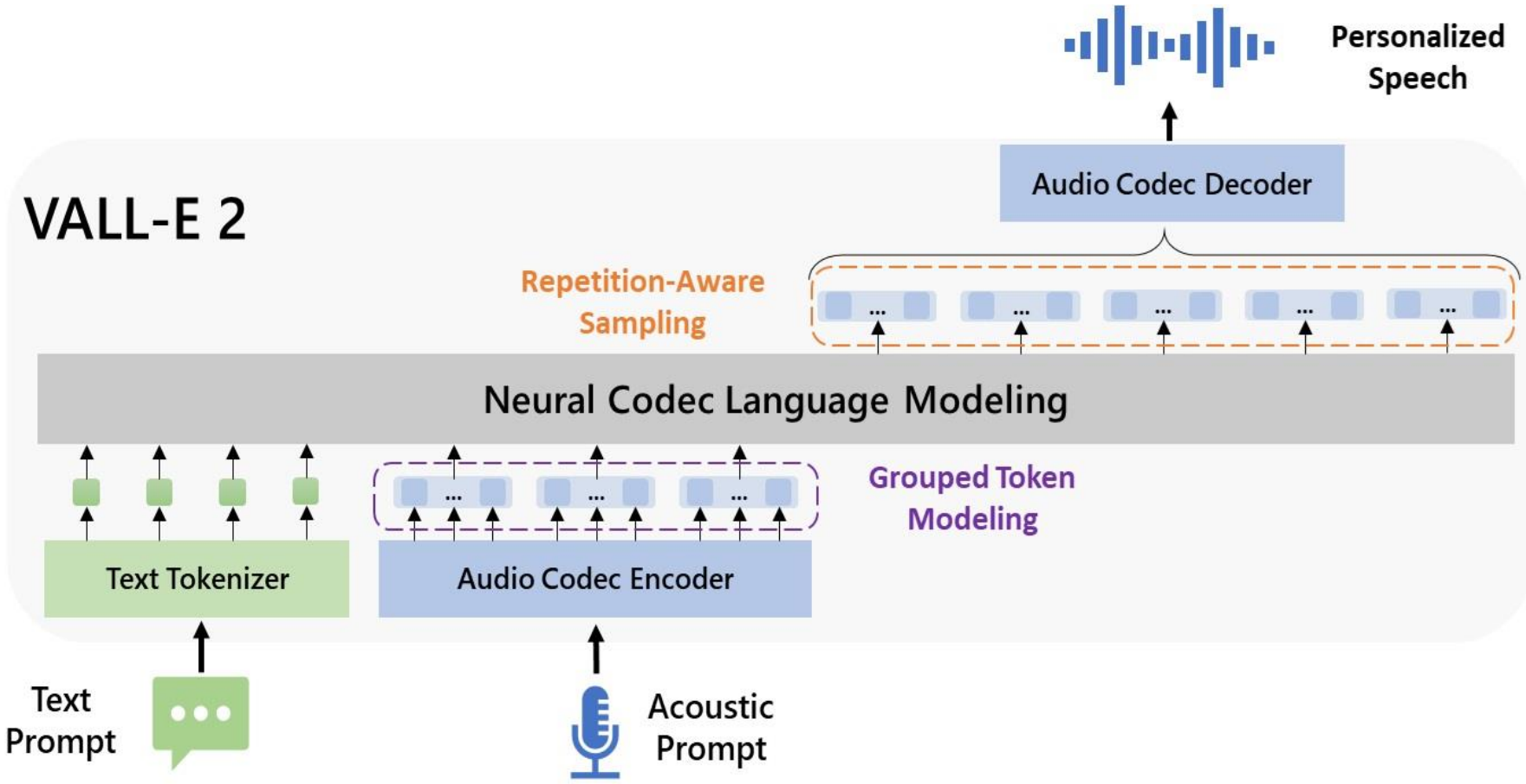
Chen et. al., **WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing**, 2021.
Zhang et. al., **SpeechLM: Enhanced Speech Pre-Training with Unpaired Textual Data**, 2022.
Zhu et. al., **VATLM: Visual-Audio-Text Pre-Training with Unified Masked Prediction for Speech Representation Learning**, 2022.

# VALL-E

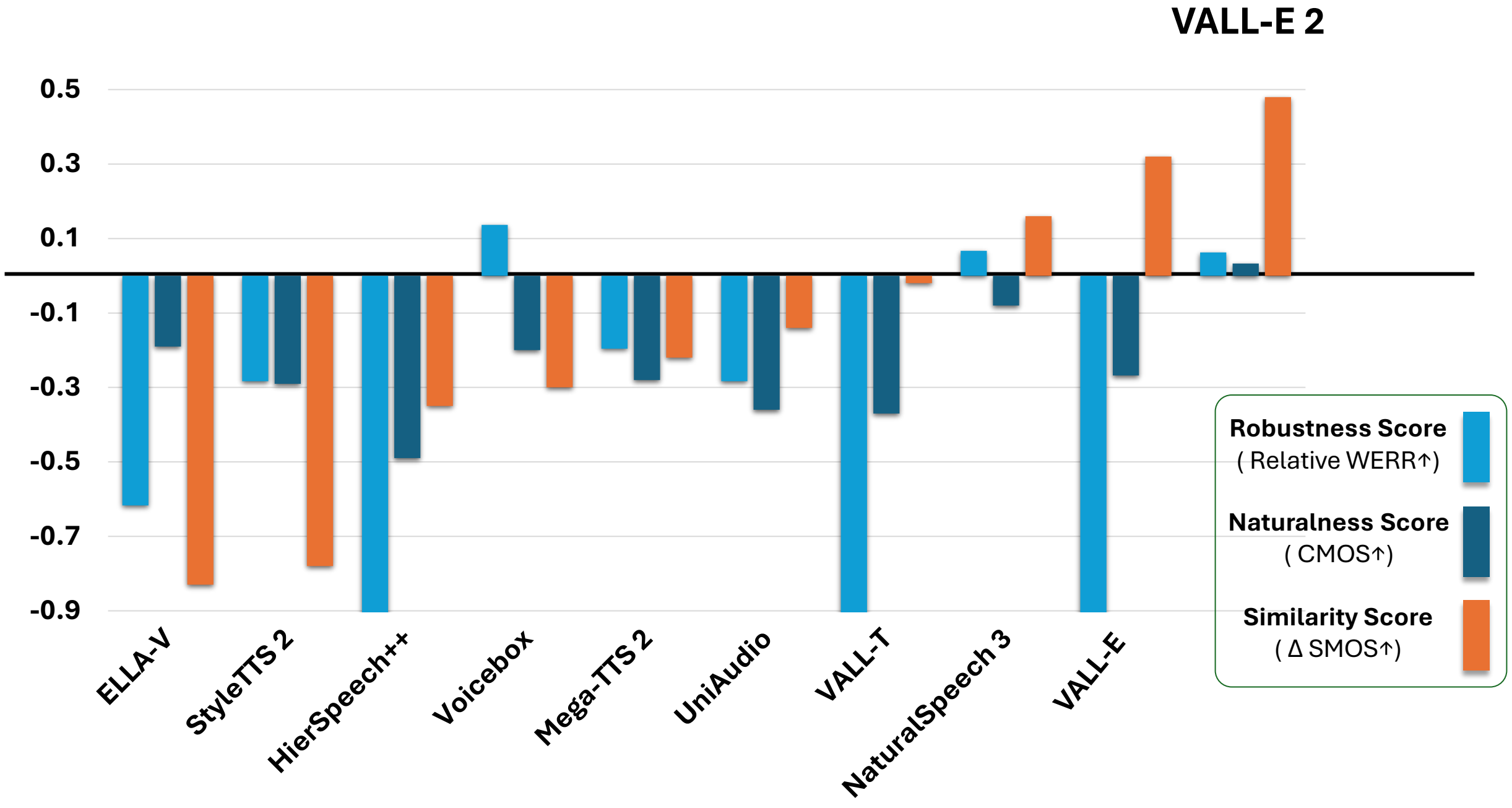



Wang et al., **Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers**. 2023

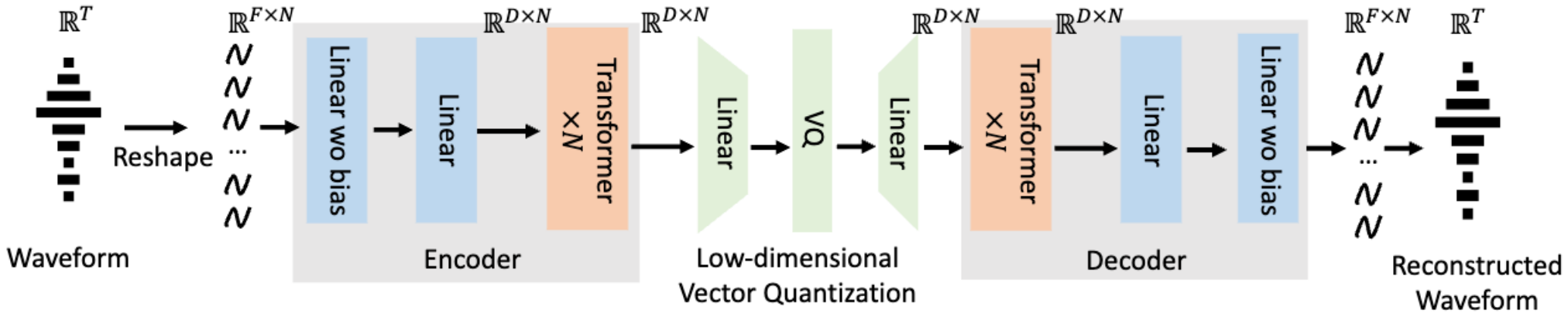
# VALL-E 2




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↑)

# TS3-Codec - **T**ransformer-Based **S**imple **S**treaming **S**ingle Codec




## Properties

1. Streaming[1]
2. Low computation[2]
3. Single codebook
4. Low token rate (bitrate) [3]

## Advantages

1. Full duplex speech LMs
2. Save computation for speech LMs
3. Avoid complicated speech LM decoding strategies
4. Easy the speech LM training

[1]Fixed left context window for self-attention
[2]TS3-Codec (**1.6B** paras): 60.52G MACs, while convolutional based BigCodec (**160M** paras): 61.1G MACs
[3]Bitrate=0.6k, token rate= 40

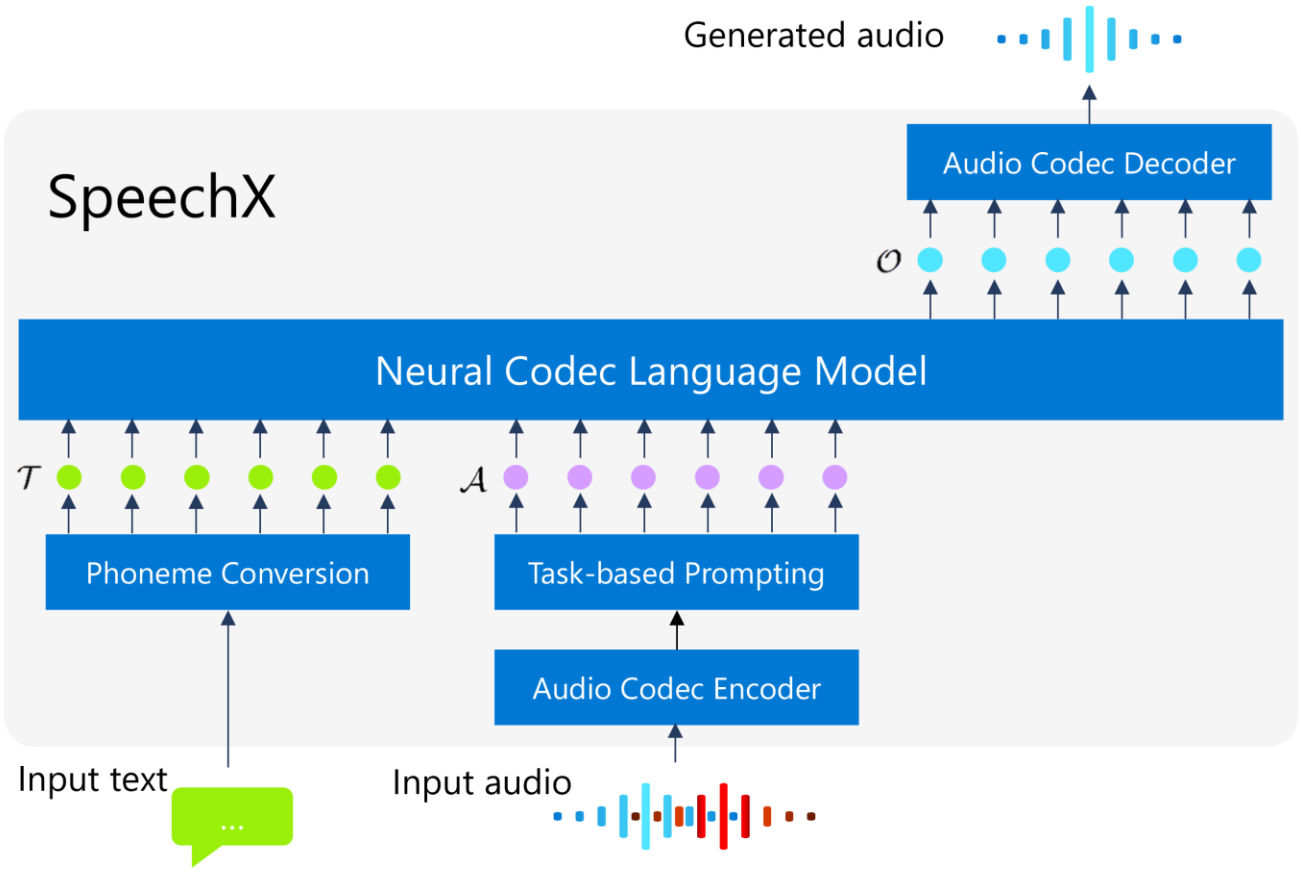
Wu, et al., **TS3-Codec: Transformer-Based Simple Streaming Single Codec**, *2024*.

# 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.

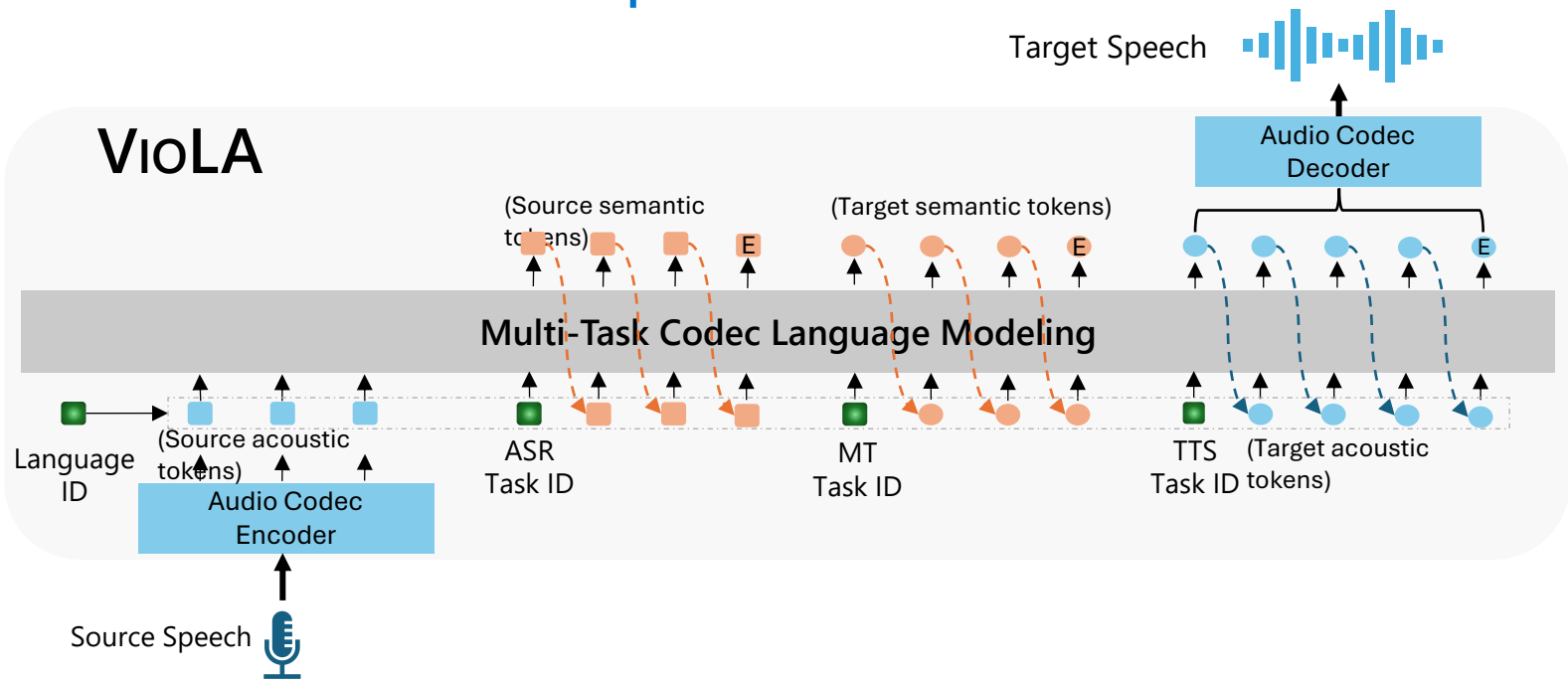



| Task | Input text | Input audio | Output audio |
|---|---|---|---|
| Noise suppression | Transcription (optional) | Noisy speech | Clean speech |
| Speech removal | Transcription (optional) | Noisy speech | Noise |
| Target speaker extraction | Transcription (optional) | Speech mixture, Enrollment 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**, *arXiv:2308.06873,* 2023.

# Multi-modal Model with Discrete Audio Inputs: VioLA




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

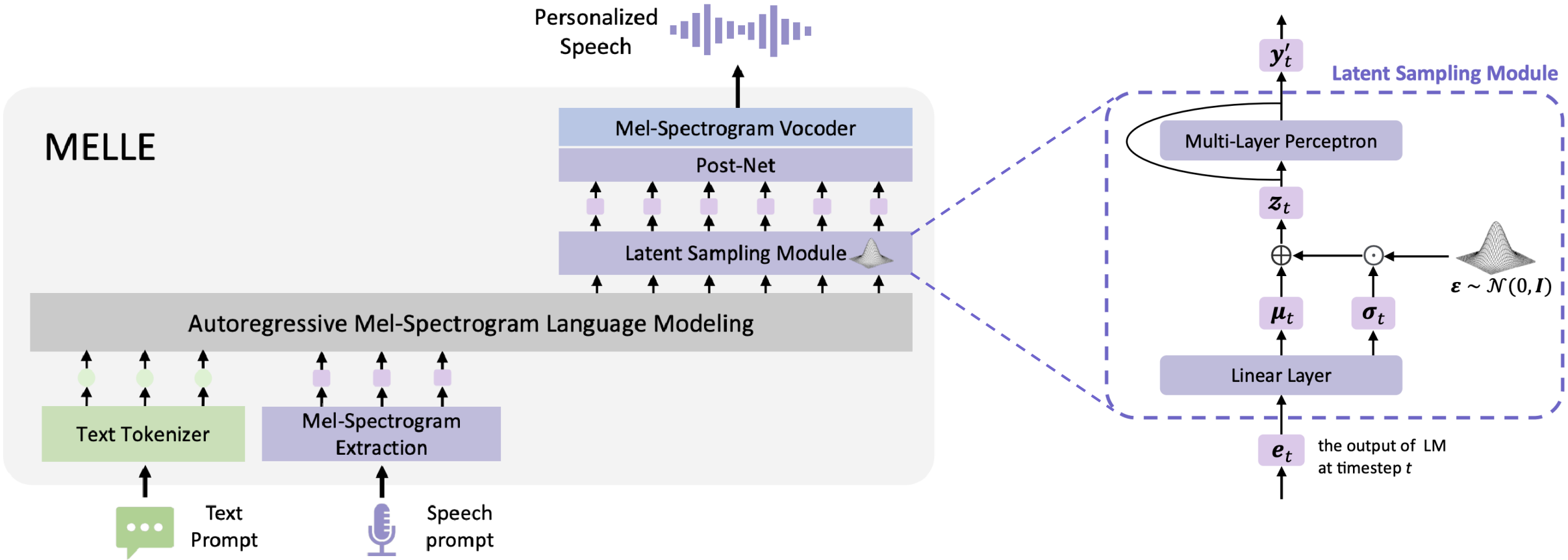
| Feature | PER |
|---|---|
| Fbank | 9.61 |
| Codec | 12.83 |

Wang, et al., **VioLA: Unified codec language models for speech recognition, synthesis, and translation**, arXiv:2305.16107, 2023.

# Discrete or Continuous Input?

- Discrete: More aligned with LLM which takes discrete text tokens.
- Continuous: Represent speech better to avoid information loss due to quantization.

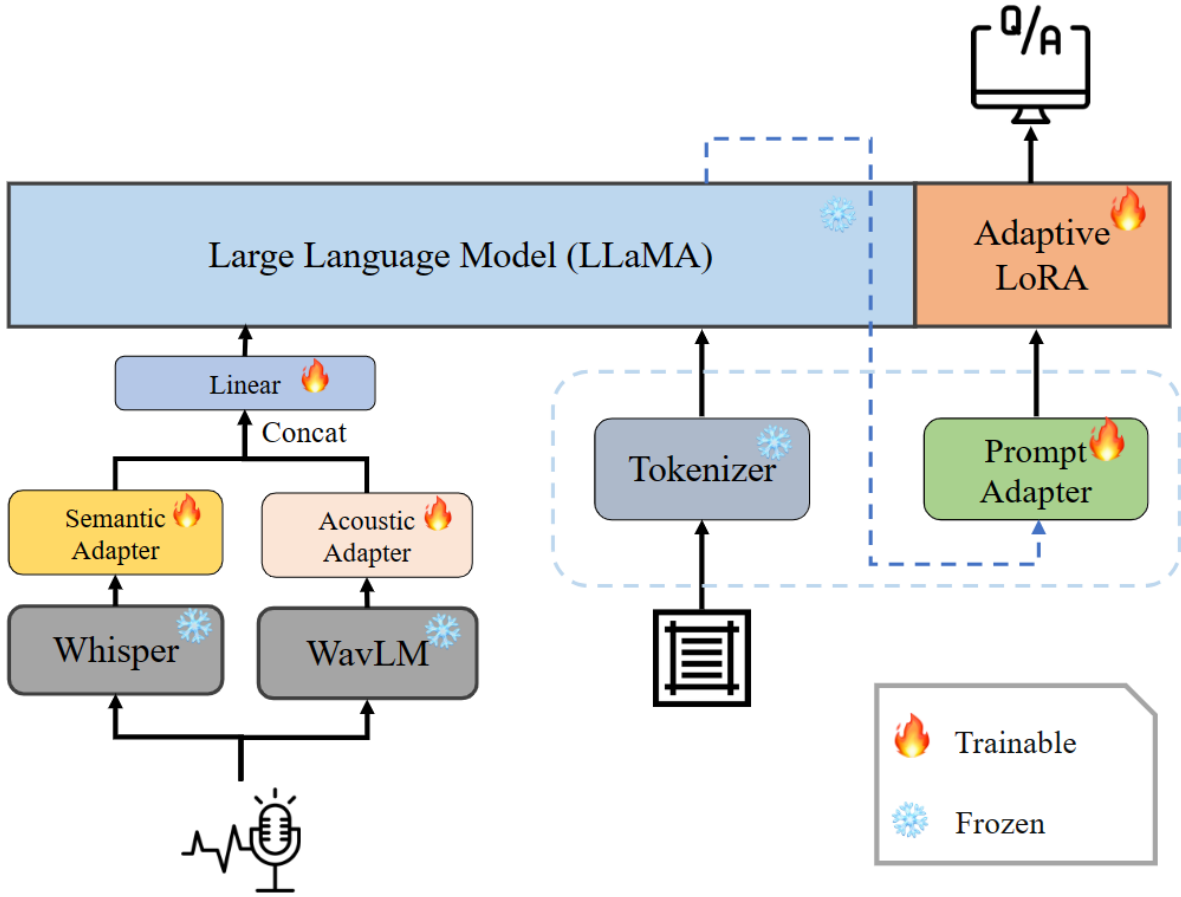
# MELL-E




For each time step, MELLE predicts a **distribution** $N(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2)$ conditioned on the history, from which a **latent state** $\boldsymbol{z}_t$ is sampled for generating the subsequent mel-spectrogram frame.

Meng et al., **Autoregressive Speech Synthesis without Vector Quantization**. 2024.

# Multi-modal Model with Continuous Audio Inputs: WavLLM




Hu, et al., **WavLLM: Towards robust and adaptive speech large language model**, *arXiv:2404.00656, 2024*.

# Problems of Most Speech-LLMs

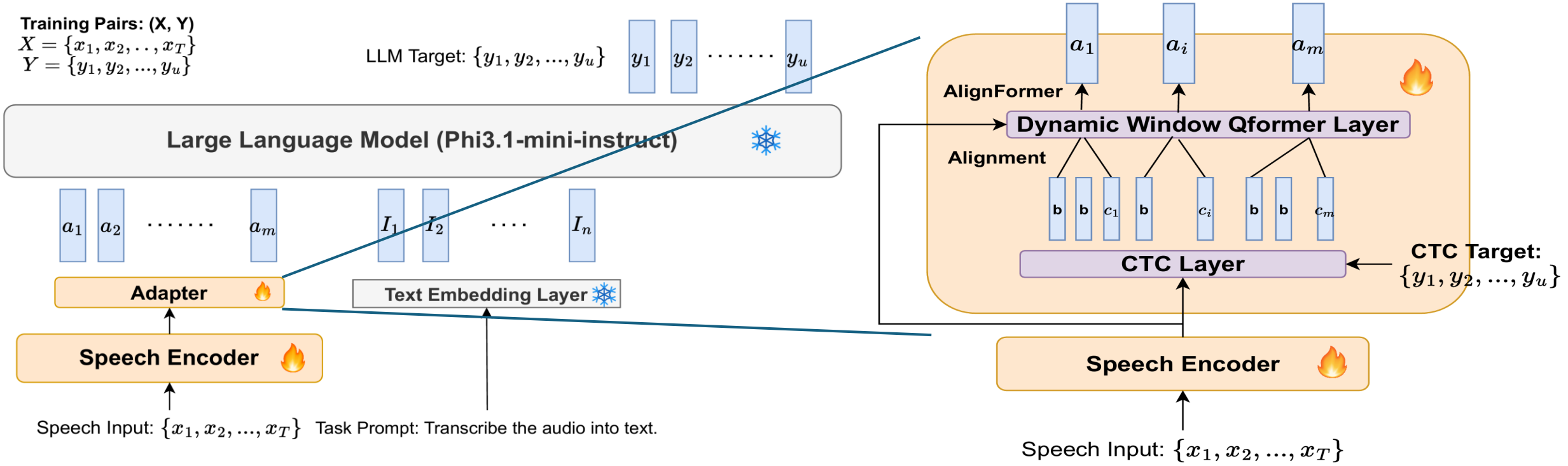
Need to collect varieties of instruction-tuning data.

May not follow the instruction unseen in training

LLM is updated during instruction tuning -> losing text capability.

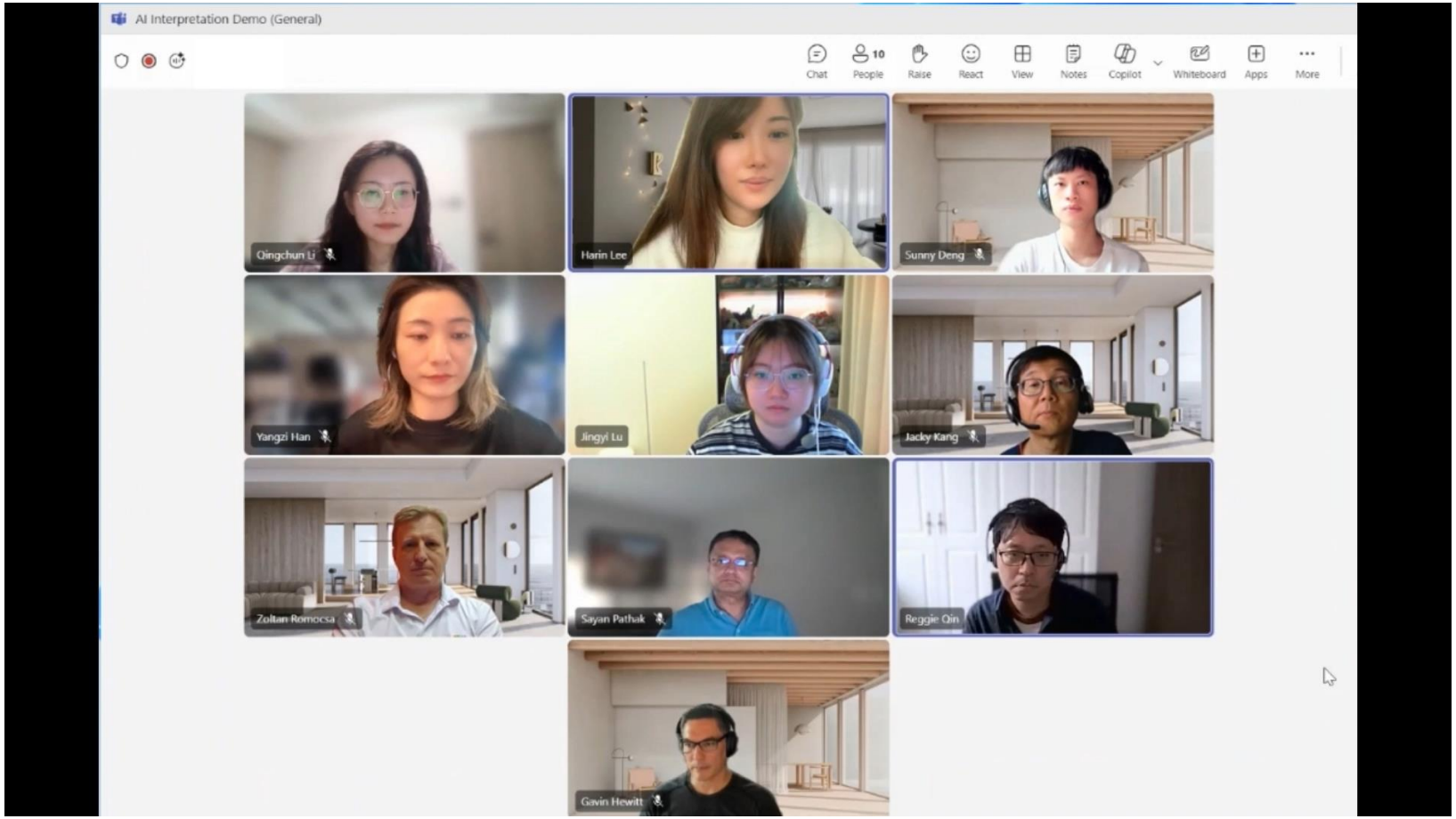
# AlignFormer: Better Zero-shot Instruction-Following Speech-LLM

100% instruction following for various speech tasks, trained with only ASR data!



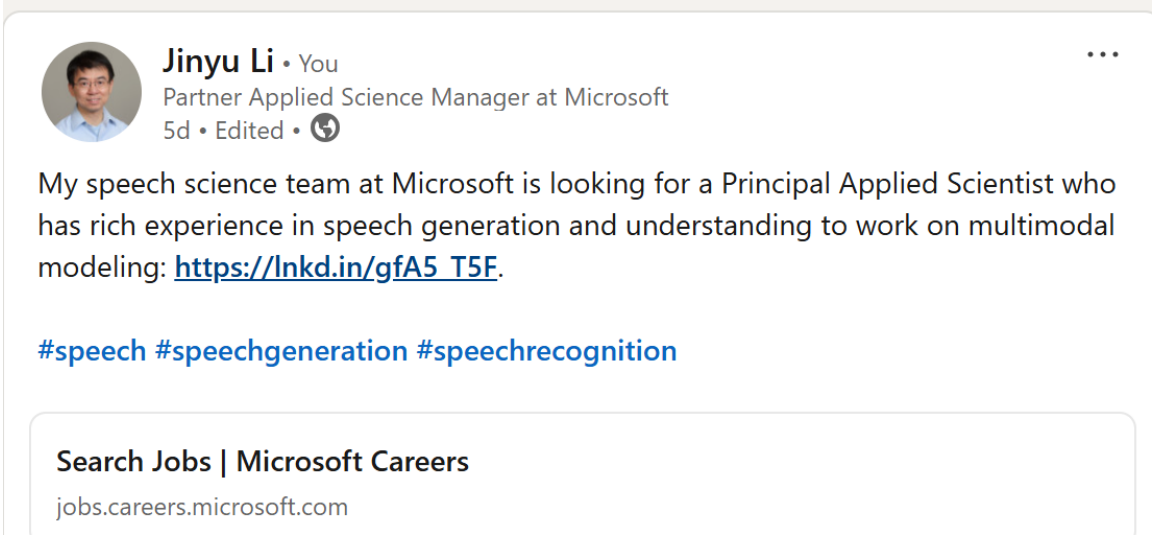

Fan, et al., **AlignFormer: Modality Matching Can Achieve Better Zero-shot Instruction-Following Speech-LLM**, *2024*.

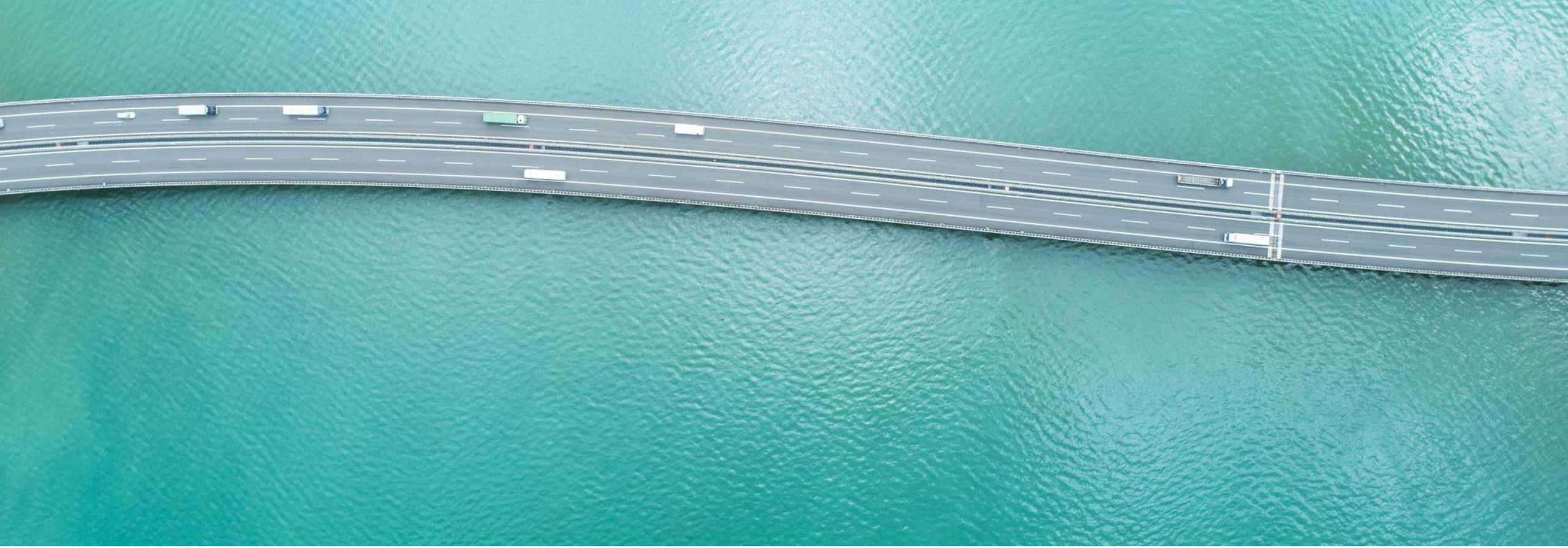
# LLM is NOT the Only Way



# Join the Journey

jinyli@microsoft.com





# Thank You!