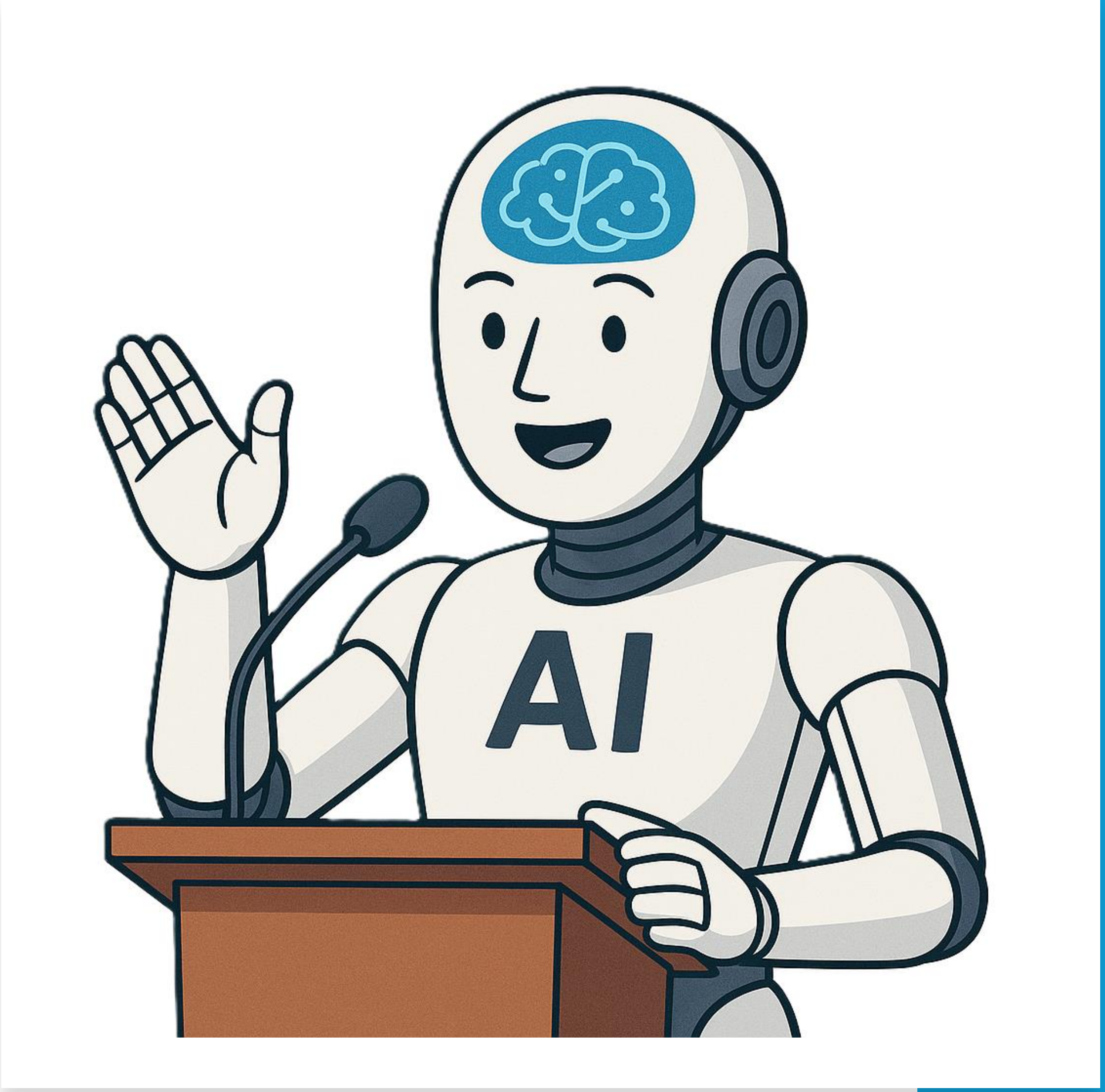
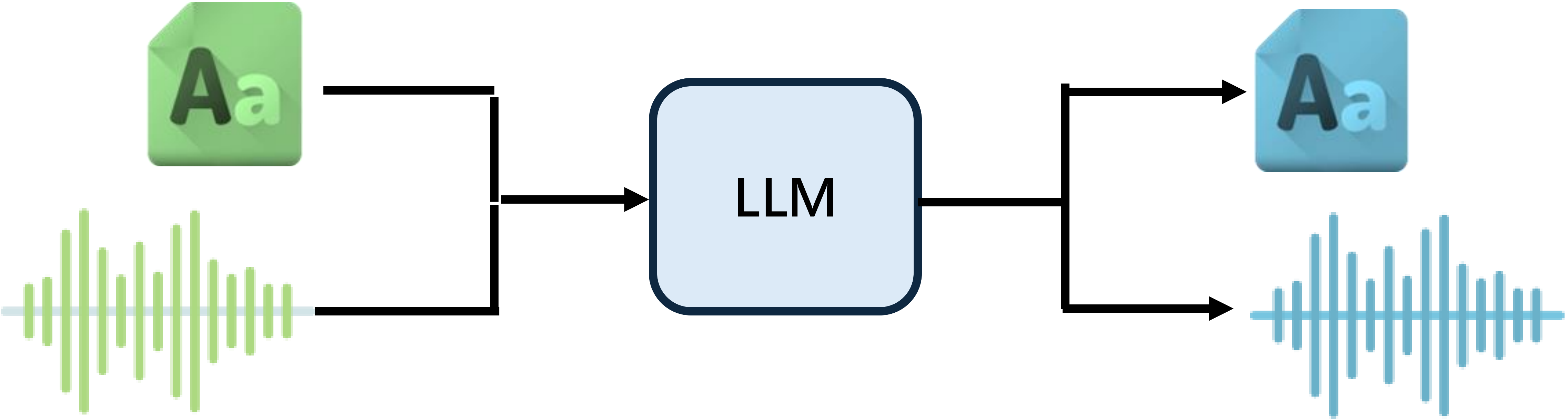
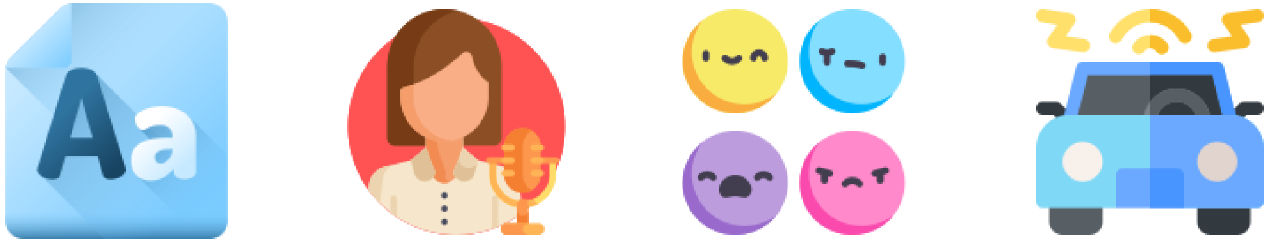
可以**聽**和**說**的
語音語言模型



Speech LLM

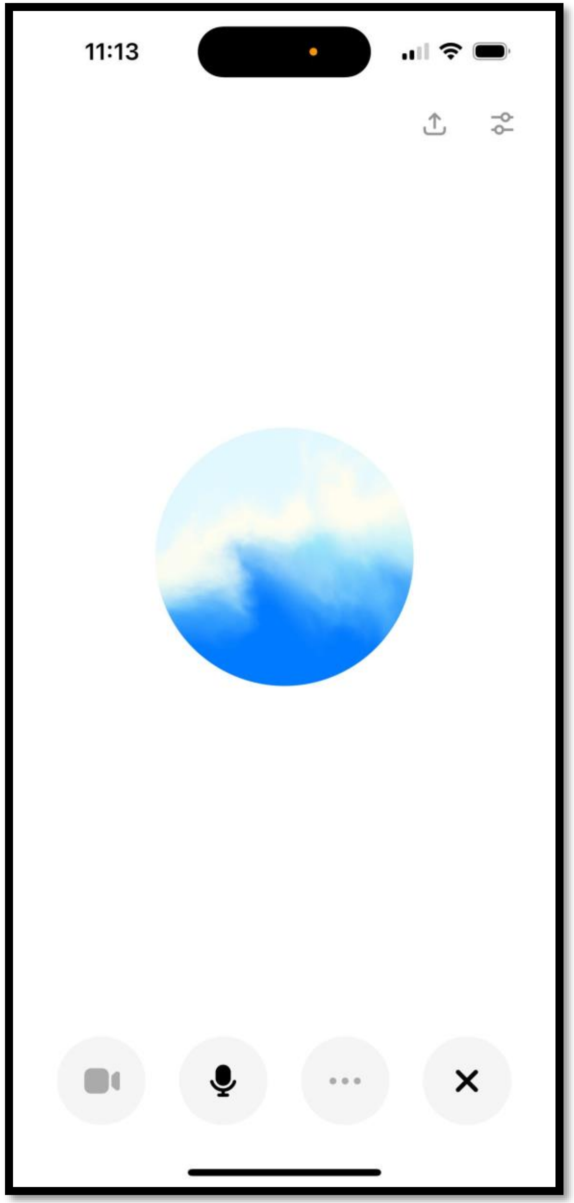


LLM

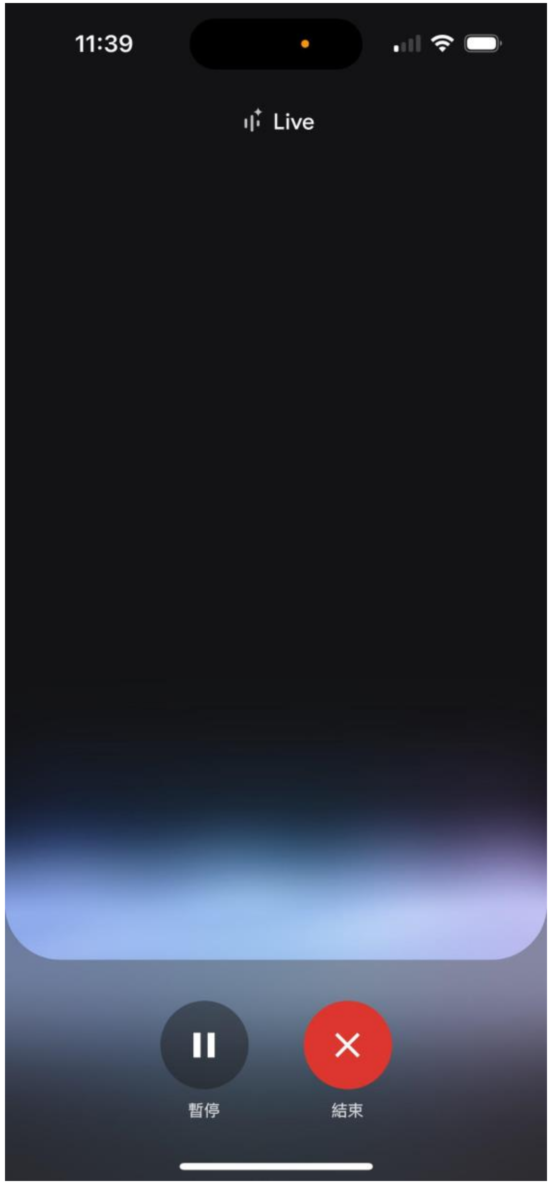


內容 語者 情緒 環境

Example



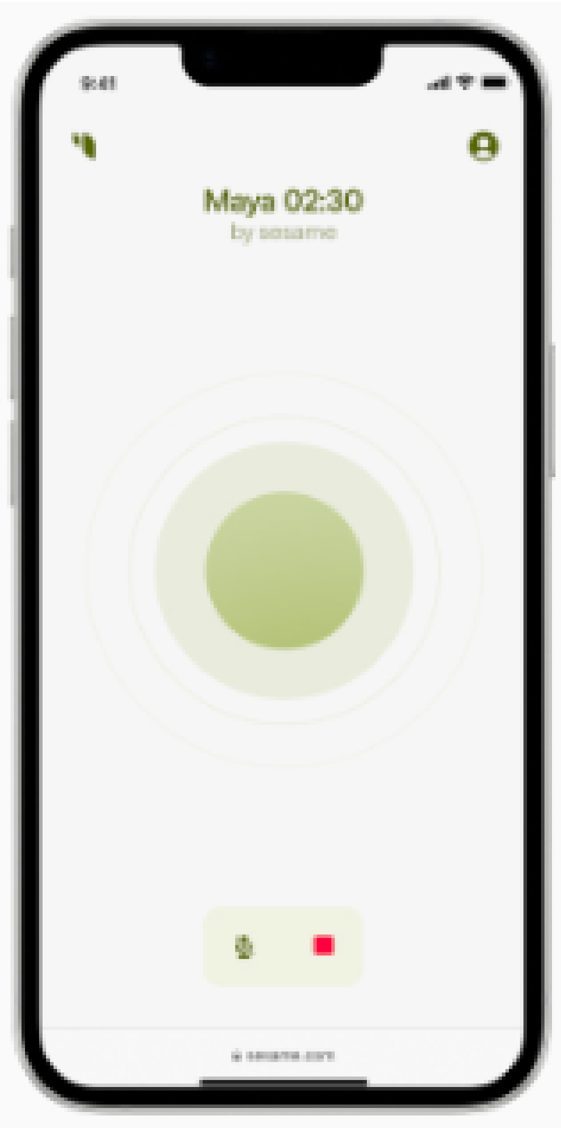
ChatGPT
voice mode



Gemini
Live

Example

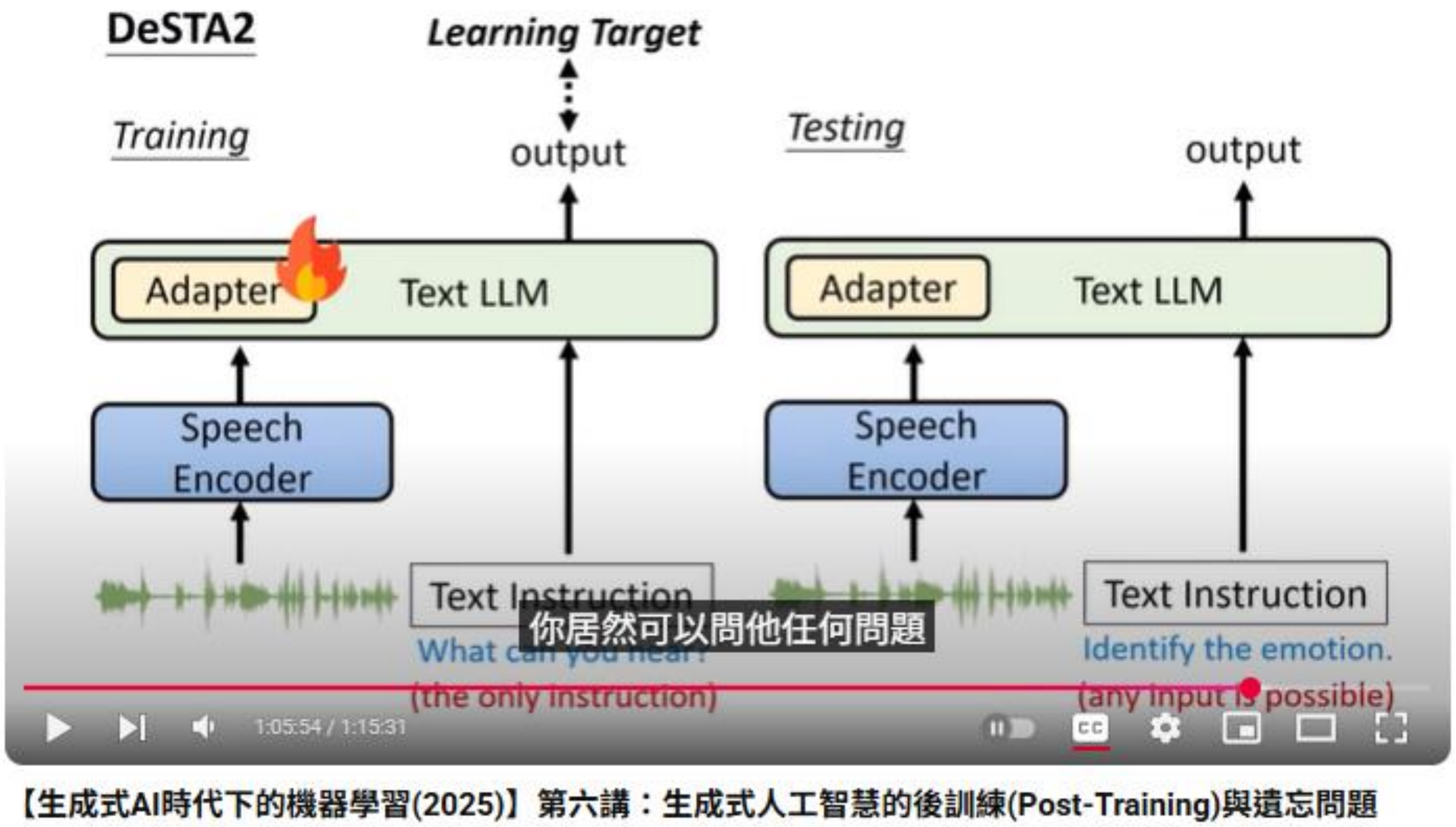
- Moshi
  - https://arxiv.org/abs/2410.00037
- GLM-4-Voice
  - https://arxiv.org/abs/2412.02612
- Step-Audio
  - https://arxiv.org/abs/2502.11946
- Qwen2.5-Omni
  - https://arxiv.org/abs/2503.20215
- Kimi-Audio
  - https://arxiv.org/abs/2504.18425
- SpeechGPT
  - https://github.com/OpenMOSS/SpeechGPT-2.0-preview
- ……



Sesame

https://www.sesame.com/research/crossing_the_uncanny_valley_of_voice

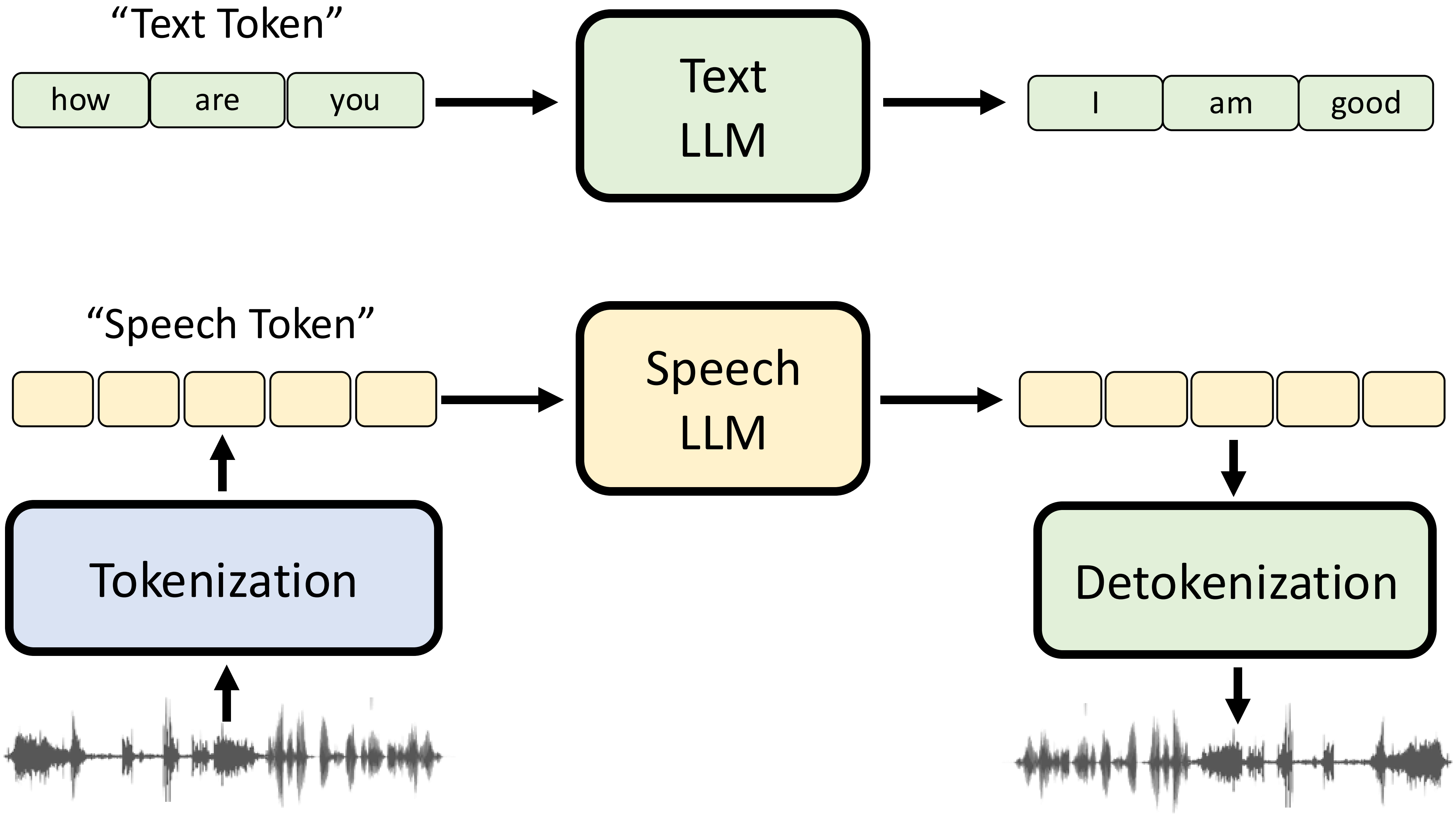
We have talked about speech input; this lecture will focus on speech generation.



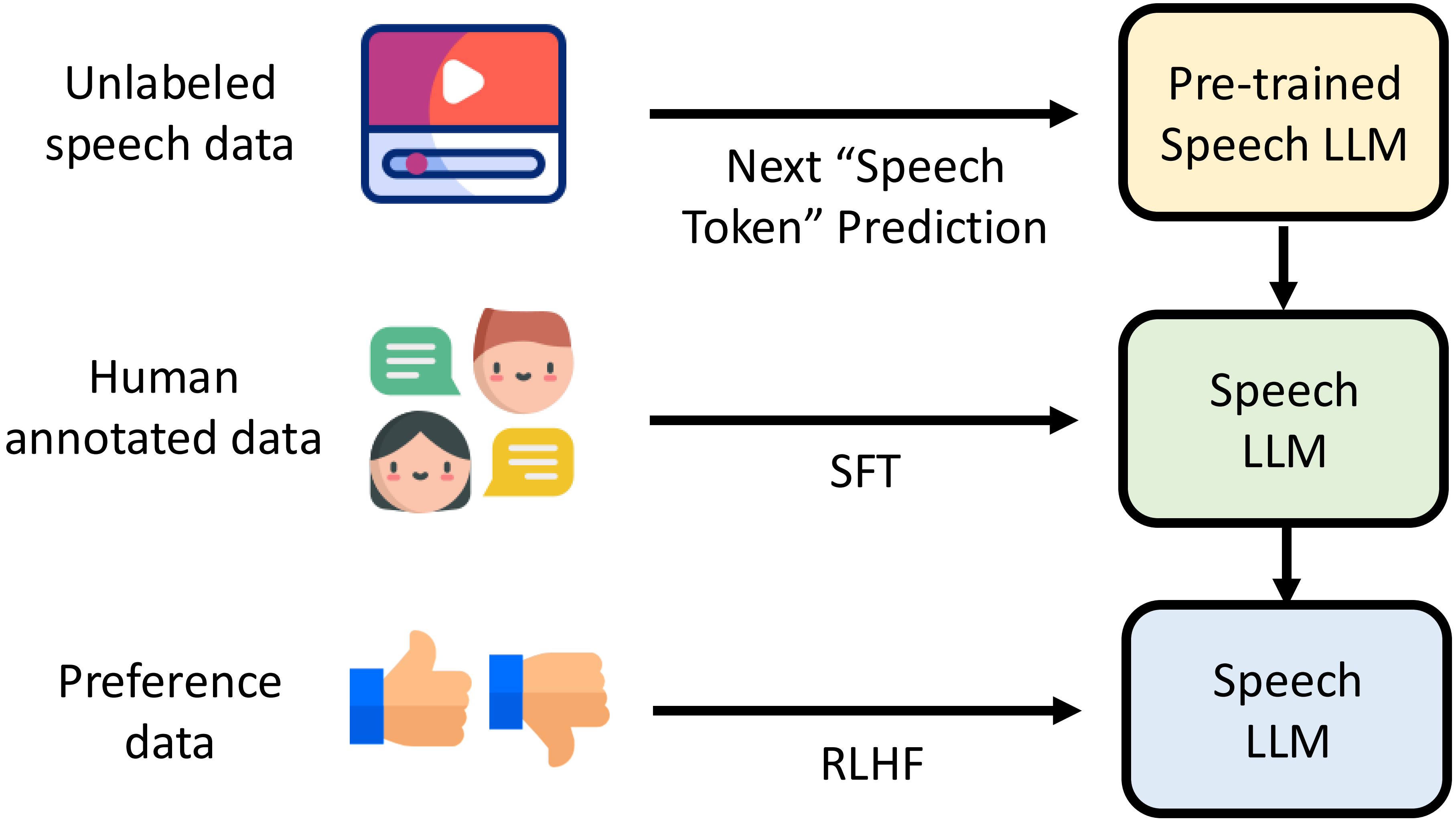
你居然可以問他任何問題

【生成式AI時代下的機器學習(2025)】第六講：生成式人工智慧的後訓練(Post-Training)與遺忘問題

https://youtu.be/Z6b5-77EfGk?si=st0d4IukGWAc__F2


"Text Token"
how
are
you
Text LLM
I
am
good
"Speech Token"
Speech LLM
Tokenization
Detokenization

How to Train Speech LLM



Unlabeled speech data

Next "Speech Token" Prediction

Pre-trained Speech LLM

Human annotated data

SFT

Speech LLM

Preference data

RLHF

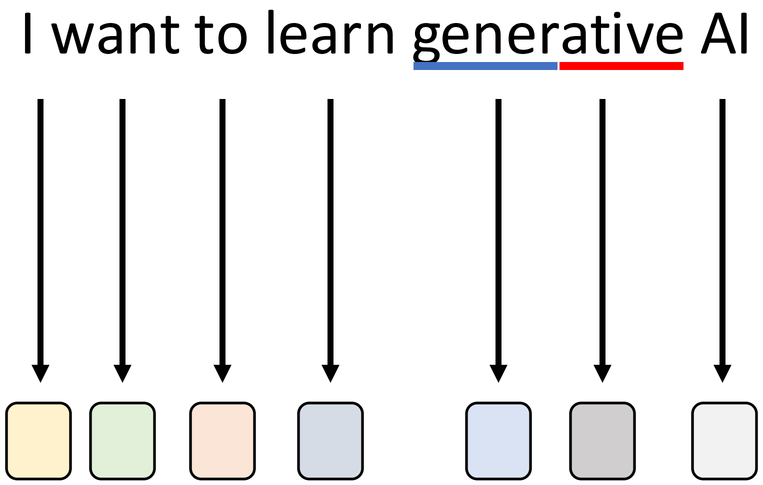
Speech LLM

語音生成的基本單位是什麼？

(Speech Token)

What is a “token” in the context of speech?

- Text

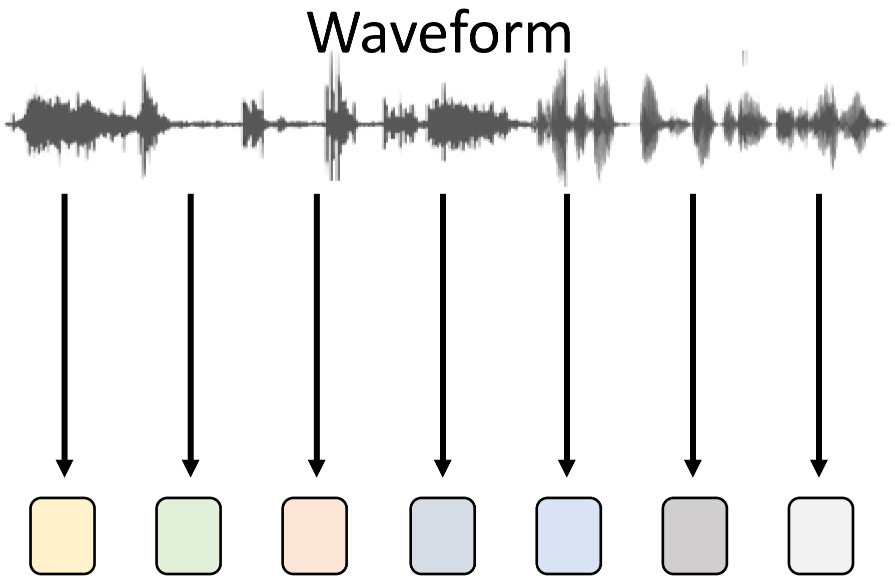


I want to learn generative AI

Token Sequence

https://platform.openai.com/tokenizer

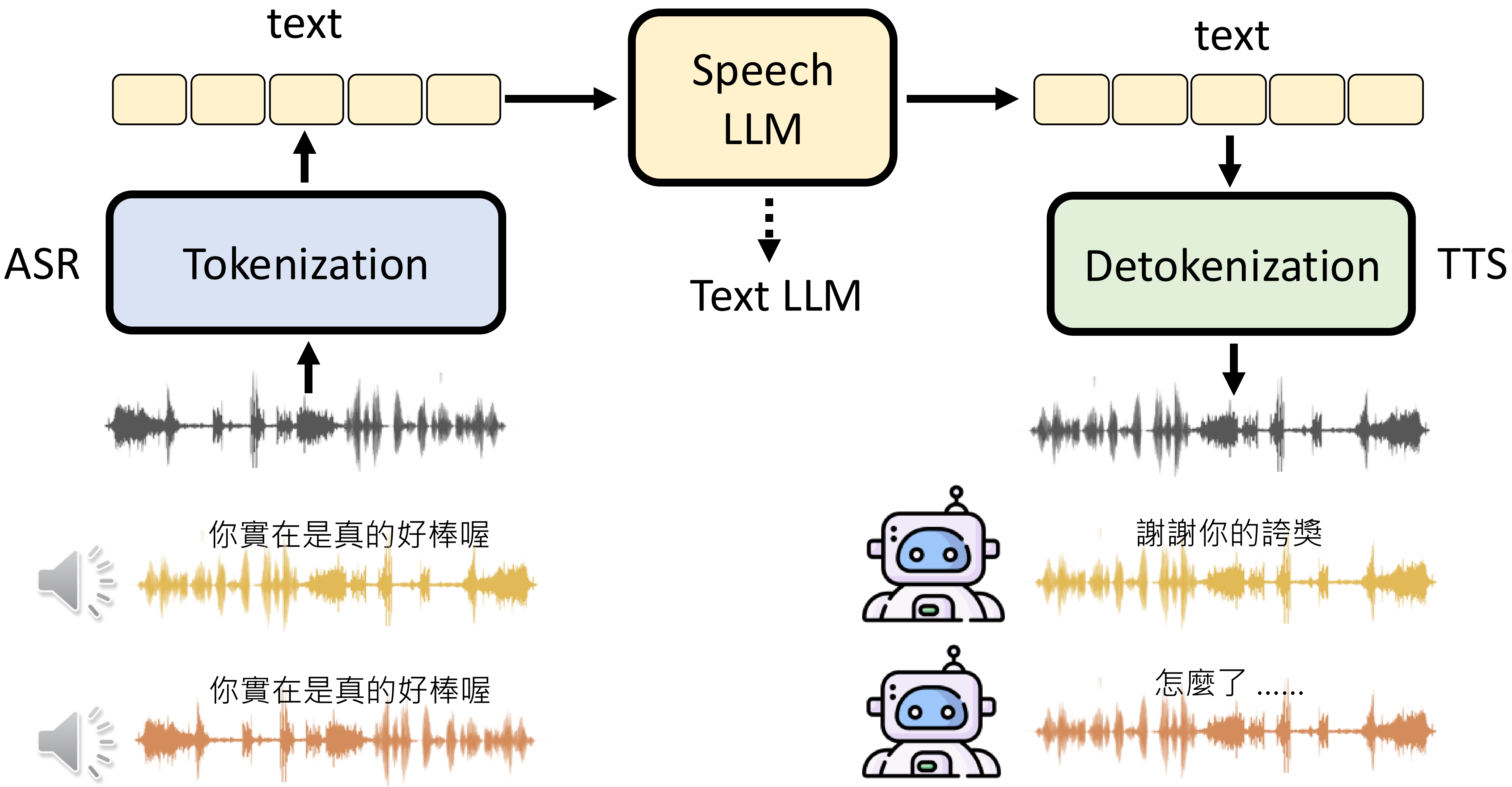
- Speech



Token Sequence

???

用 gpt-4o-mini-tts 合成



text

Speech LLM

text

ASR

Tokenization

Text LLM

Detokenization

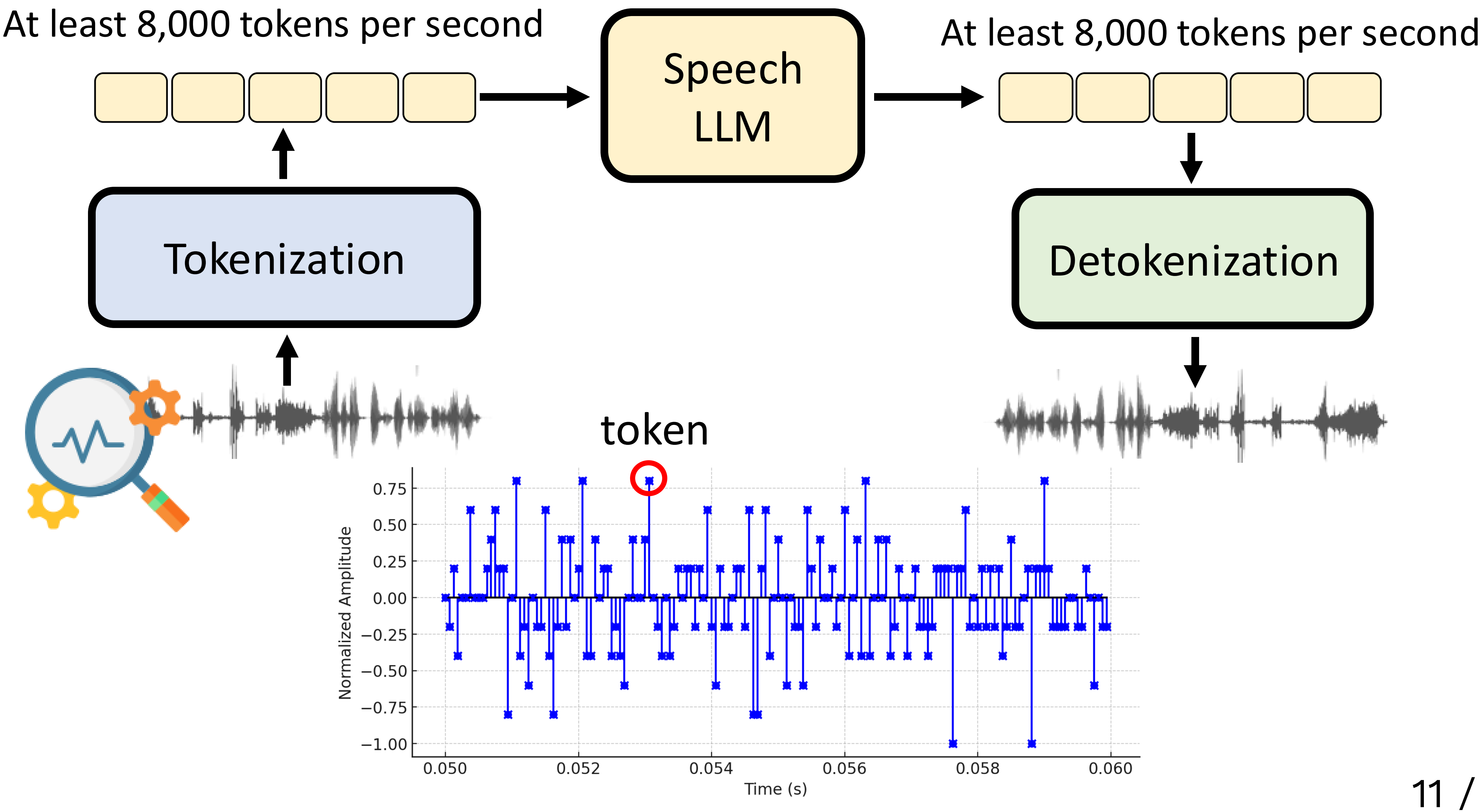
TTS

你實在是真的好棒喔

謝謝你的誇獎

你實在是真的好棒喔

怎麼了 ......



At least 8,000 tokens per second

Speech LLM

At least 8,000 tokens per second

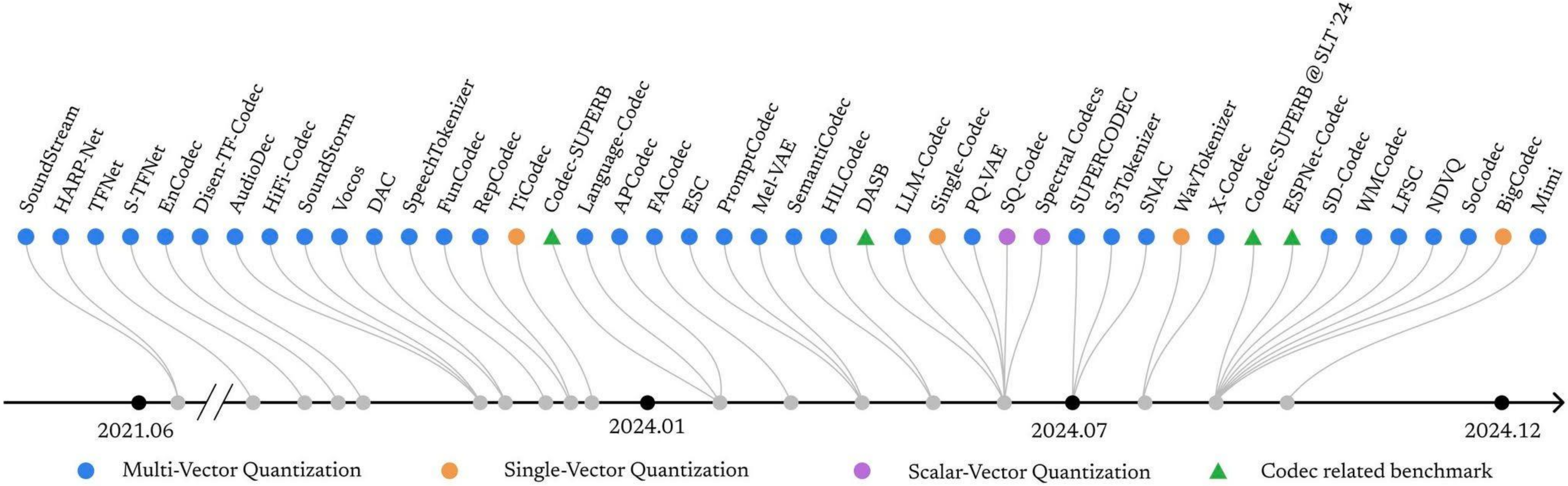
Tokenization

Detokenization

token

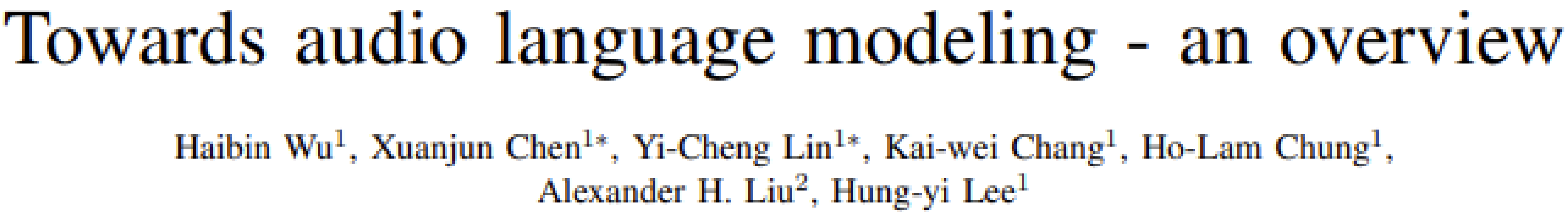
Various Types of Speech Tokenizers

Haibin Wu



Source of image: https://www.linkedin.com/in/haibin-wu-479a39252/recent-activity/all/

Overview paper about Speech Tokenization



Towards audio language modeling - an overview

Haibin Wu[1], Xuanjun Chen[1*], Yi-Cheng Lin[1*], Kai-wei Chang[1], Ho-Lam Chung[1],
Alexander H. Liu[2], Hung-yi Lee[1]

https://arxiv.org/abs/2402.13236

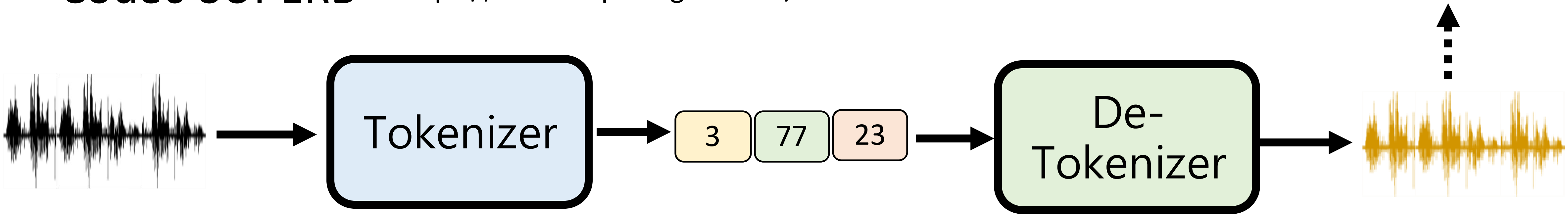
Recent Advances in Discrete Speech Tokens: A Review

Yiwei Guo, Zhihan Li, Hankun Wang, Bohan Li, Chongtian Shao, Hanglei Zhang, Chenpeng Du, Xie Chen,
Shujie Liu, Kai Yu

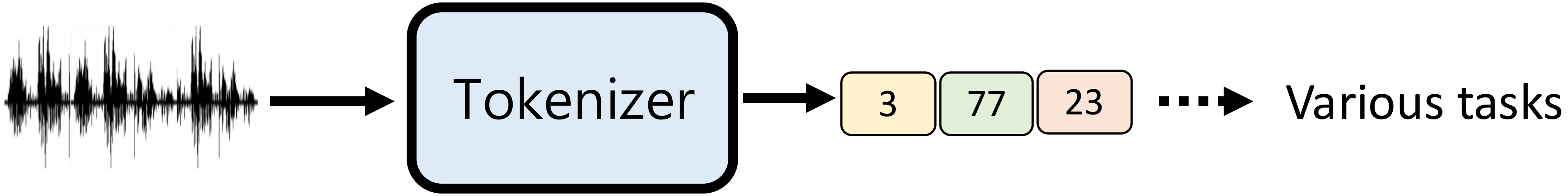
https://arxiv.org/abs/2502.06490

What is the best choice of tokens?

- Codec-SUPERB https://codecsuperb.github.io/

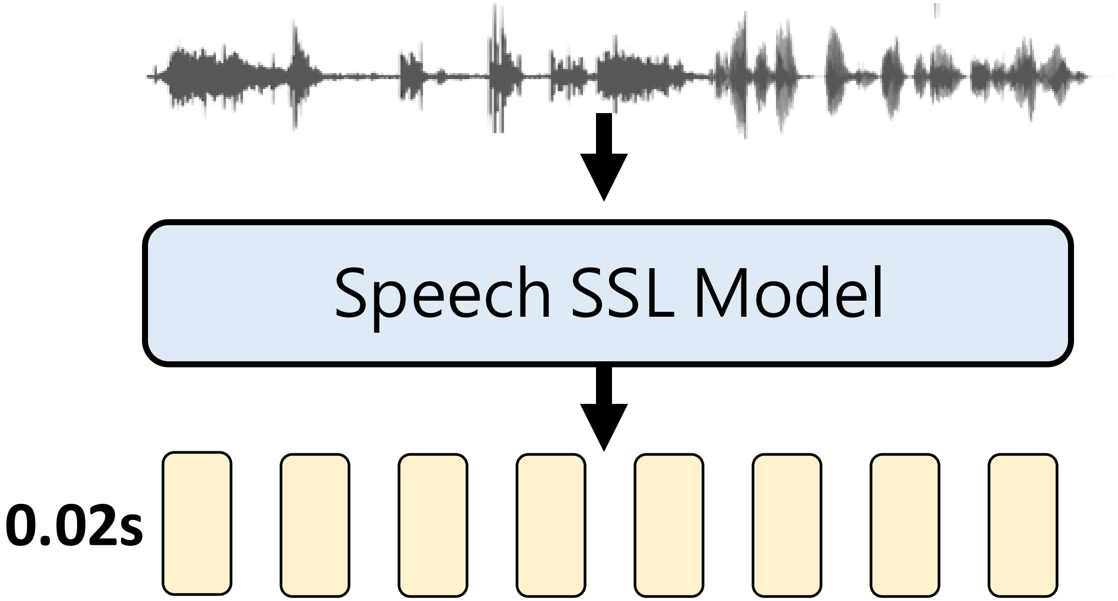


- DASB https://poonehmousavi.github.io/DASB-website/

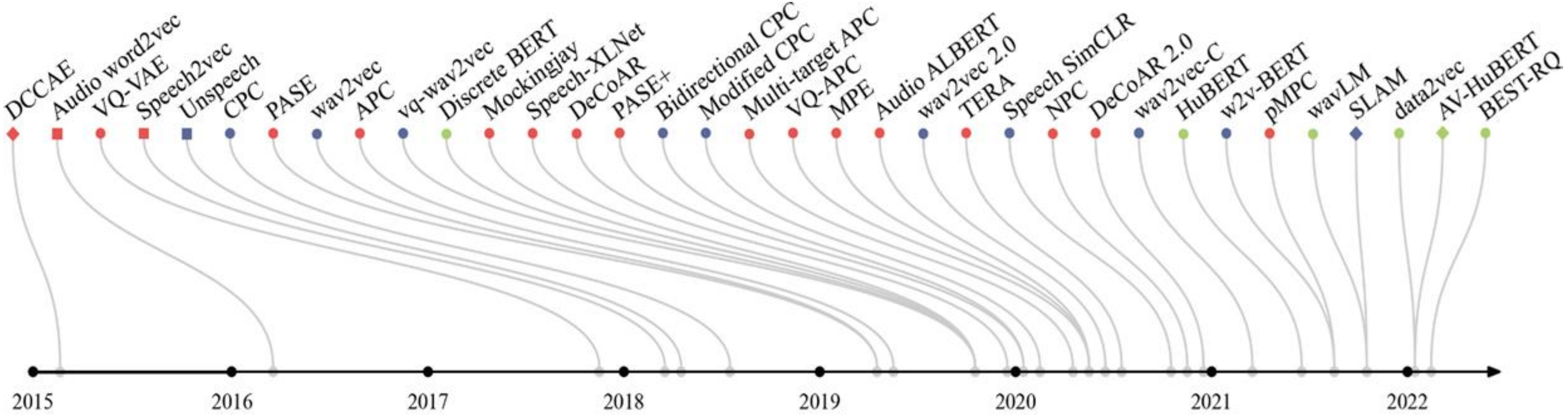


Learn more from Interspeech2024 Speech Processing Using Discrete Speech Unit Challenge https://www.wavlab.org/activities/2024/Interspeech2024-Discrete-Speech-Unit-Challenge/

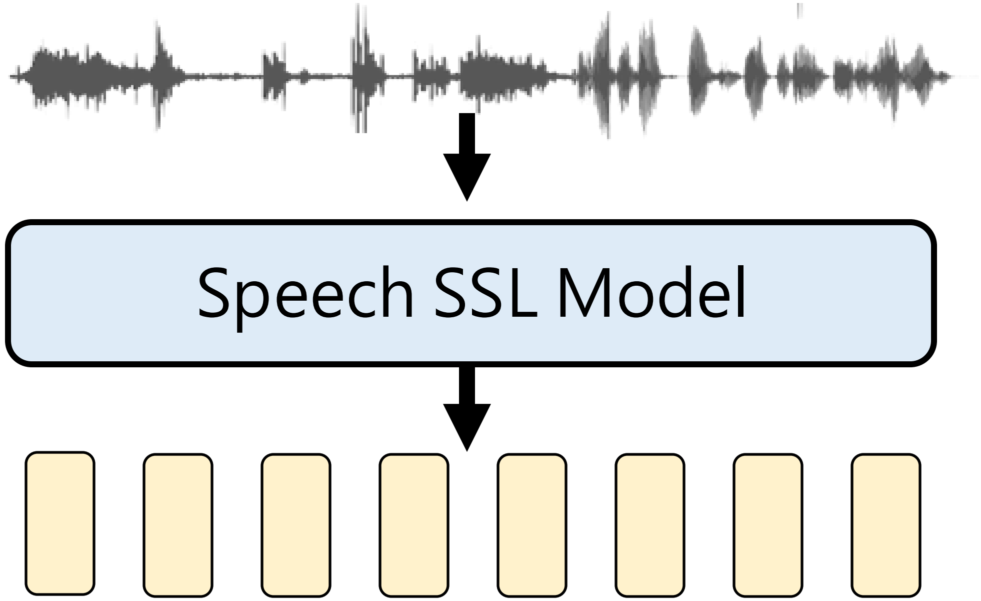
A possible pipeline of speech tokenization

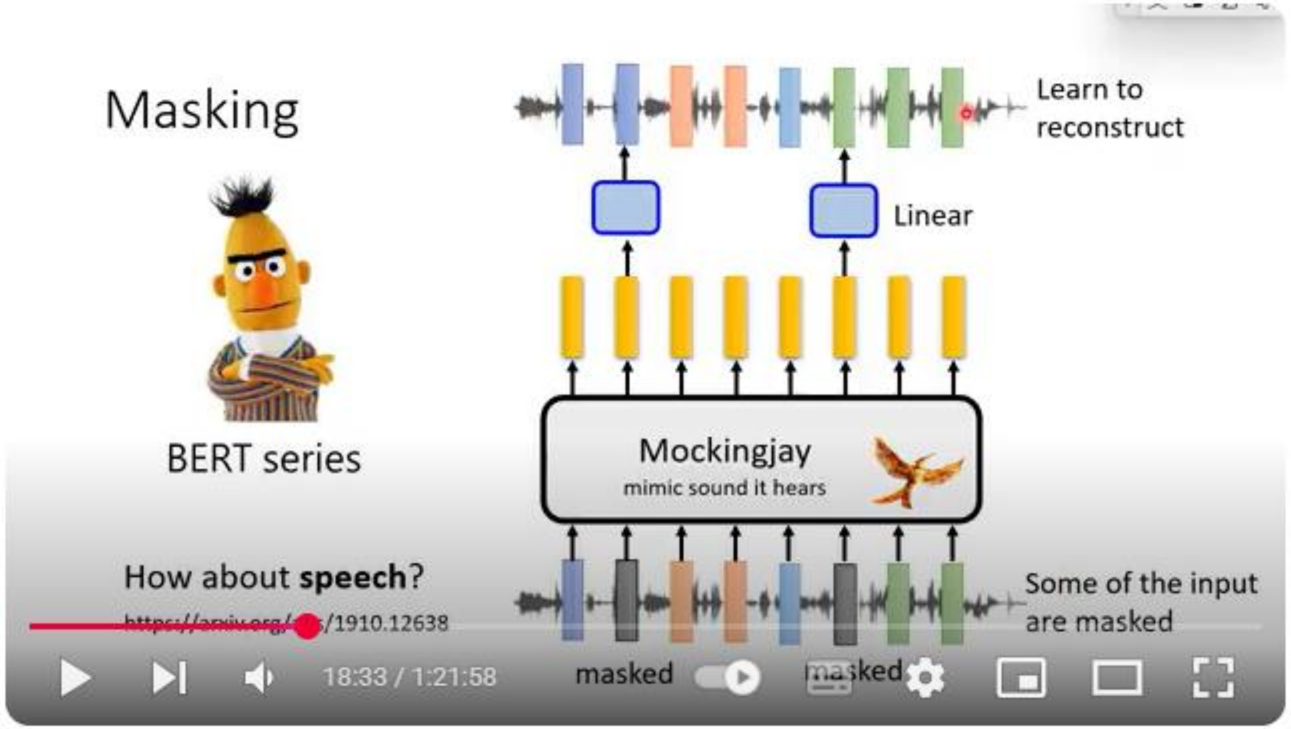


https://arxiv.org/abs/2205.10643



A possible pipeline of speech tokenization

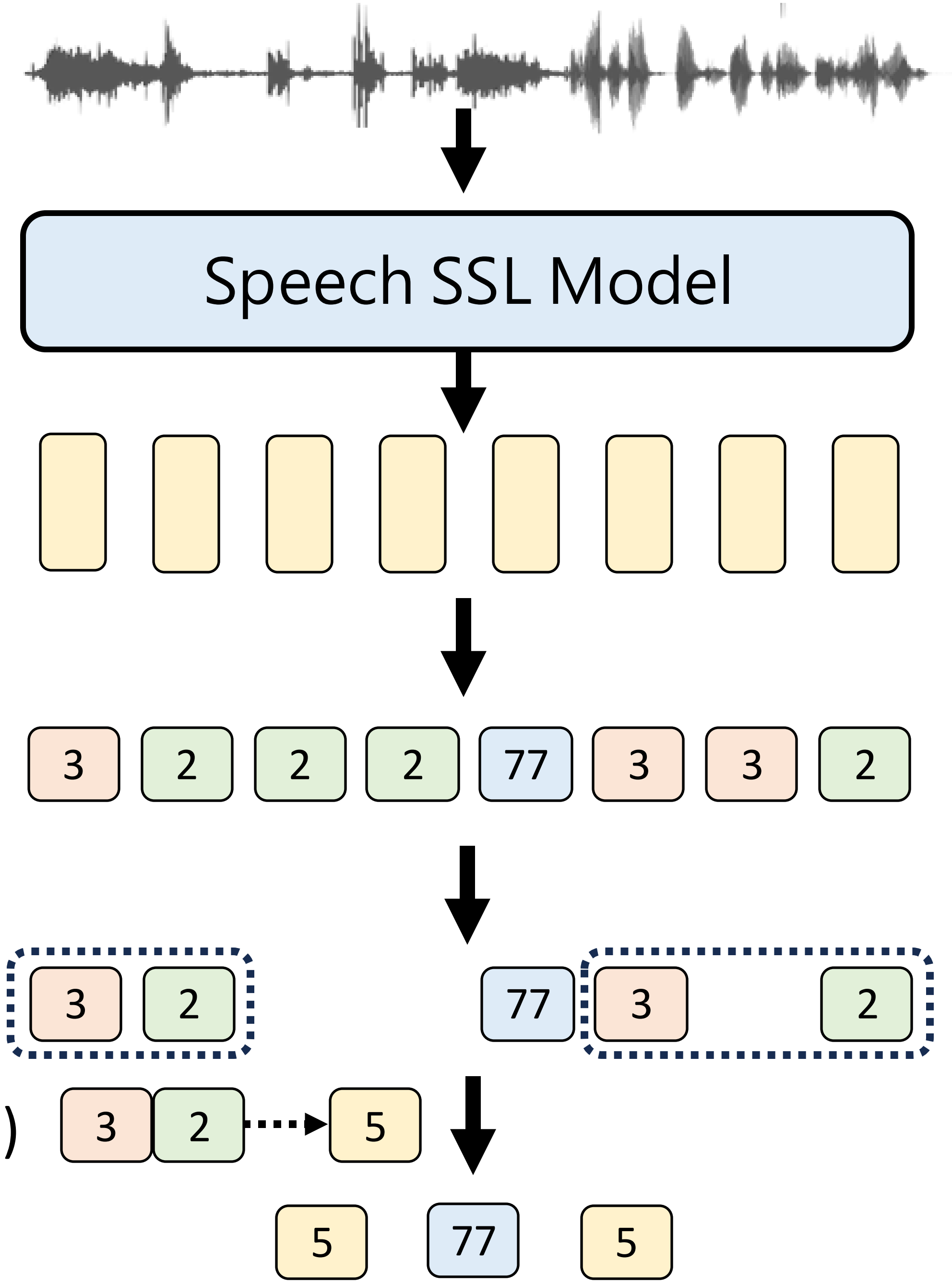




【機器學習 2022】語音與影像上的神奇自督導式學習 (Self-supervised Learning) 模型

https://www.youtube.com/watch?v=lMIN1iKYNmA

A possible pipeline of speech tokenization

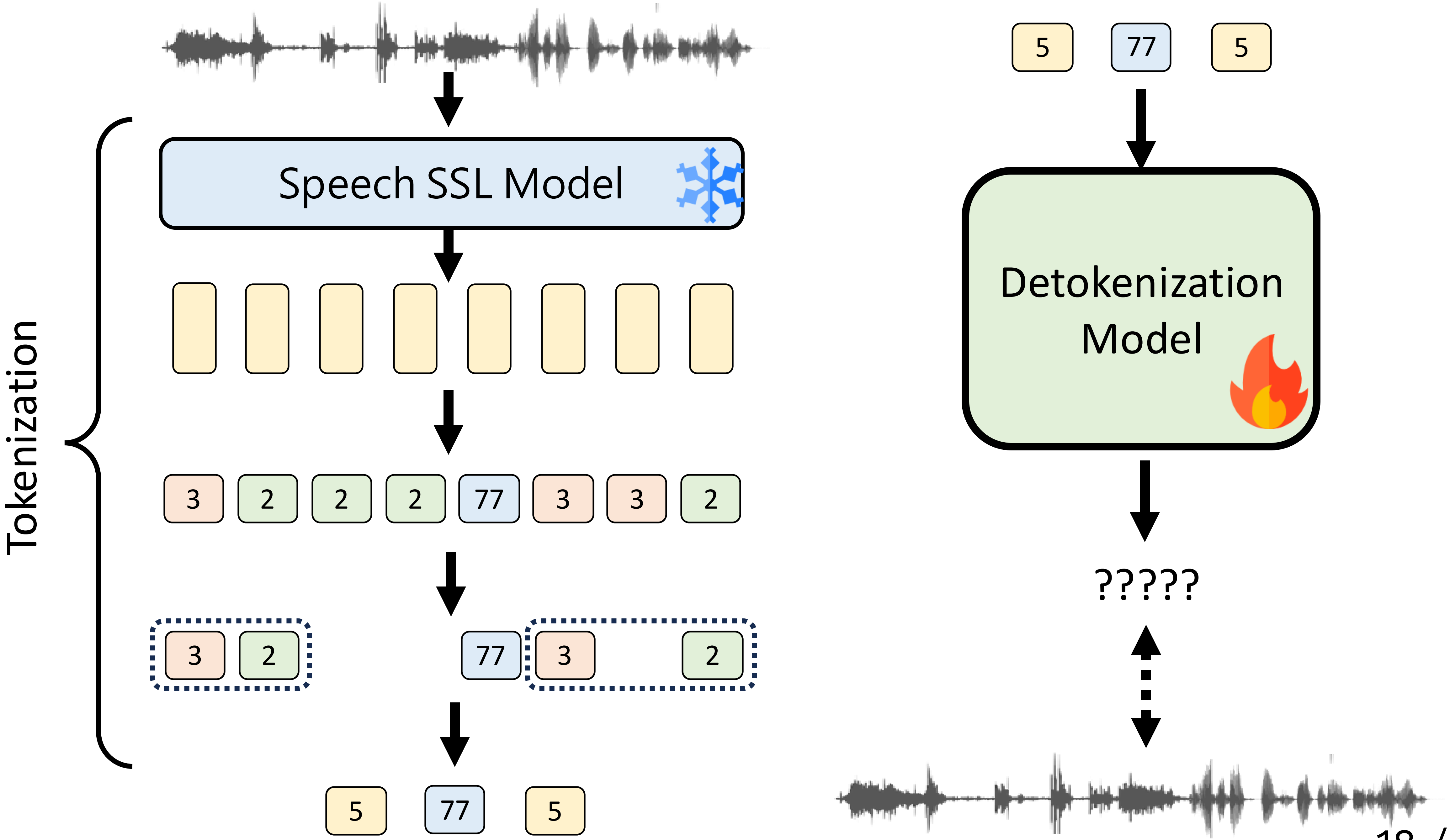


Quantization: K-means or VQ-layer

Deduplicate

BPE (Byte Pair Encoding)

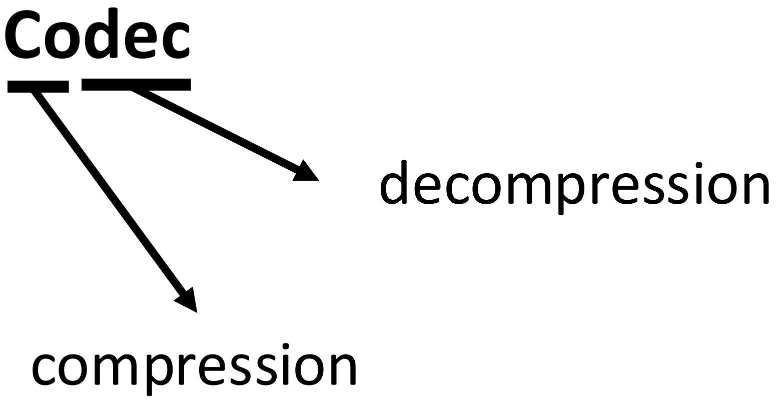
https://arxiv.org/abs/2310.14580
http://arxiv.org/abs/2205.01086
https://ieeexplore.ieee.org/abstract/document/10096788

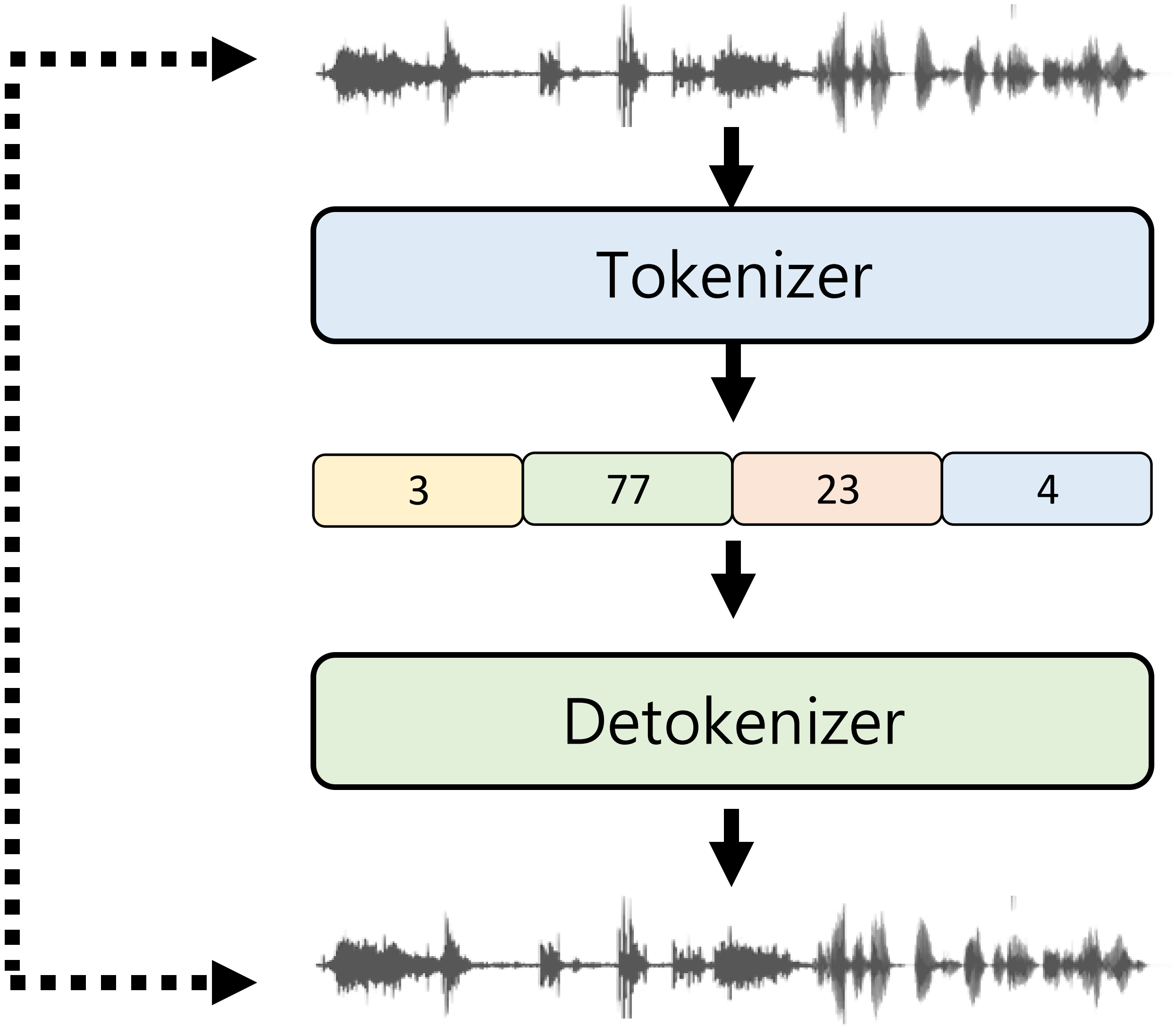

Tokenization
Speech SSL Model
3
2
2
2
77
3
3
2
3
2
77
3
2
5
77
5
5
77
5
Detokenization Model
?????

Another possible pipeline of speech tokenization

Neural Speech Codec

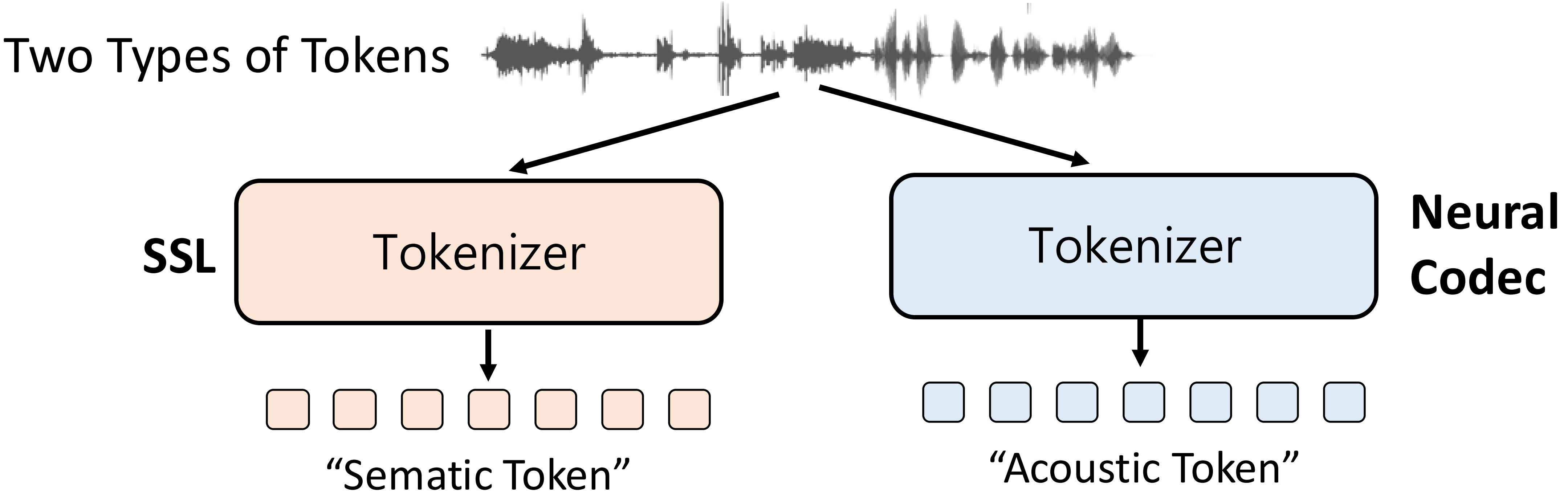
The tokenizer and detokenizer are learned jointly.





Audio LM
https://arxiv.org/abs/2209.03143

Various Types of Speech Tokenizers

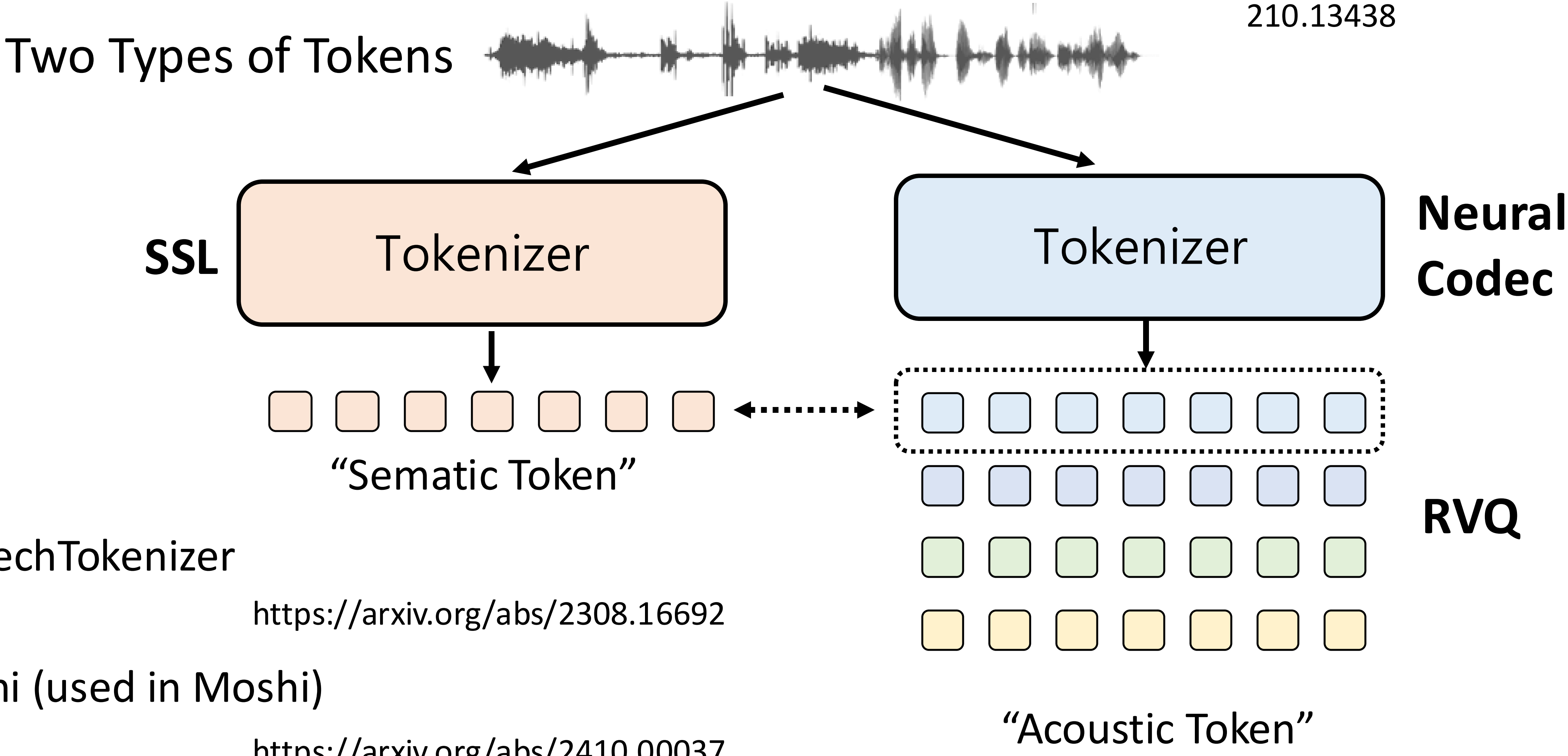


- "Semantic" does not refer to its usual meaning in linguistics. Instead, "semantic tokens" are closer to content information (usually containing phonetic information).
- The distinction between the two types can be vague. 'Semantic tokens' also include acoustic information, and vice versa.

Various Types of Speech Tokenizers

RVQ (Residual vector Quantization)

https://arxiv.org/abs/2210.13438



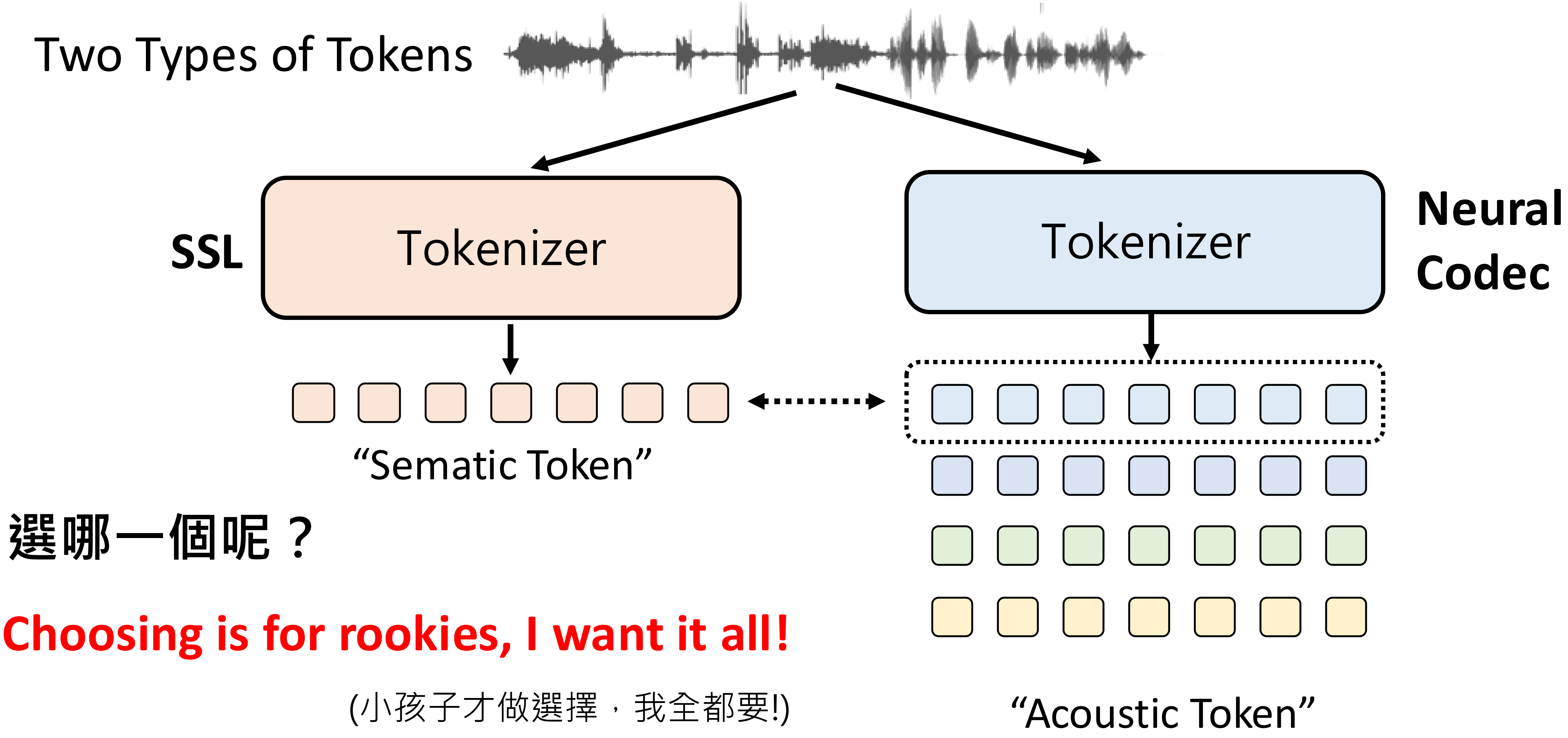
SpeechTokenizer

https://arxiv.org/abs/2308.16692

Mimi (used in Moshi)

https://arxiv.org/abs/2410.00037

Various Types of Speech Tokenizers

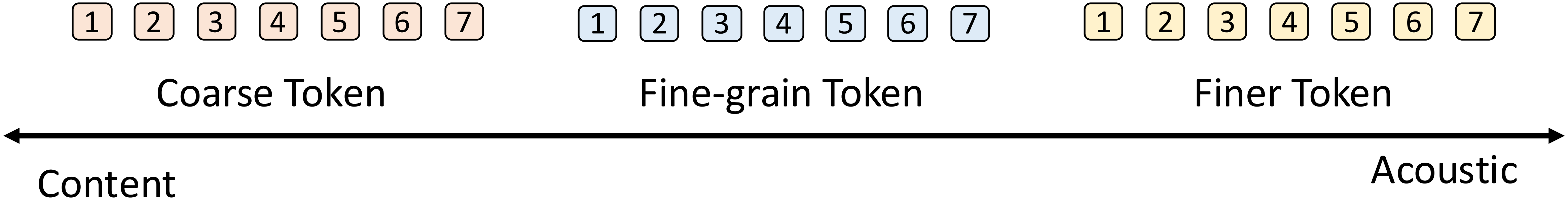


選哪一個呢？

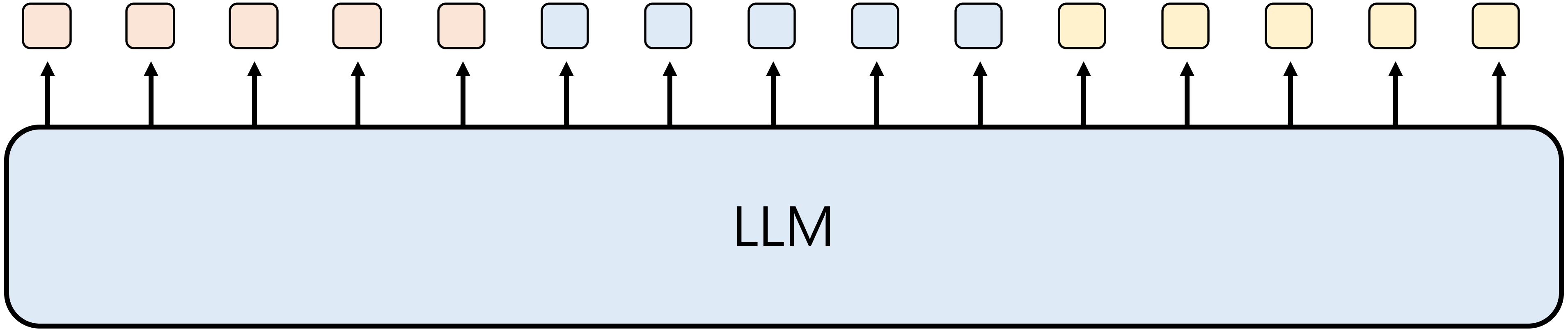
Choosing is for rookies, I want it all!

(小孩子才做選擇，我全都要!)

Choice of Decoding Strategies



Assumption: All tokens are of equal length for simplicity.



e.g., AudioLM, VALLE

https://arxiv.org/abs/2209.03143

https://arxiv.org/abs/2301.02111

Choice of Decoding Strategies

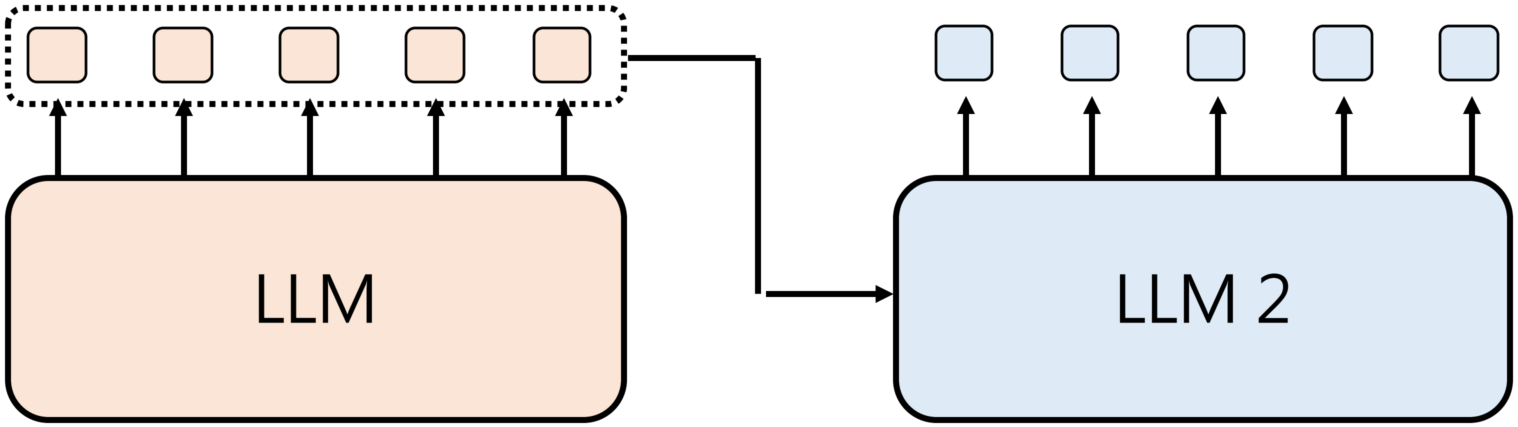
1 2 3 4 5 6 7
1 2 3 4 5 6 7
1 2 3 4 5 6 7

Coarse Token

Fine-grain Token

Finer Token

In VALLE, LLM 2 is a non-autoregressive language model.

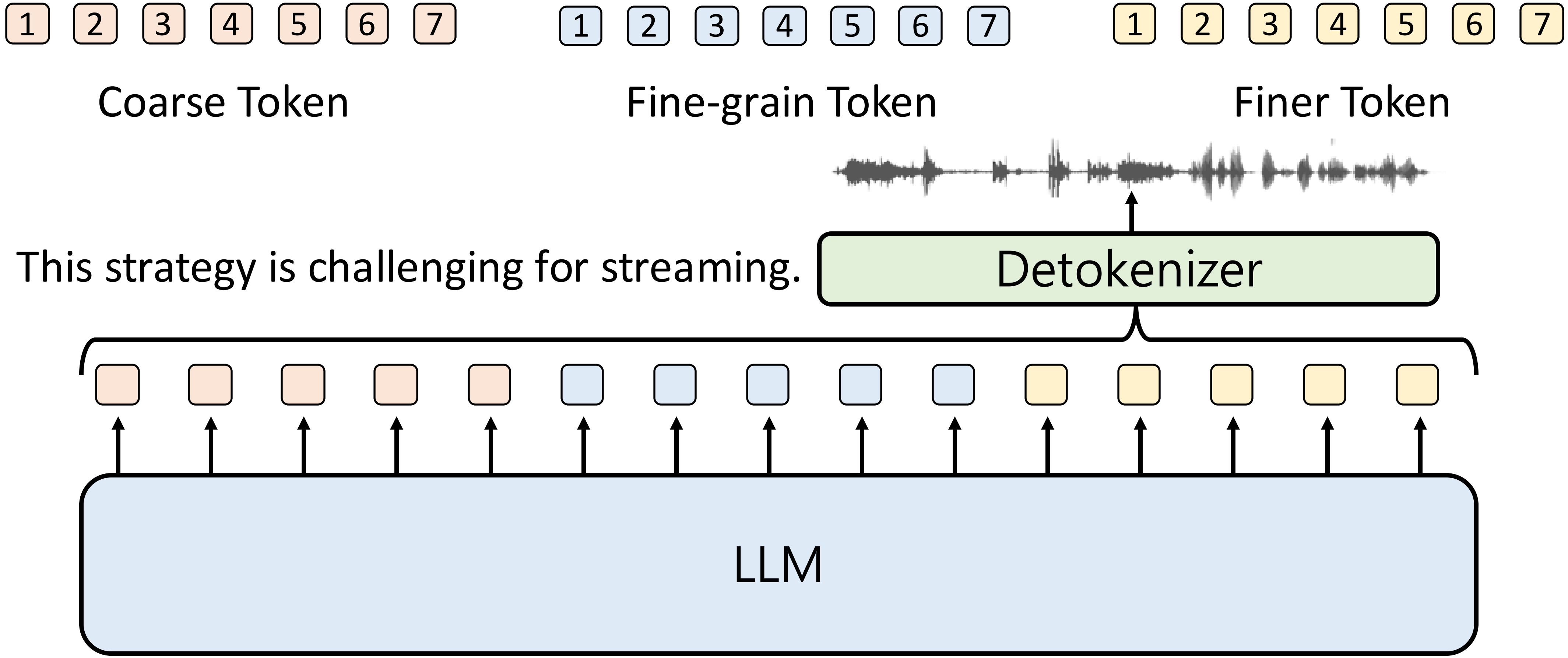


e.g., AudioLM, VALLE

https://arxiv.org/abs/2209.03143

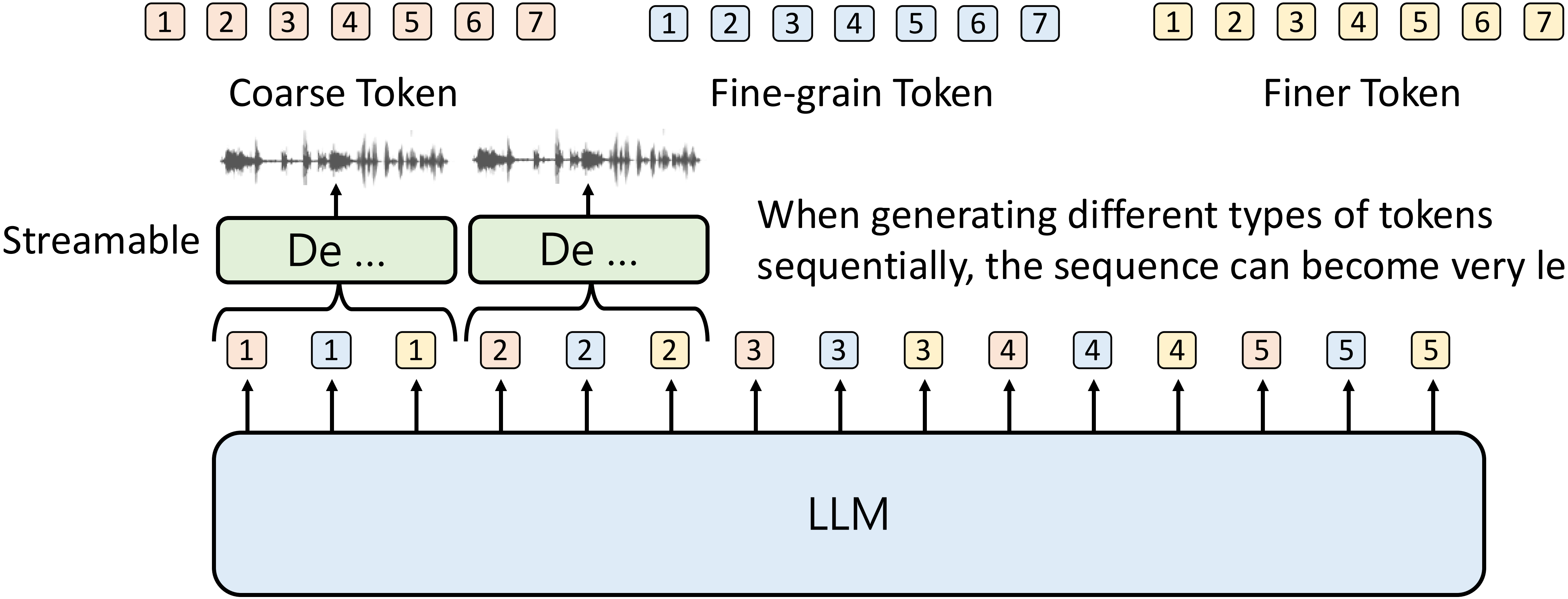
https://arxiv.org/abs/2301.02111

Choice of Decoding Strategies



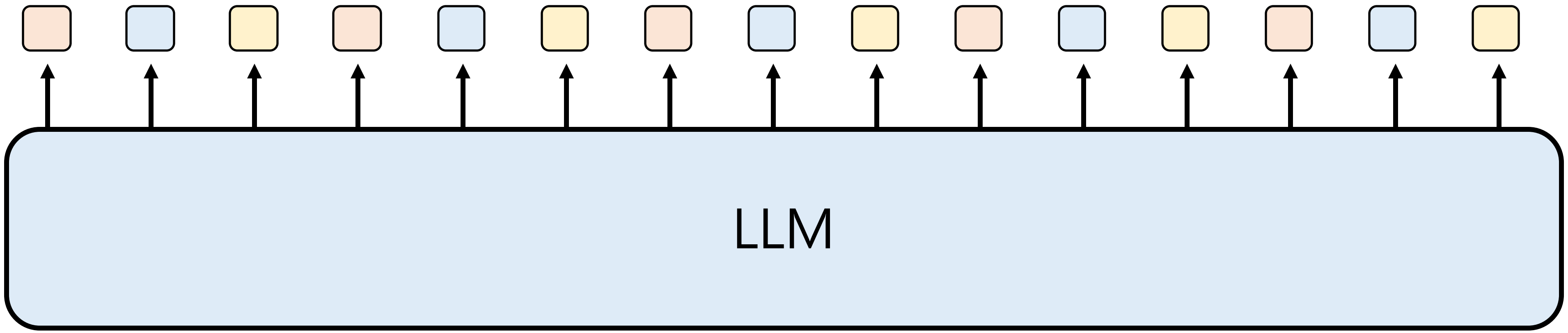
This strategy is challenging for streaming.

Choice of Decoding Strategies



When generating different types of tokens sequentially, the sequence can become very lengthy.

https://arxiv.org/abs/2402.05755

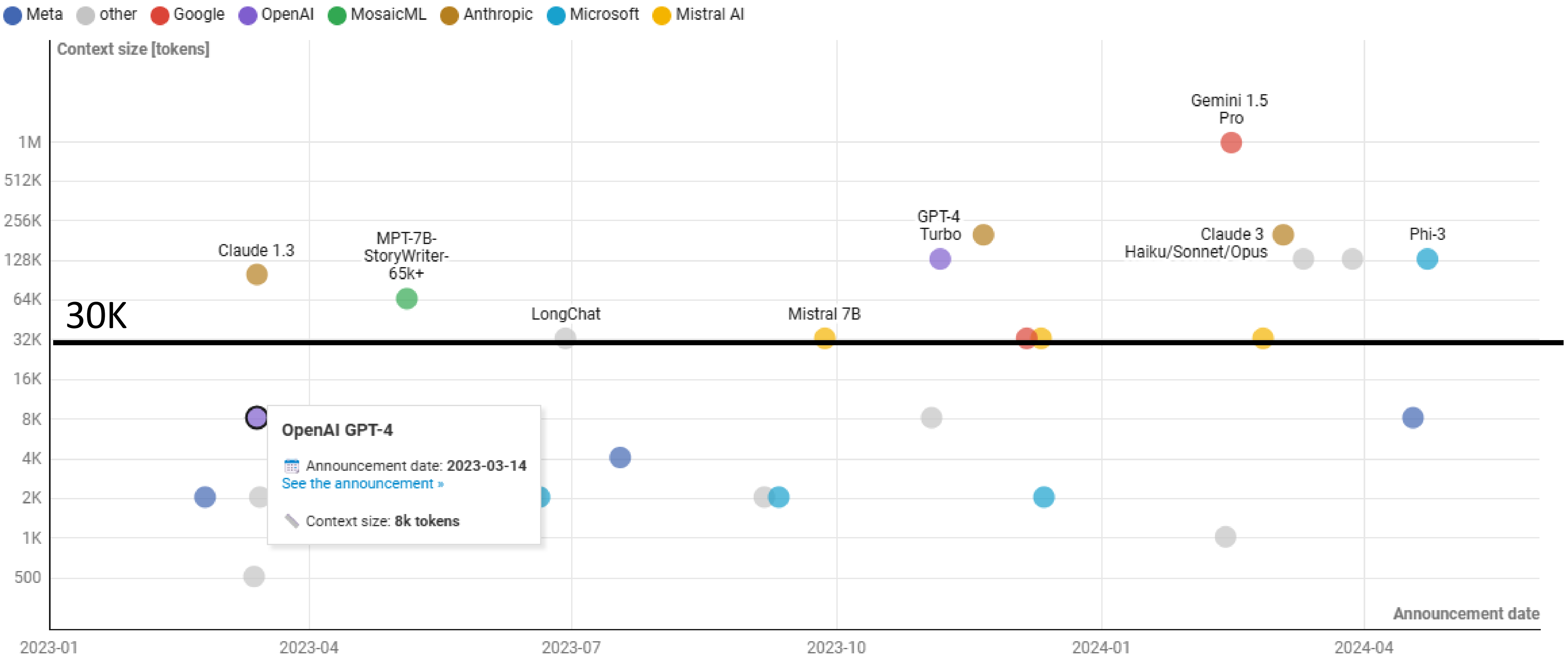


sequence length
= token per second x types of tokens x dialogue length

Take Moshi as example

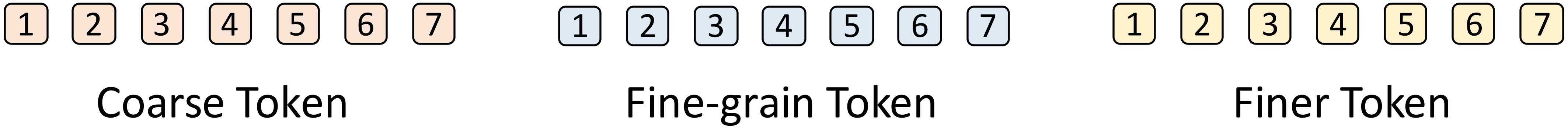
12.5Hz 8 5 mins (300 seconds)

= **30k**

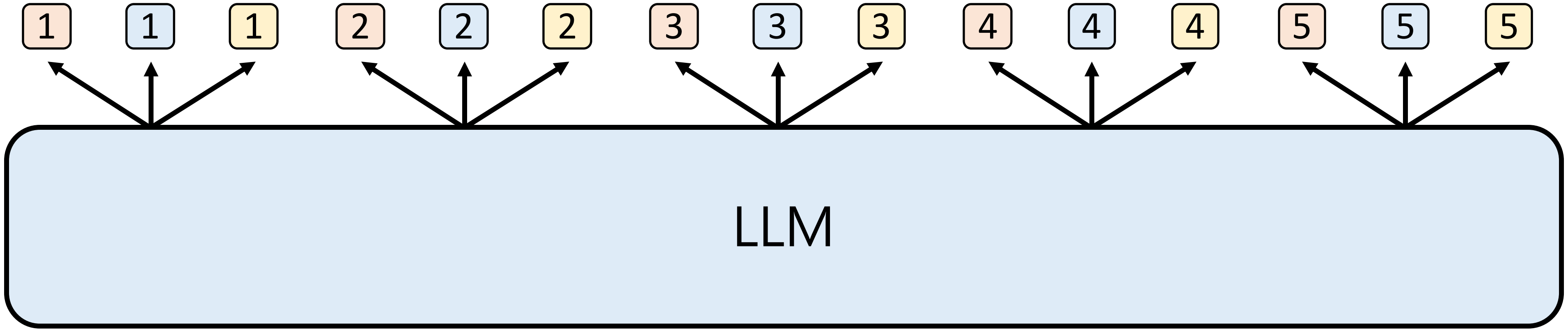


Source of image: https://towardsdatascience.com/towards-infinite-llm-context-windows-e099225abaaf

Choice of Decoding Strategies

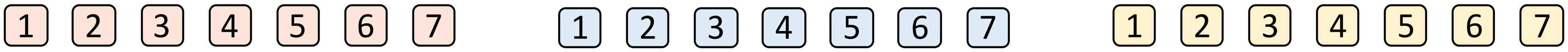


Generate multiple types of tokens in one step



https://arxiv.org/abs/2402.05755

Choice of Decoding Strategies



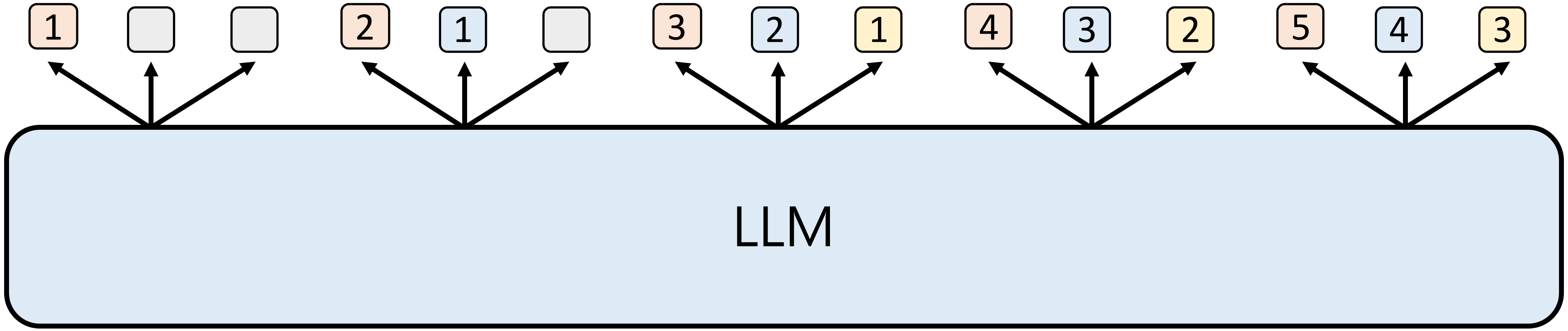
Coarse Token

Fine-grain Token

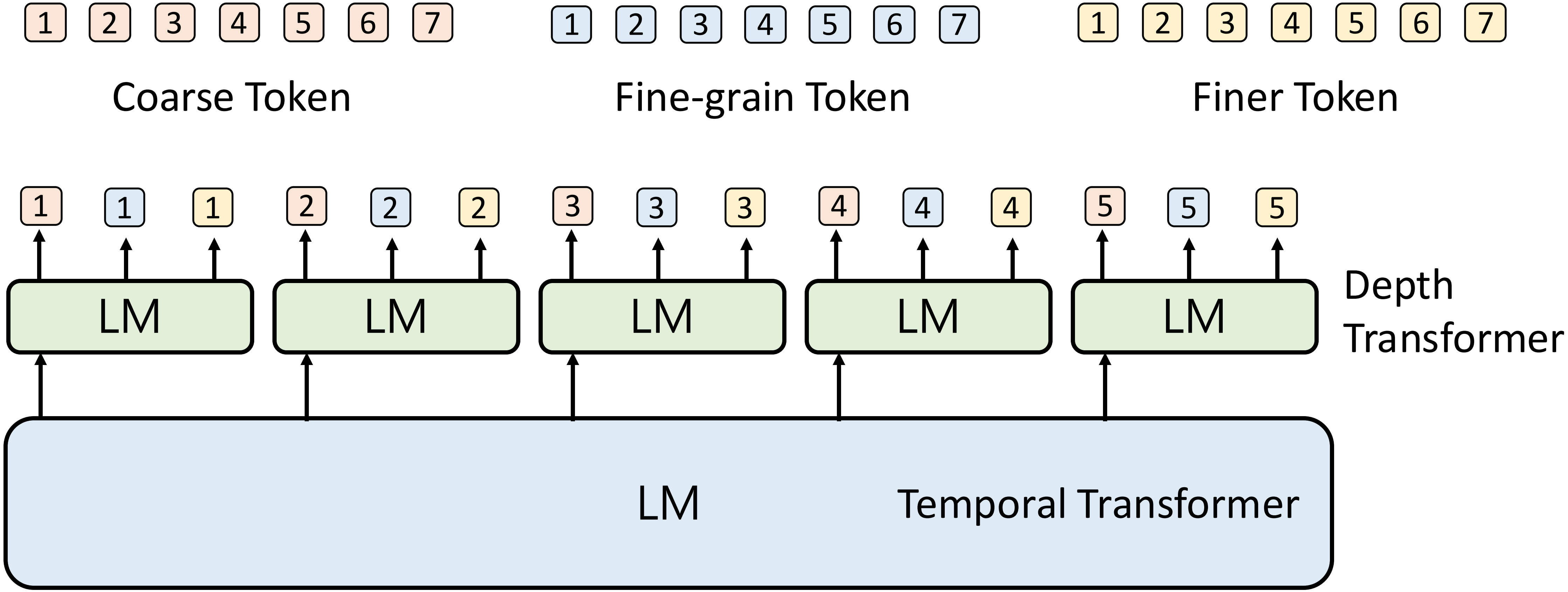
Finer Token

Acoustic Delay

https://arxiv.org/abs/2306.05284
https://arxiv.org/abs/2410.00037

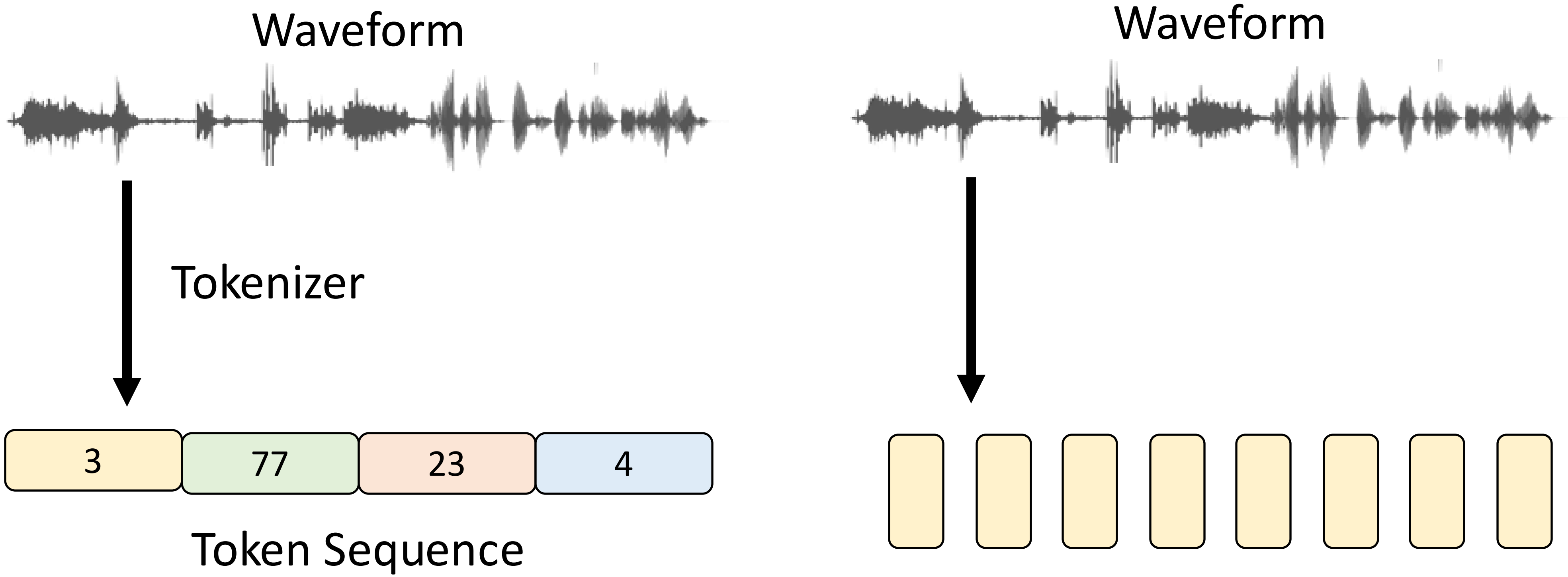


Choice of Decoding Strategies



https://arxiv.org/abs/2109.03264
https://arxiv.org/abs/2410.00037

Why discrete tokens?



Waveform

Tokenizer

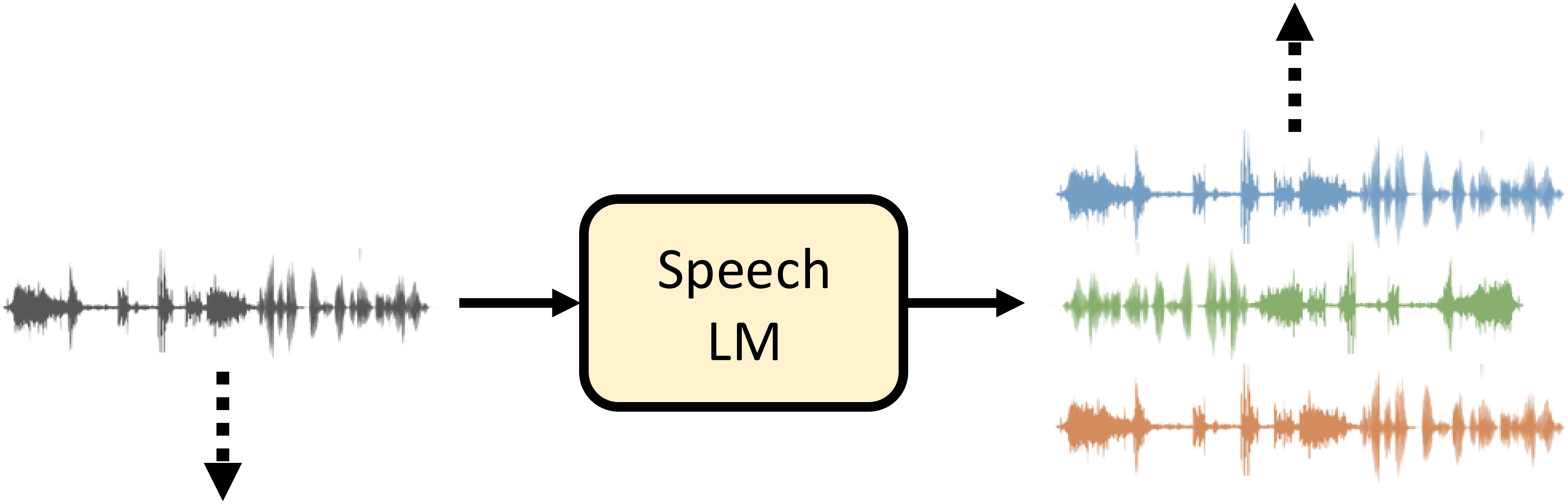
3 | 77 | 23 | 4

Token Sequence

Waveform

How about Continuous Representation?

Why discrete tokens?



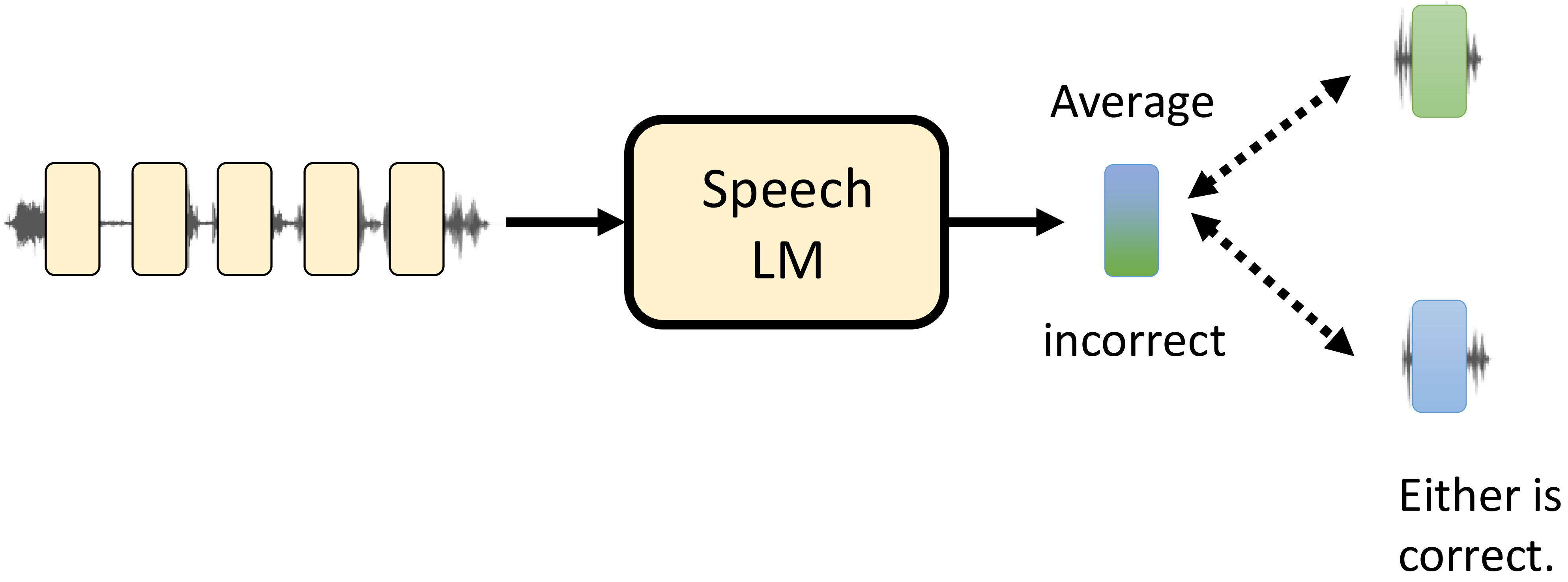
The discrete tokens are crucial for generation.

For understanding, there is no remarkable difference between continuous representations and discrete tokens.

Given the same input, there can be many possible outputs.

Why discrete tokens?

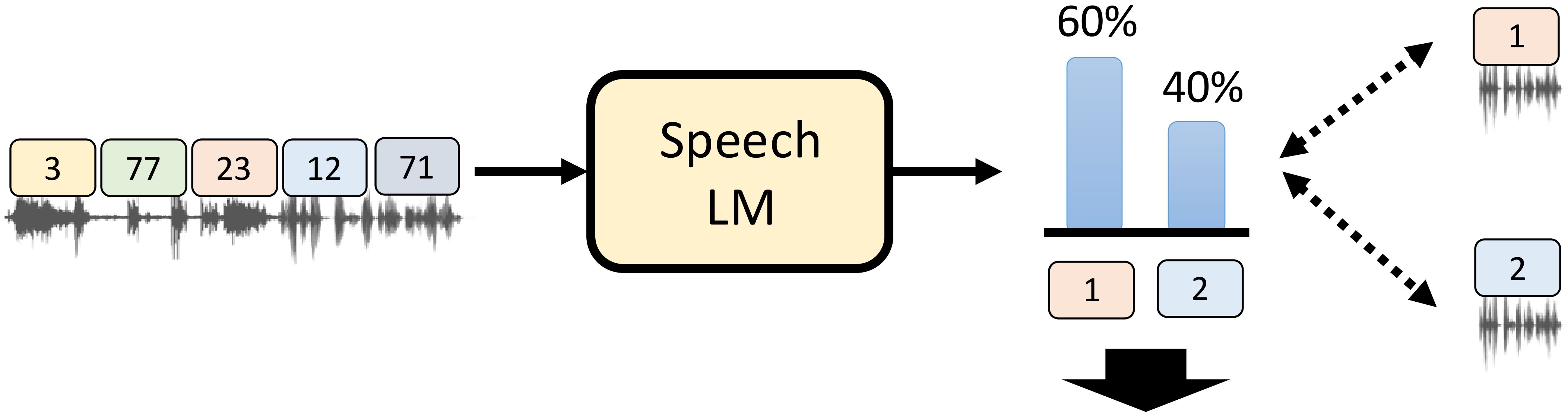
- Let's say we train a speech LM to generate continuous representations.



Why discrete tokens?

- How do discrete tokens solve the issue?

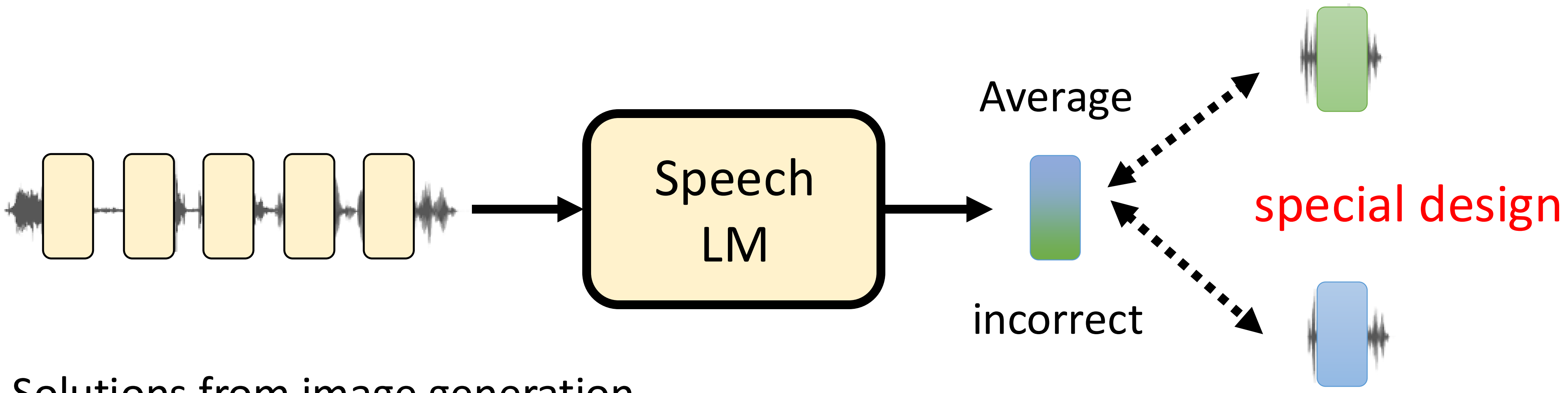
We learn a probability distribution



Sampling from the distribution during inference.

Why discrete tokens?

- Let's say we train a speech LM to generate continuous representations.



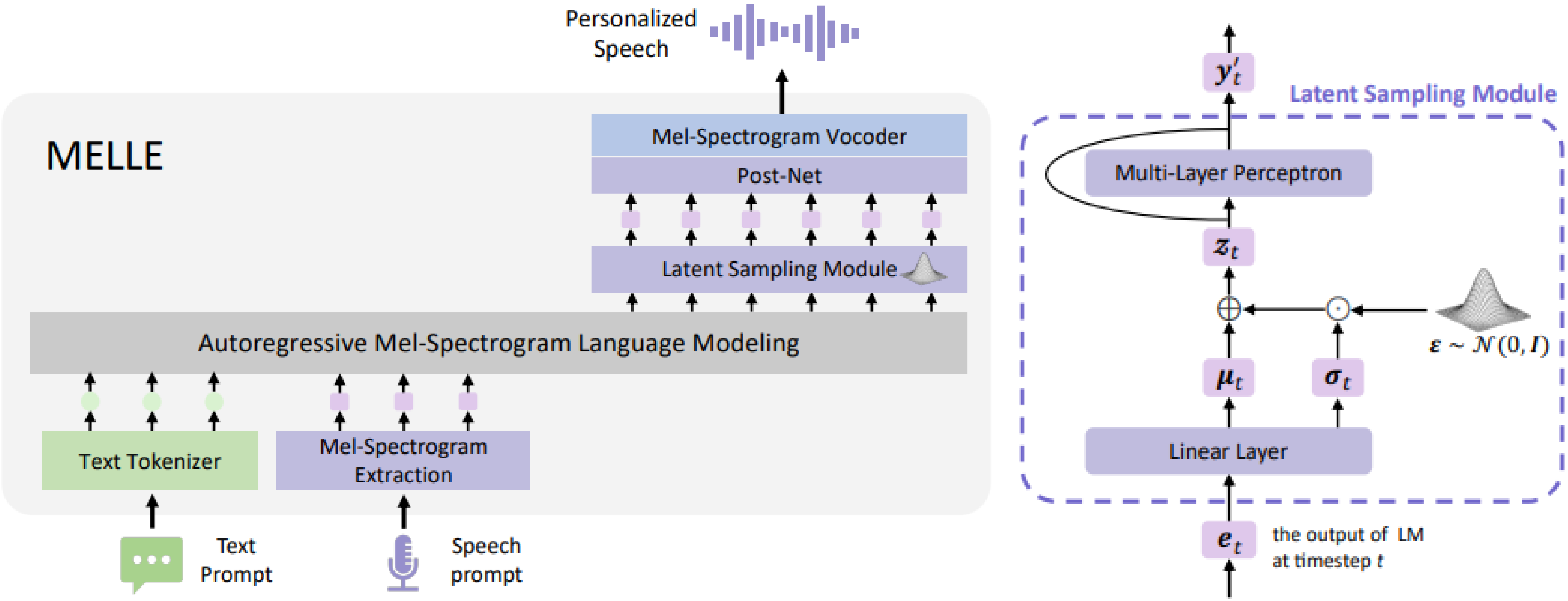
Either is correct.

Solutions from image generation.

https://arxiv.org/abs/2406.11838
https://arxiv.org/abs/2312.02116
https://arxiv.org/abs/2403.05196

MELLE

https://arxiv.org/pdf/2407.08551

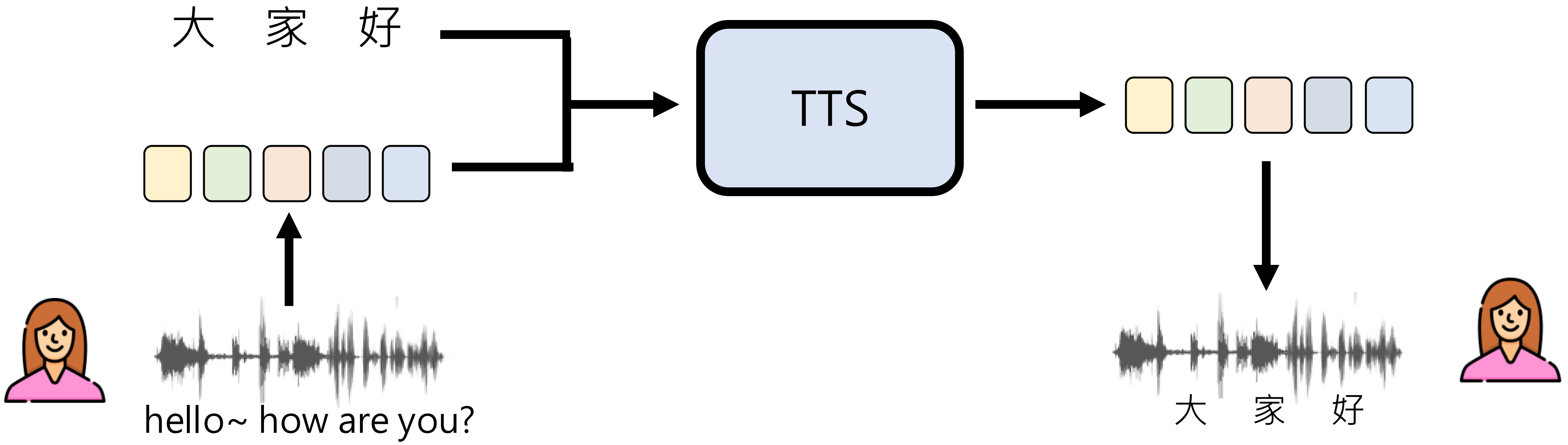


Good performance in Text-to-Speech (TTS)

Breezy Voice

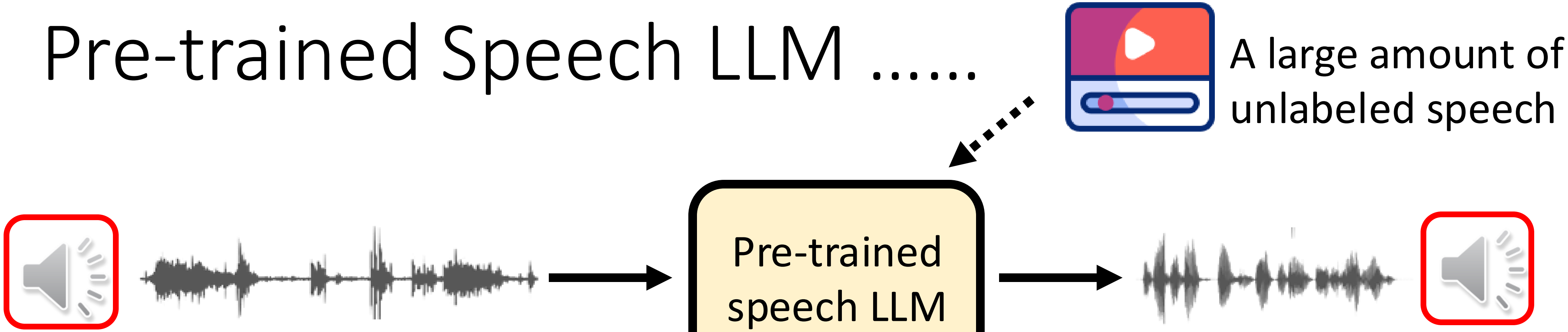
GitHub: https://github.com/mtkresearch/BreezyVoice

Paper: https://arxiv.org/abs/2501.17790



The source of the real audio is from the BIIC Podcast.

Pre-trained Speech LLM ......



A large amount of unlabeled speech data

Pre-trained speech LLM

He assassinated the president

He assassinated the president and gave mister johnson the last charge of improvement in his writing possible three point eight nine.

https://arxiv.org/abs/2306.02207

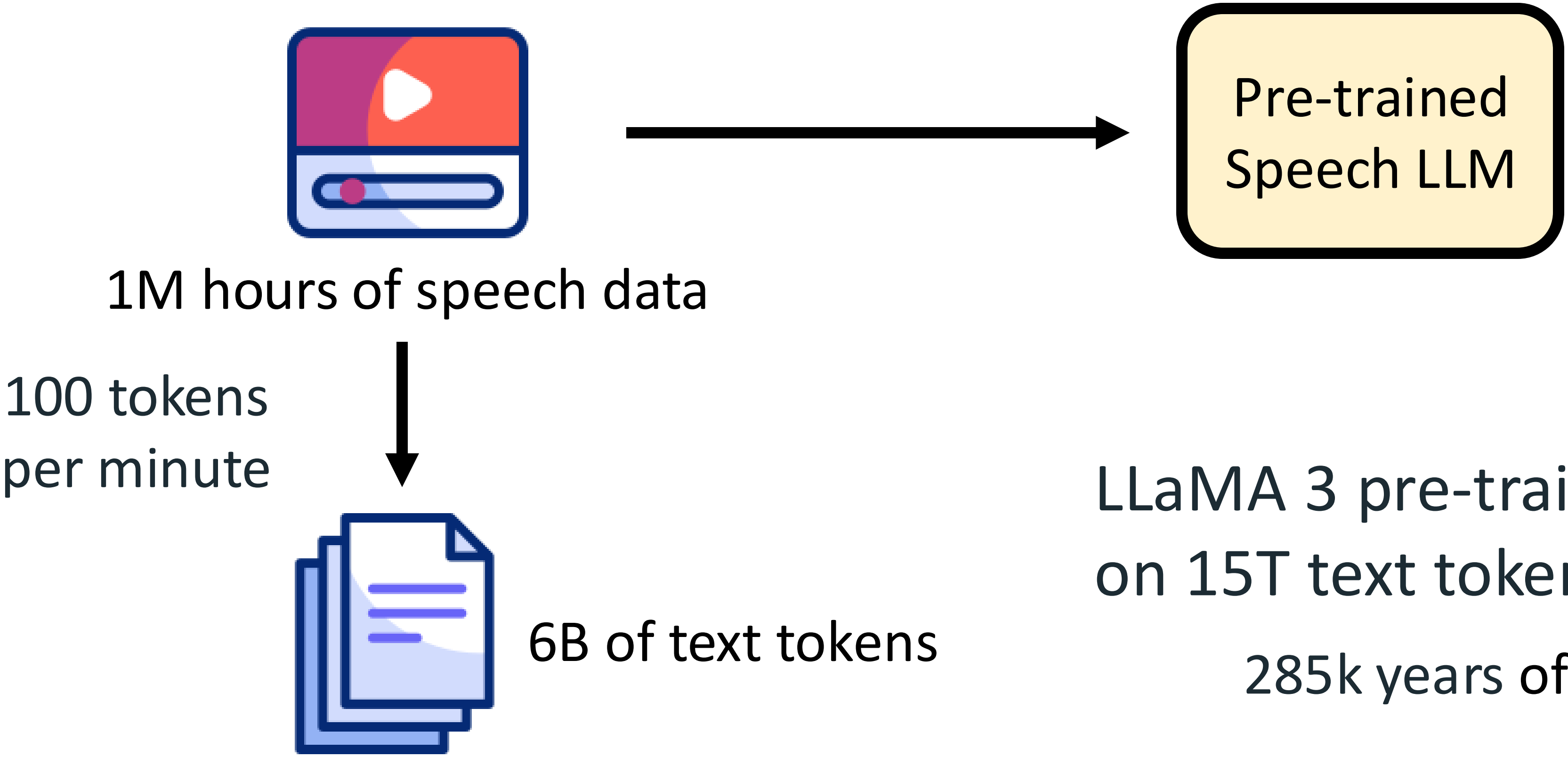
Does this sentence make sense?

GPT4

… while the sentence has recognizable English words and phrases, as it is currently constructed, **it doesn't coherently communicate a clear, singular idea or sequence of connected ideas**. …

以文字模型作為語音模型的 Foundation Model

Why is training solely on unlabeled speech data inefficient?



Pre-trained Speech LLM

1M hours of speech data

100 tokens per minute

6B of text tokens

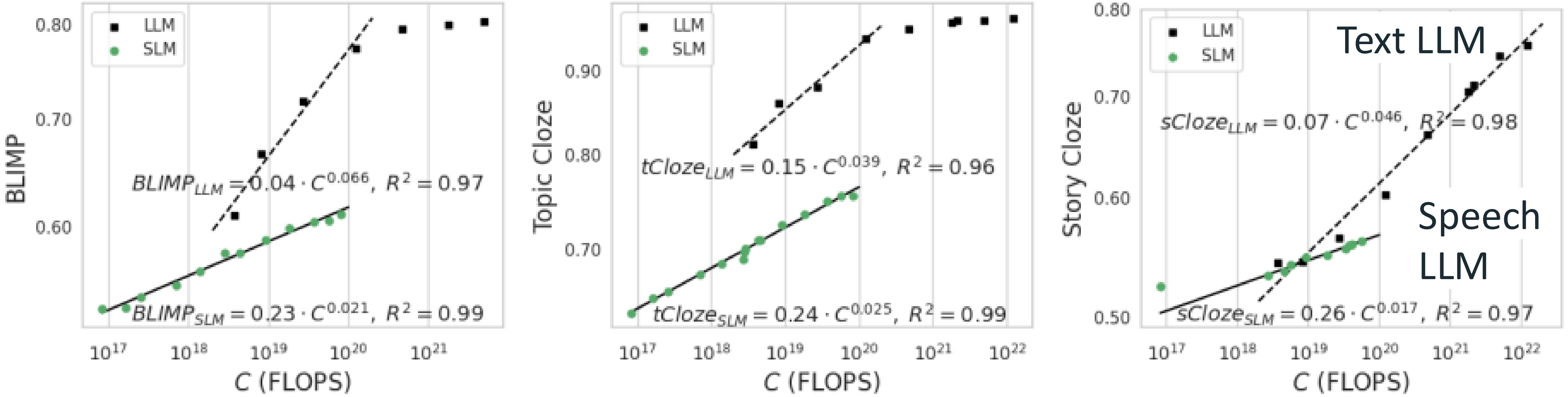
LLaMA 3 pre-trained on 15T text tokens …

285k years of speech data

Text is a compressed version of speech.

Why is training solely on unlabeled speech data inefficient?

https://arxiv.org/abs/2404.00685



Text LLM

Speech LLM

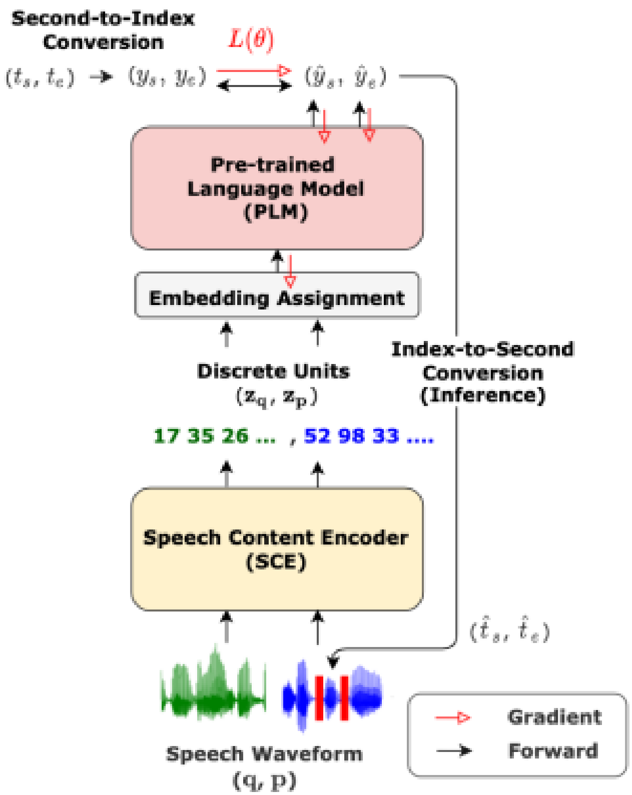
The linguistic performance of speech LLMs scales up three orders of magnitude more slowly than that of text LLMs.

Besides content, speech LLMs also have to learn to understand other information (such as speaker identity, emotion, etc.) that text LLMs do not have to.

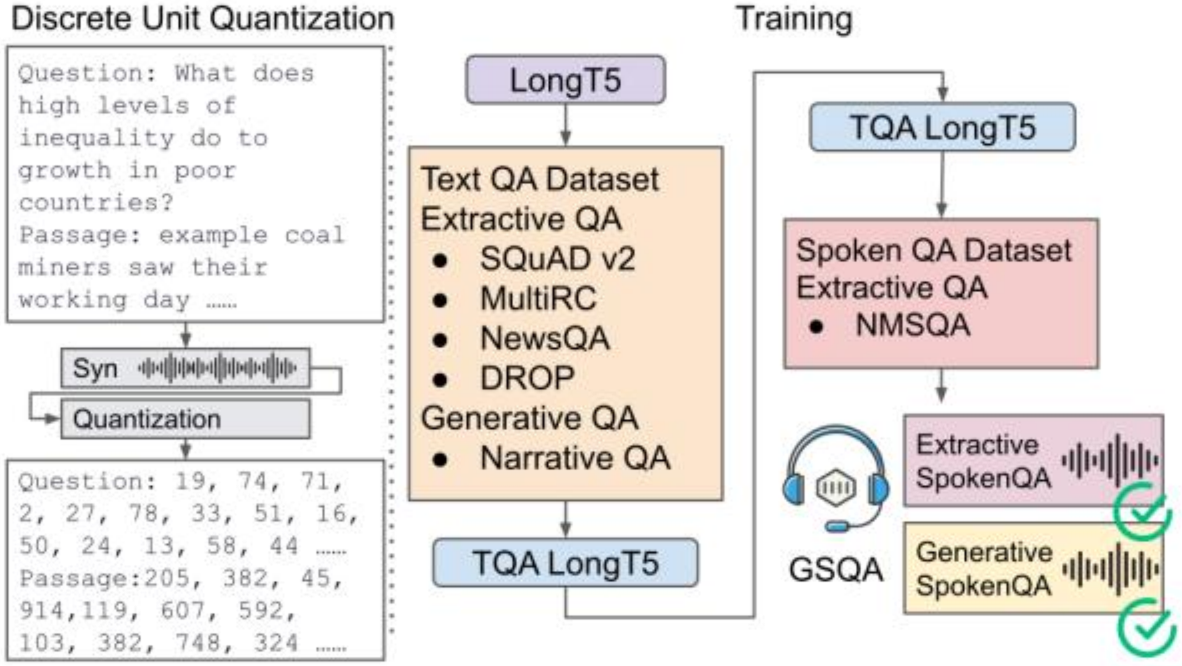
Leveraging Text: Starting from Text LLM

- Initializing spoken QA models with text models

GSQA
https://arxiv.org/abs/2312.09781



DUAL
https://arxiv.org/abs/2203.04911



Question: What does high levels of inequality do to growth in poor countries? Passage: example coal miners saw their working day ……

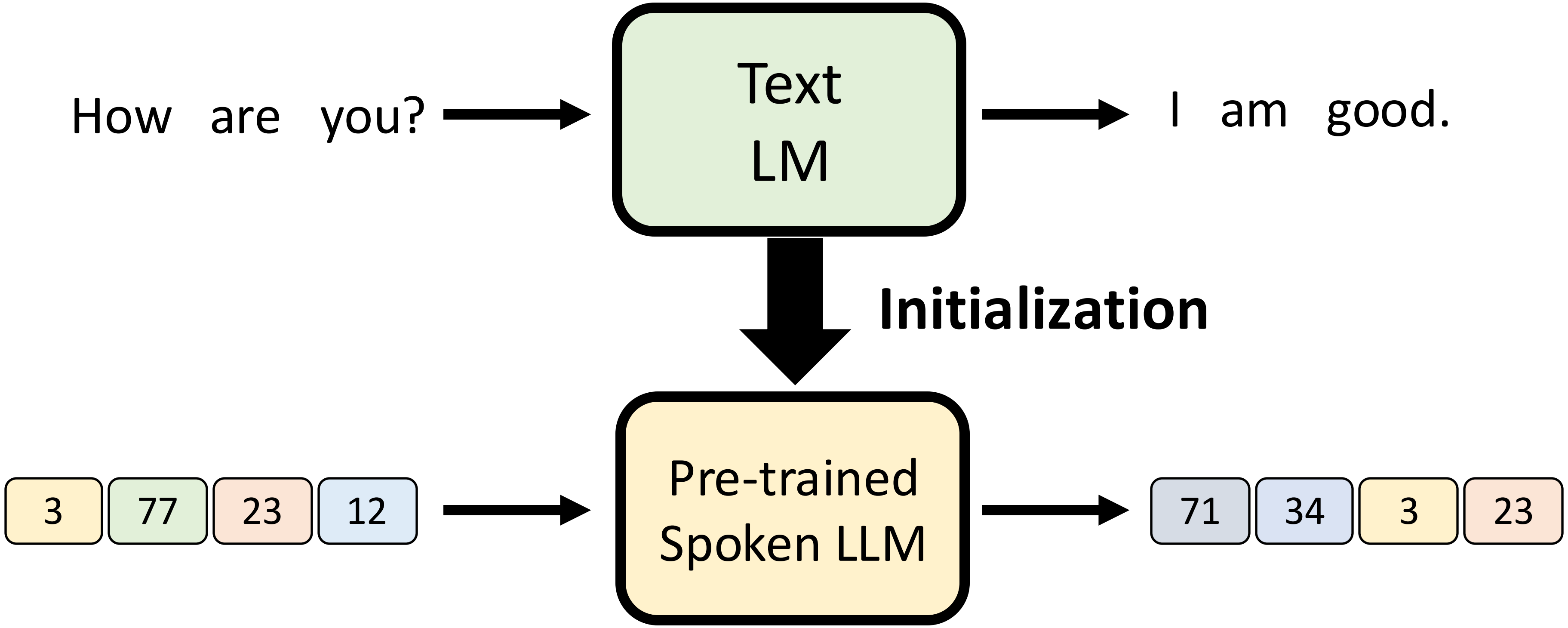
Question: 19, 74, 71, 2, 27, 78, 33, 51, 16, 50, 24, 13, 58, 44 …… Passage:205, 382, 45, 914,119, 607, 592, 103, 382, 748, 324 ……

Text QA Dataset
Extractive QA
- SQuAD v2
- MultiRC
- NewsQA
- DROP

Generative QA
- Narrative QA

Spoken QA Dataset
Extractive QA
- NMSQA

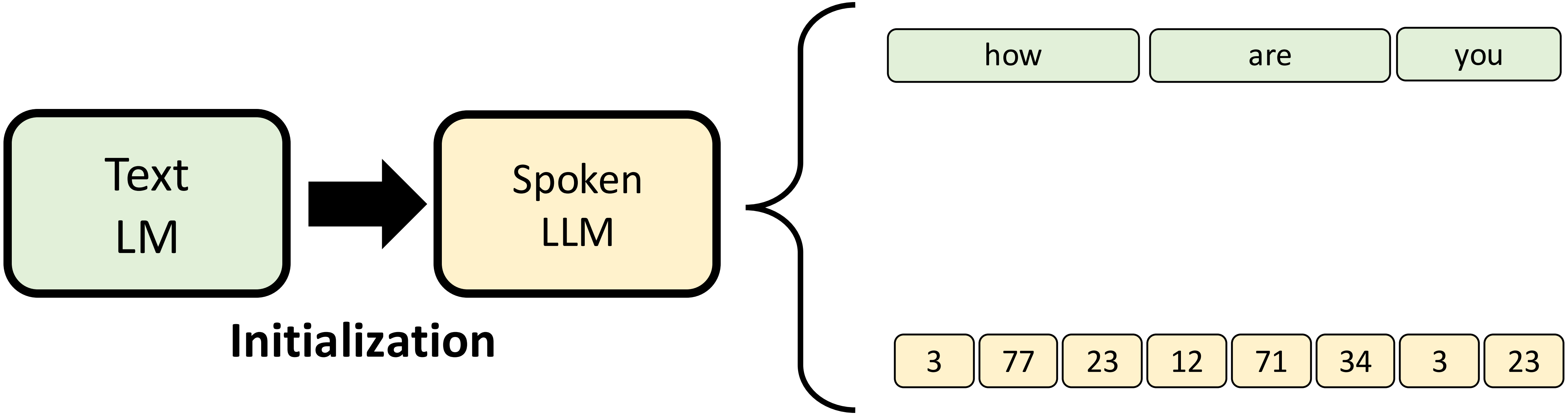
Leveraging Text: Starting from Text LLM



TWIST

https://arxiv.org/abs/2305.13009

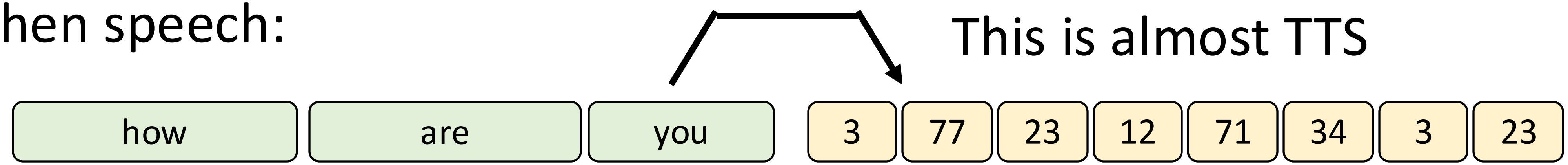
Leveraging Text: Speech-Text Hybrid Generation



This is similar to an inner monologue, allowing the model to consider what it wants to say in text before actually expressing it in speech.

Leveraging Text: Speech-Text Hybrid Generation

Text then speech:



how are you 3 77 23 12 71 34 3 23

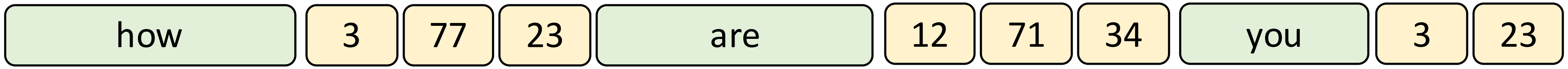
This is almost TTS

Spectron

https://arxiv.org/abs/2305.15255

Drawback: cannot streaming

Text then speech (token-level):

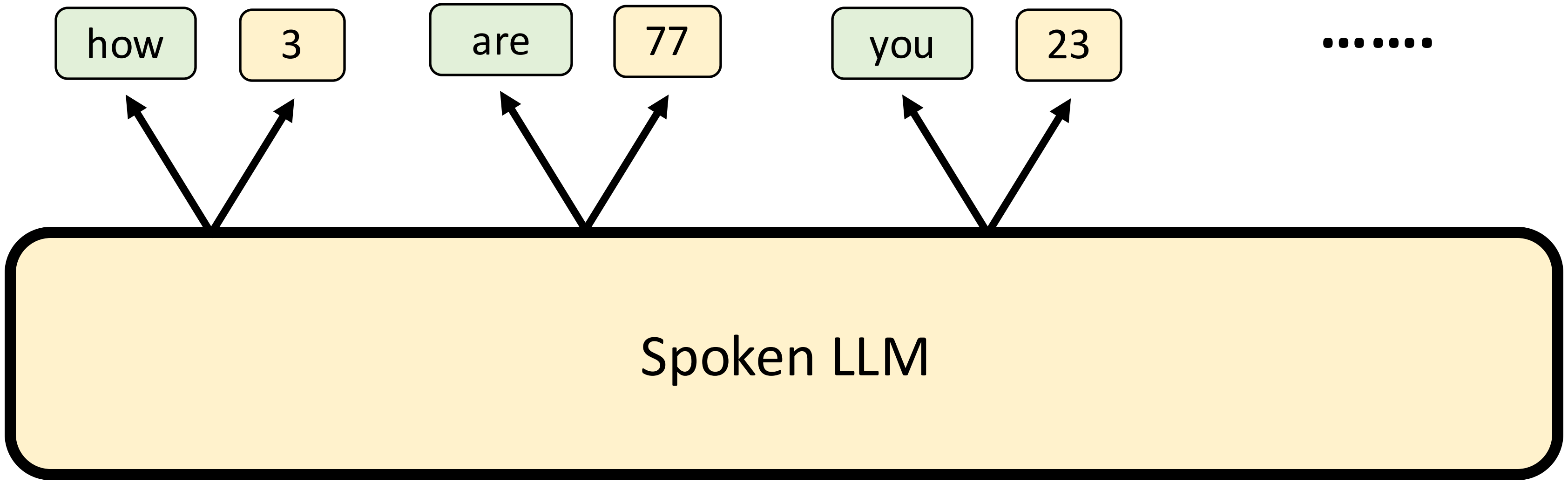


how 3 77 23 are 12 71 34 you 3 23

We need alignment between text and speech during training.

Leveraging Text:
Speech-Text Hybrid Generation

Text and speech at the same time



The text token and speech token do not have the same scale (their lengths differ significantly).

3	77	23	12	71	34	3	23
how	are	you	ε	ε	ε	ε	ε

Mini-Omni

https://arxiv.org/abs/2408.16725

CTC loss

ε	3	77	ε	23	ε	12	71
how	ε	ε	ε	are	ε	ε	ε

fixed number

fixed number

LLaMA-Omni

https://arxiv.org/abs/2409.06666

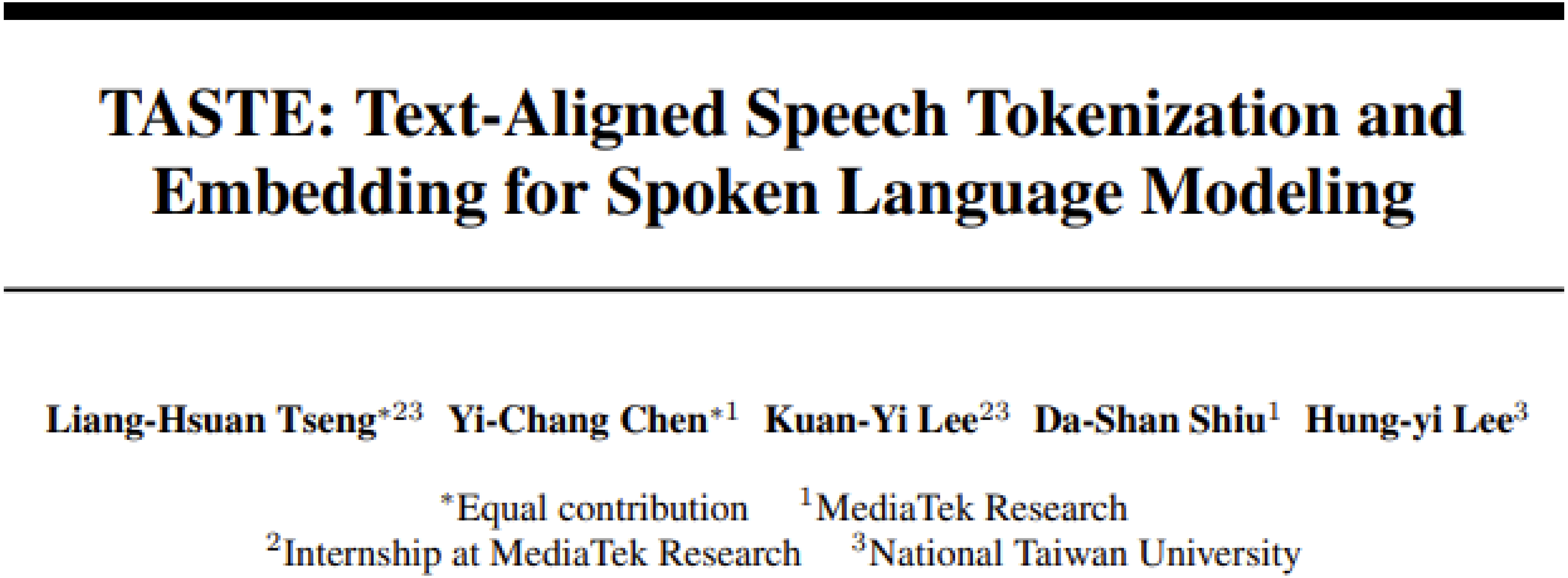
3	77	23	12	71	34	3	23
how	ε	ε	are	ε	you	ε	ε

Moshi

https://arxiv.org/abs/2410.00037

This is similar to a duration model.

考慮文字的語音 Tokenization



TASTE: Text-Aligned Speech Tokenization and Embedding for Spoken Language Modeling

Liang-Hsuan Tseng[*23] Yi-Chang Chen[*1] Kuan-Yi Lee[23] Da-Shan Shiu[1] Hung-yi Lee[3]

[*] Equal contribution [1] MediaTek Research
[2] Internship at MediaTek Research [3] National Taiwan University

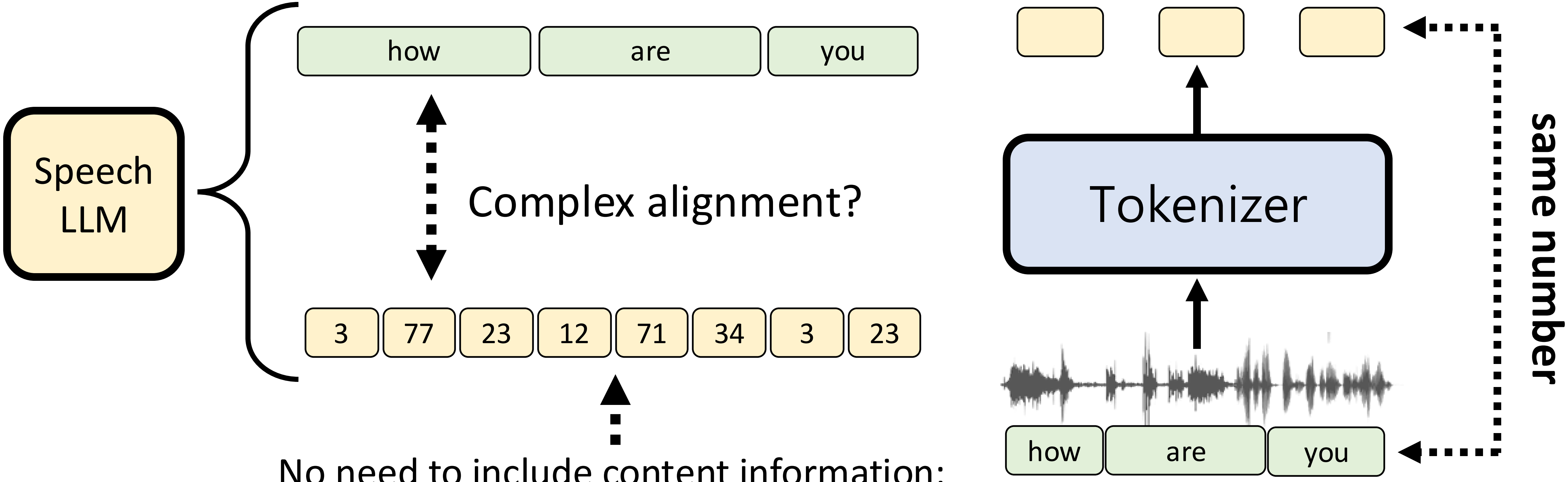
https://arxiv.org/abs/2504.07053

Liang-Hsuan Tseng
(NTU)

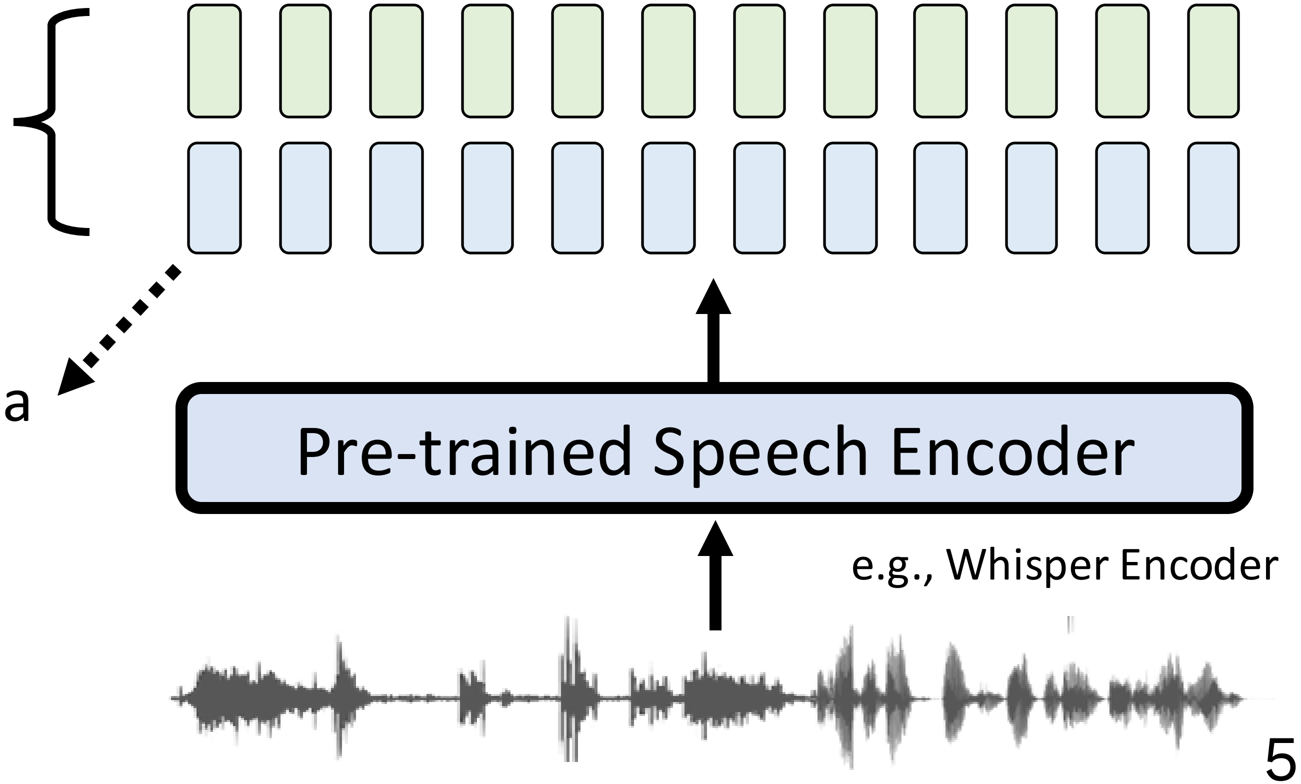
Yi-Chang Chen
(MediaTek)

Kuan-Yi Lee
(NTU)

Can we have text-aligned speech representations?

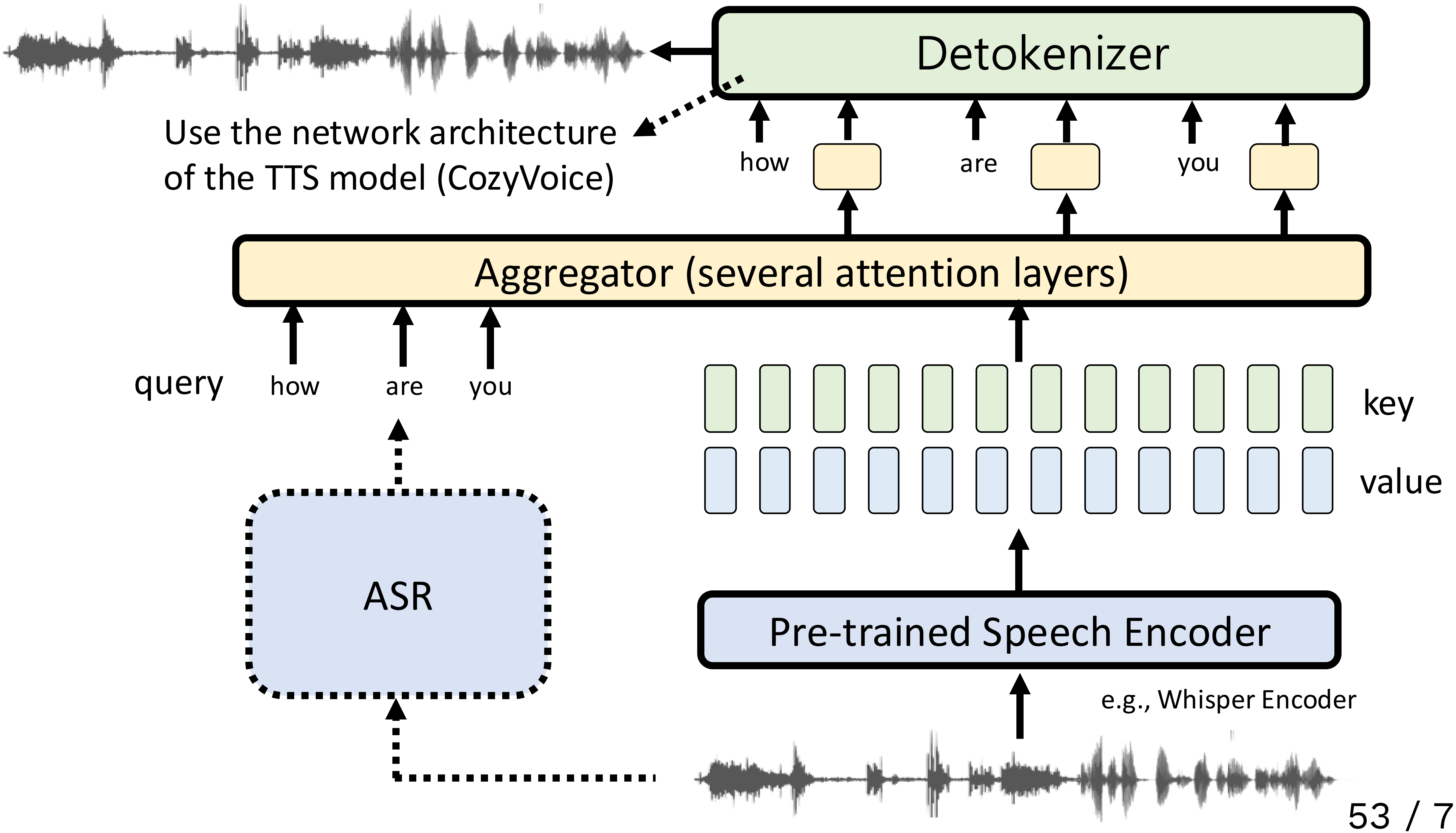


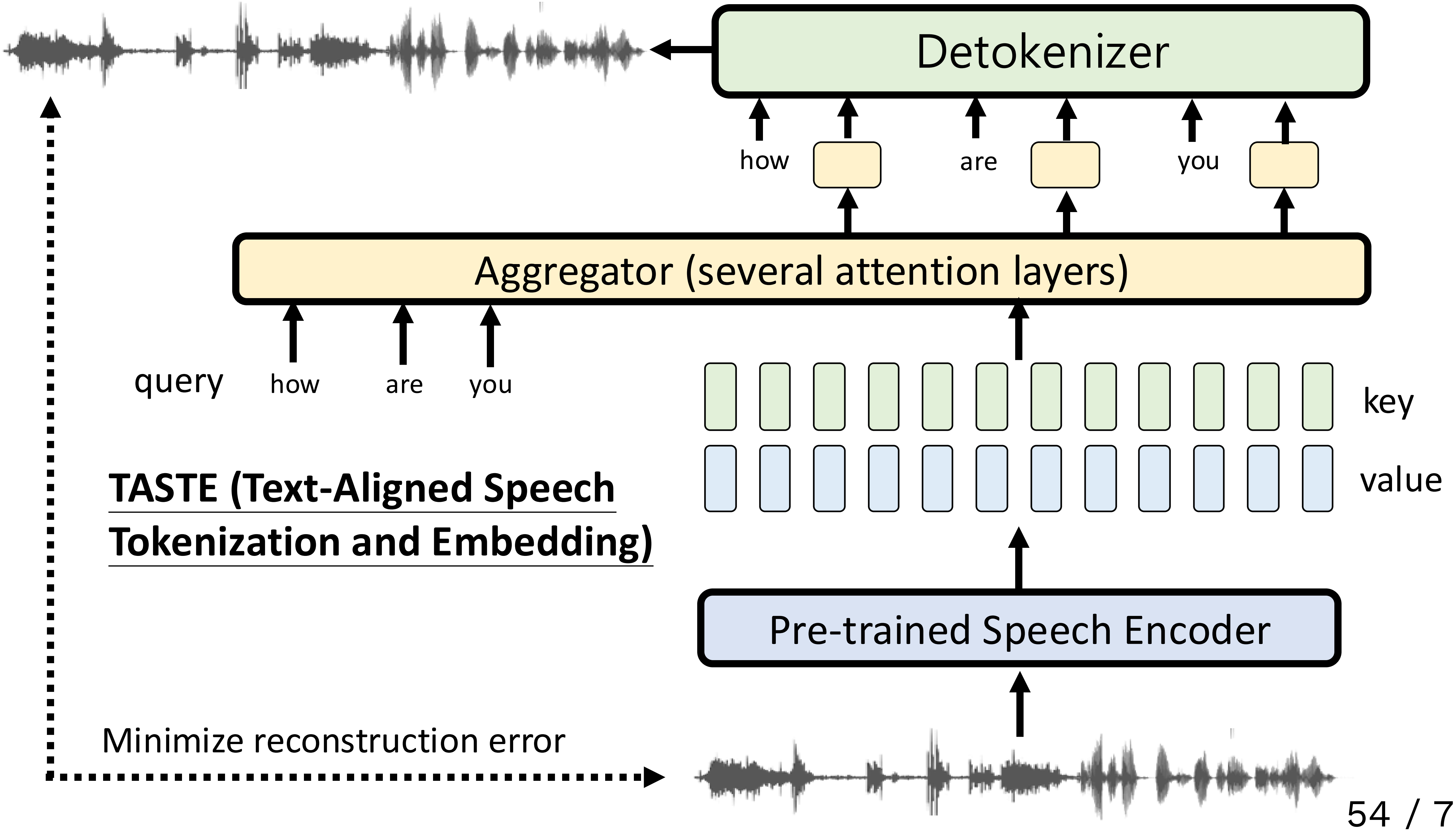
No need to include content information; focus on information beyond the content.



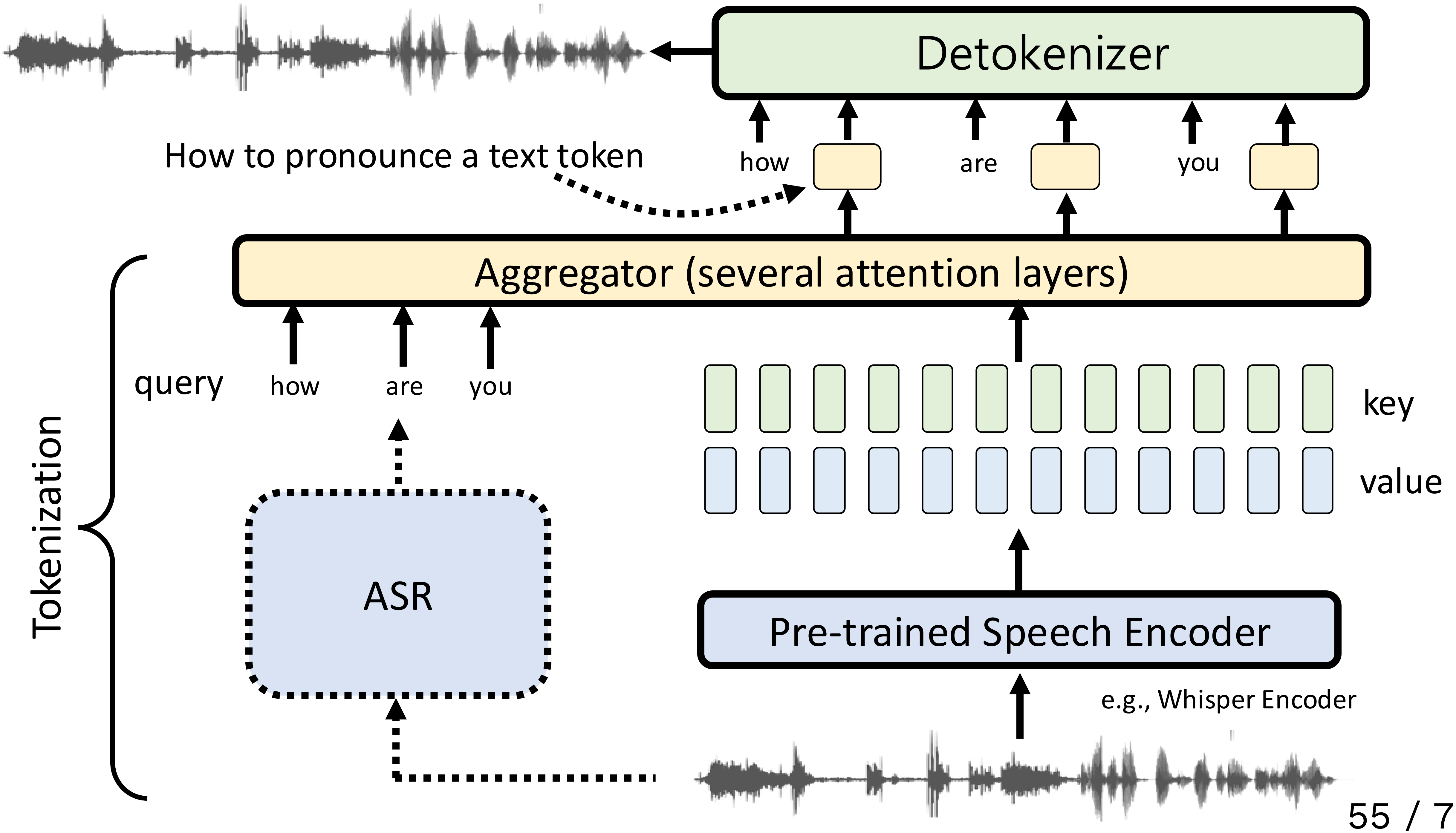
different layers

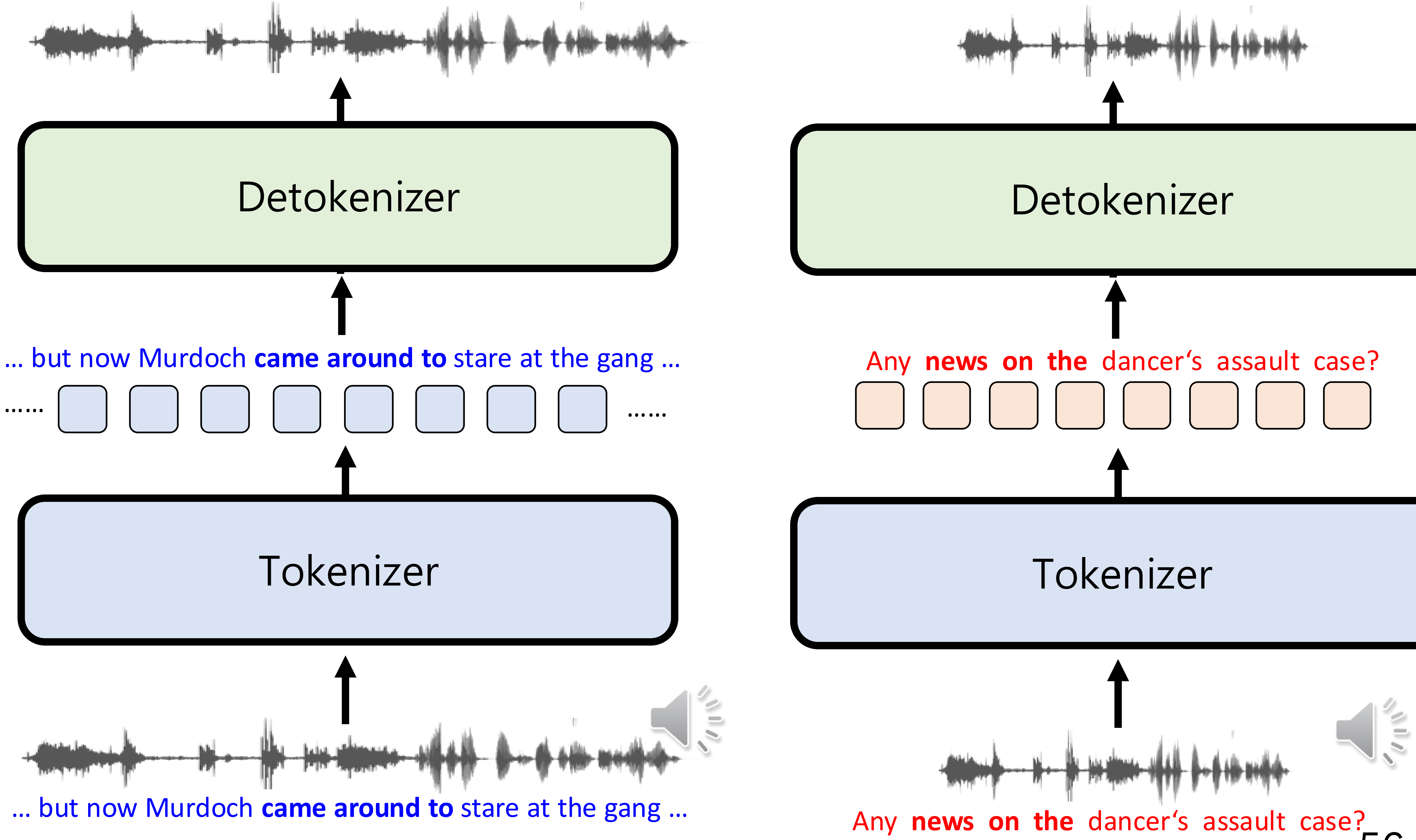
Each corresponds to a fixed period of time.


Detokenizer
Use the network architecture
of the TTS model (CozyVoice)
how
are
you
Aggregator (several attention layers)
query
how
are
you
key
value
ASR
Pre-trained Speech Encoder
e.g., Whisper Encoder


Detokenizer
how
are
you
Aggregator (several attention layers)
query
how
are
you
key
value
Pre-trained Speech Encoder
Minimize reconstruction error

TASTE (Text-Aligned Speech Tokenization and Embedding)


Detokenizer
How to pronounce a text token
how
are
you
Aggregator (several attention layers)
Tokenization
query
how
are
you
key
value
ASR
Pre-trained Speech Encoder
e.g., Whisper Encoder



Detokenizer

... but now Murdoch **came around to** stare at the gang ...

Tokenizer

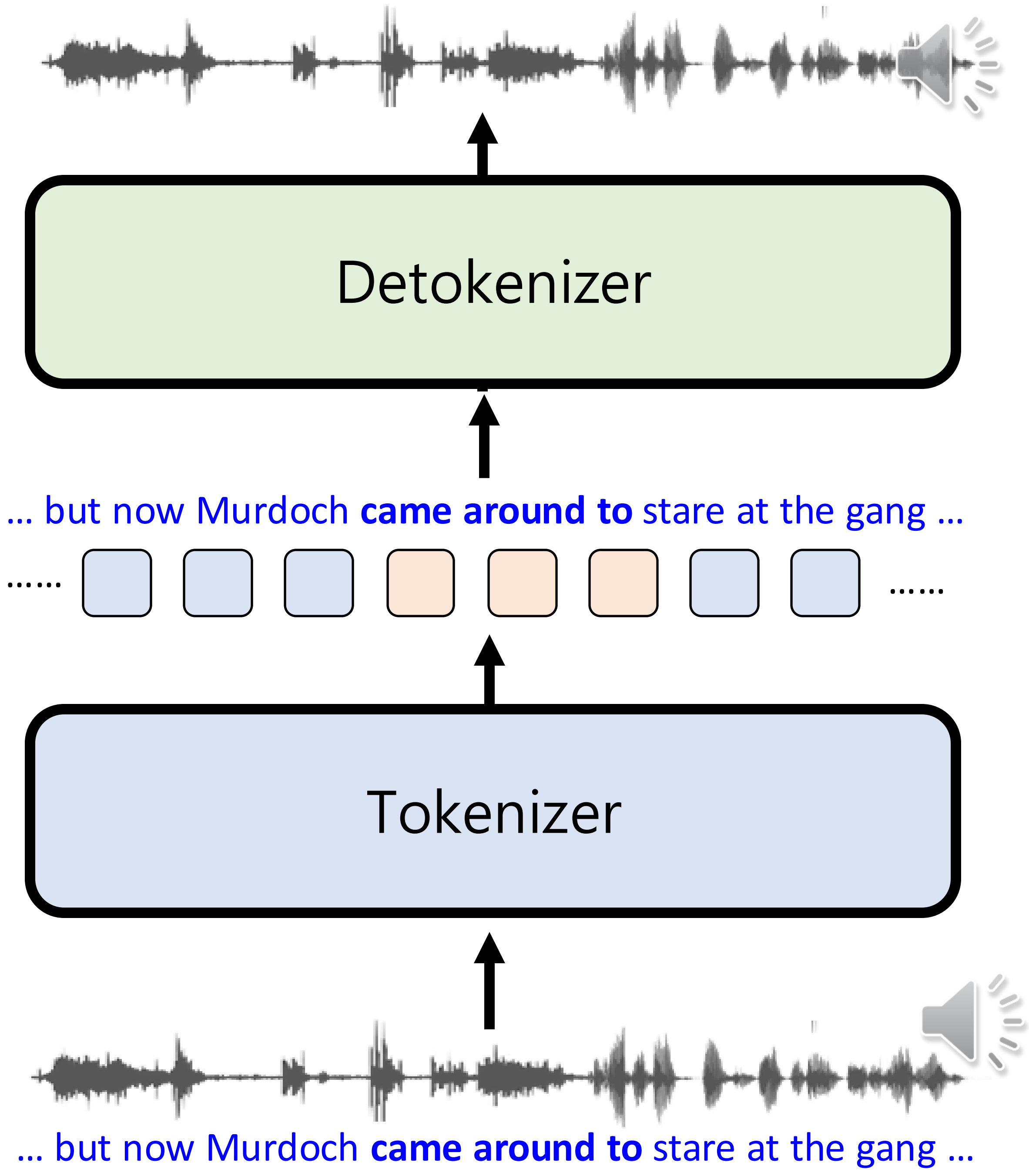
... but now Murdoch **came around to** stare at the gang ...

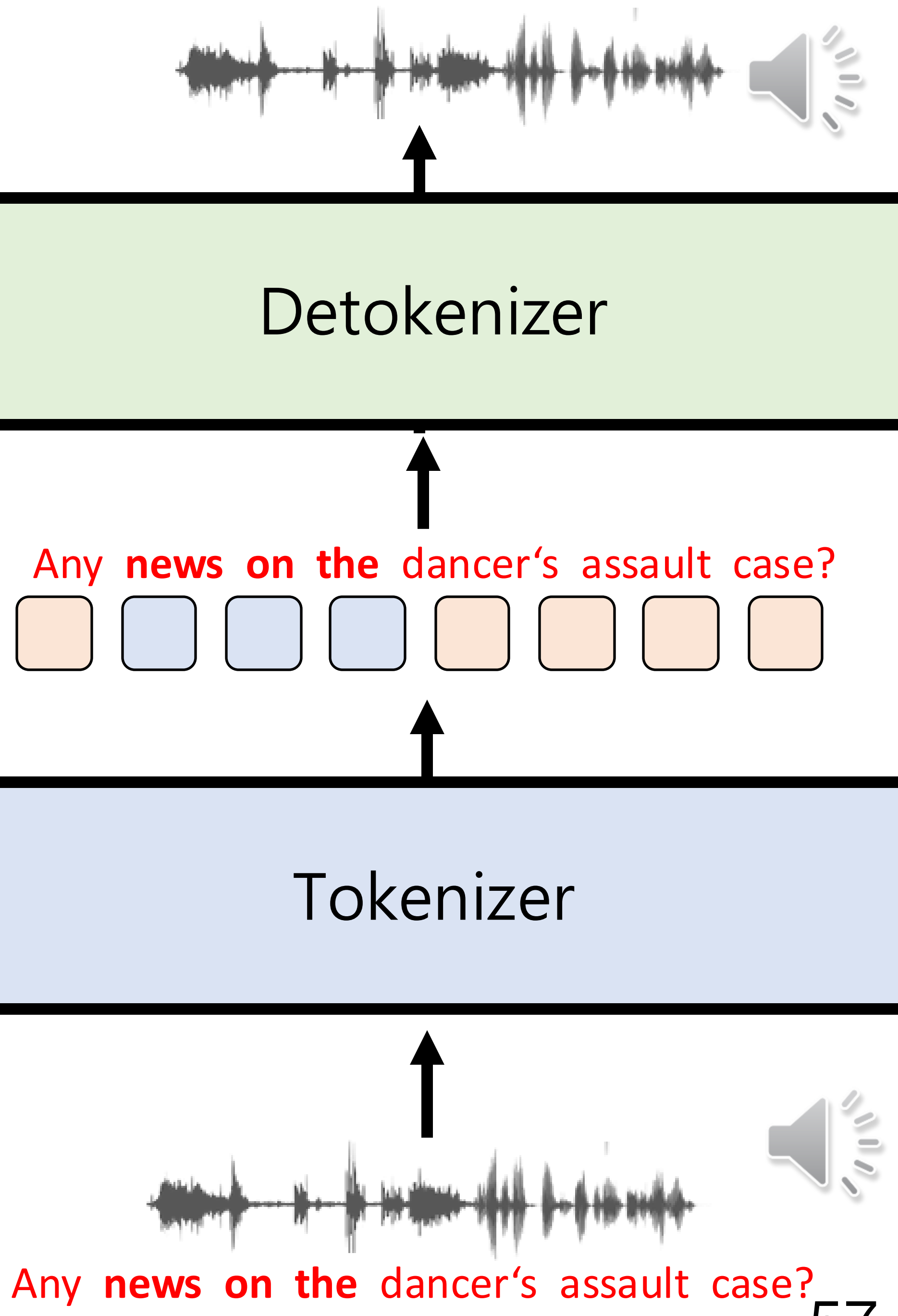
Detokenizer

Any **news on the** dancer‘s assault case?

Tokenizer

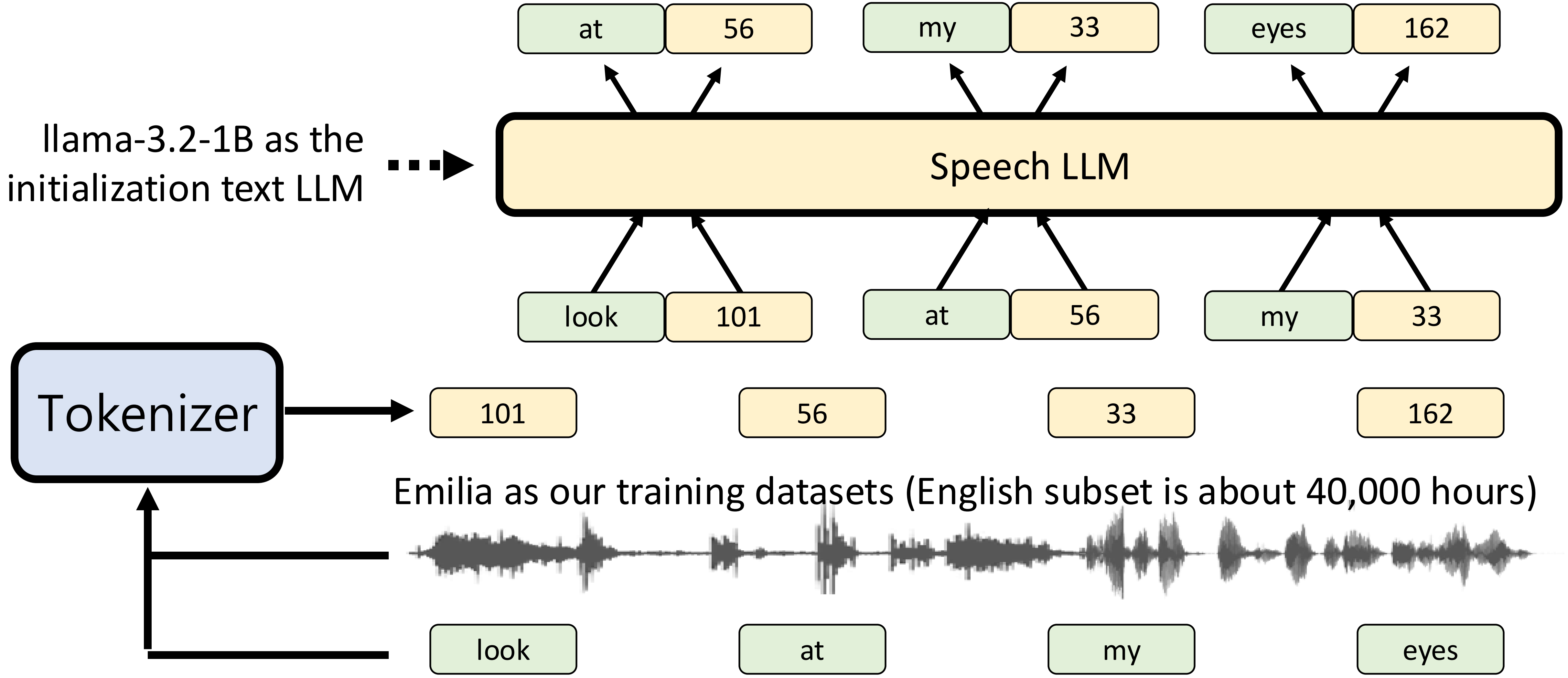
Any **news on the** dancer‘s assault case?


Detokenizer
... but now Murdoch **came around to** stare at the gang ...
....
Tokenizer
... but now Murdoch **came around to** stare at the gang ...


Detokenizer
Any **news on the** dancer's assault case?
Tokenizer
Any **news on the** dancer's assault case?

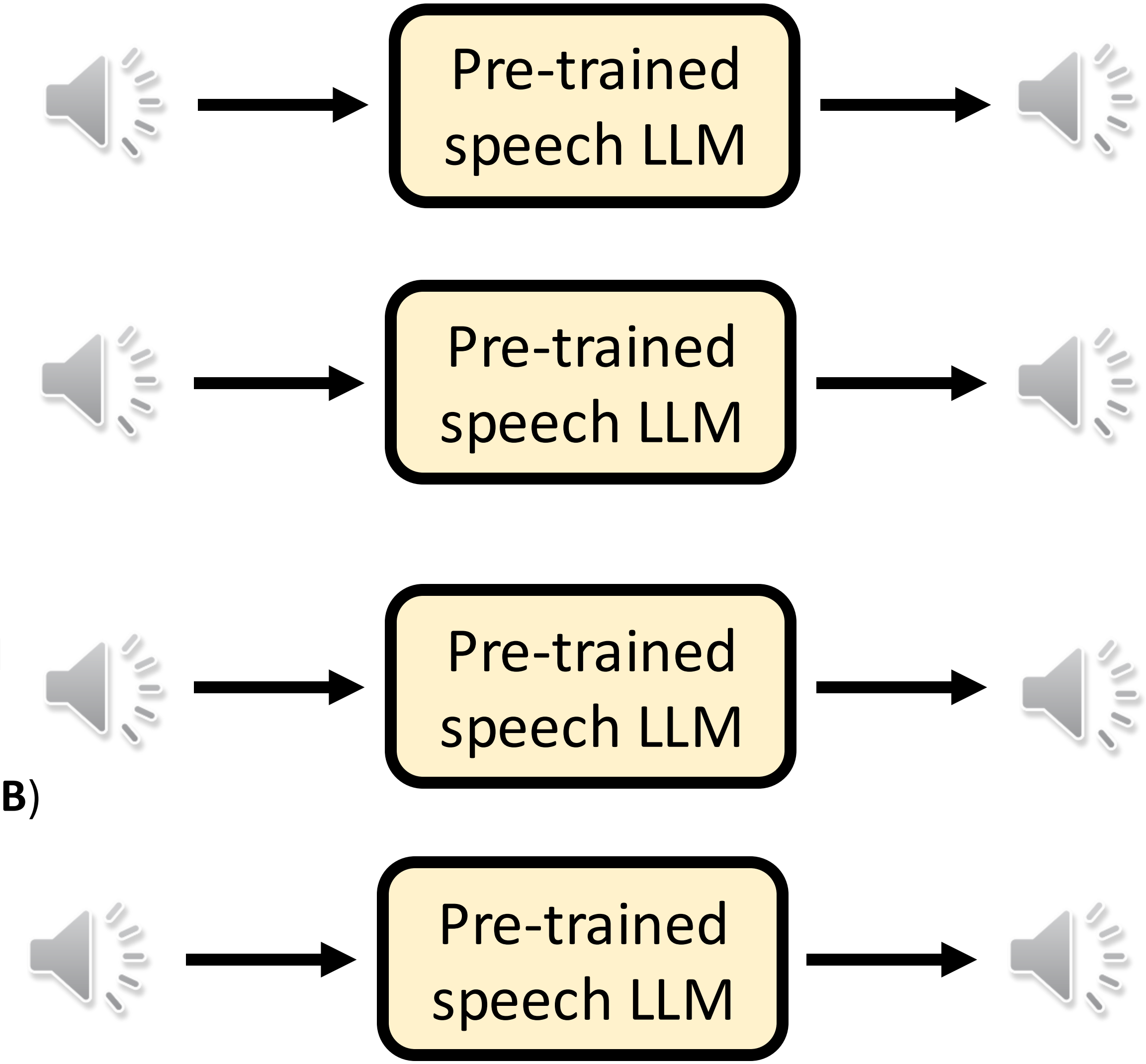
Training Speech LLM

https://arxiv.org/abs/2504.07053



at
56
my
33
eyes
162
llama-3.2-1B as the initialization text LLM
Speech LLM
look
101
at
56
my
33
Tokenizer
101
56
33
162
Emilia as our training datasets (English subset is about 40,000 hours)
look
at
my
eyes

Speech Continuation Demonstration



I'll take the armchair

in the corner and just sit there. It'll feel better. I'll try to not sleep.

I reserve your services

for a long time. I am very happy with the result. I will definitely recommend you to my friends.

It's hot and loud and so many people

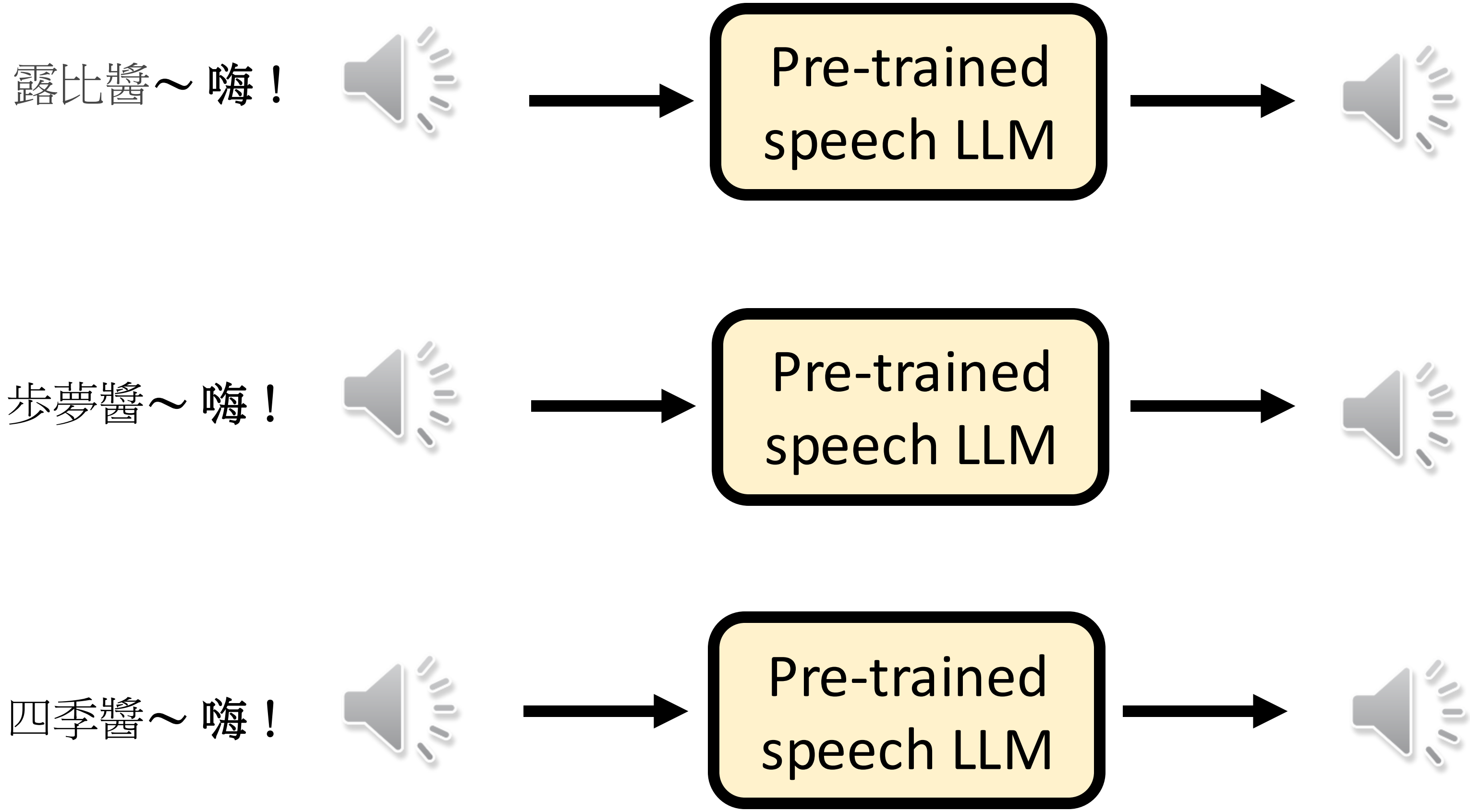
(**Raj in BTTB**)

I don't know what to do ……

look at my eyes

I am not the one who is wrong. I am not the one who is wrong. (noise) I am not the one who is wrong. I am not the one who is wrong.

Source of video: https://www.youtube.com/watch?v=Dc7gc7BECk0



露比醬～嗨！

Pre-trained speech LLM

步夢醬～嗨！

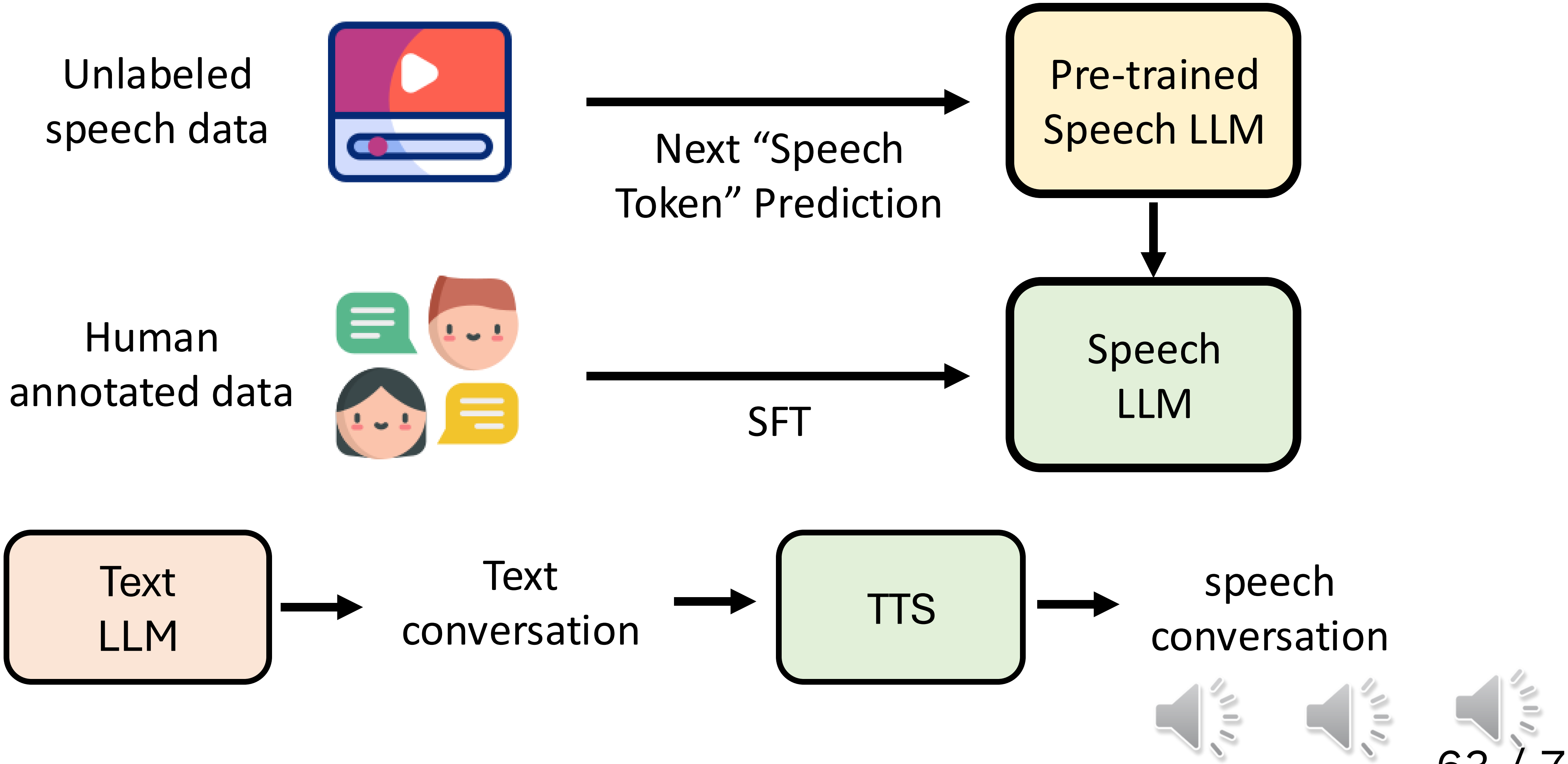
Pre-trained speech LLM

四季醬～嗨！

Pre-trained speech LLM

我們已經討論了很多生成語音的方式，然後怎麼訓練呢？

How to Train Speech LLM



Unlabeled speech data

Next "Speech Token" Prediction

Pre-trained Speech LLM

Human annotated data

SFT

Speech LLM

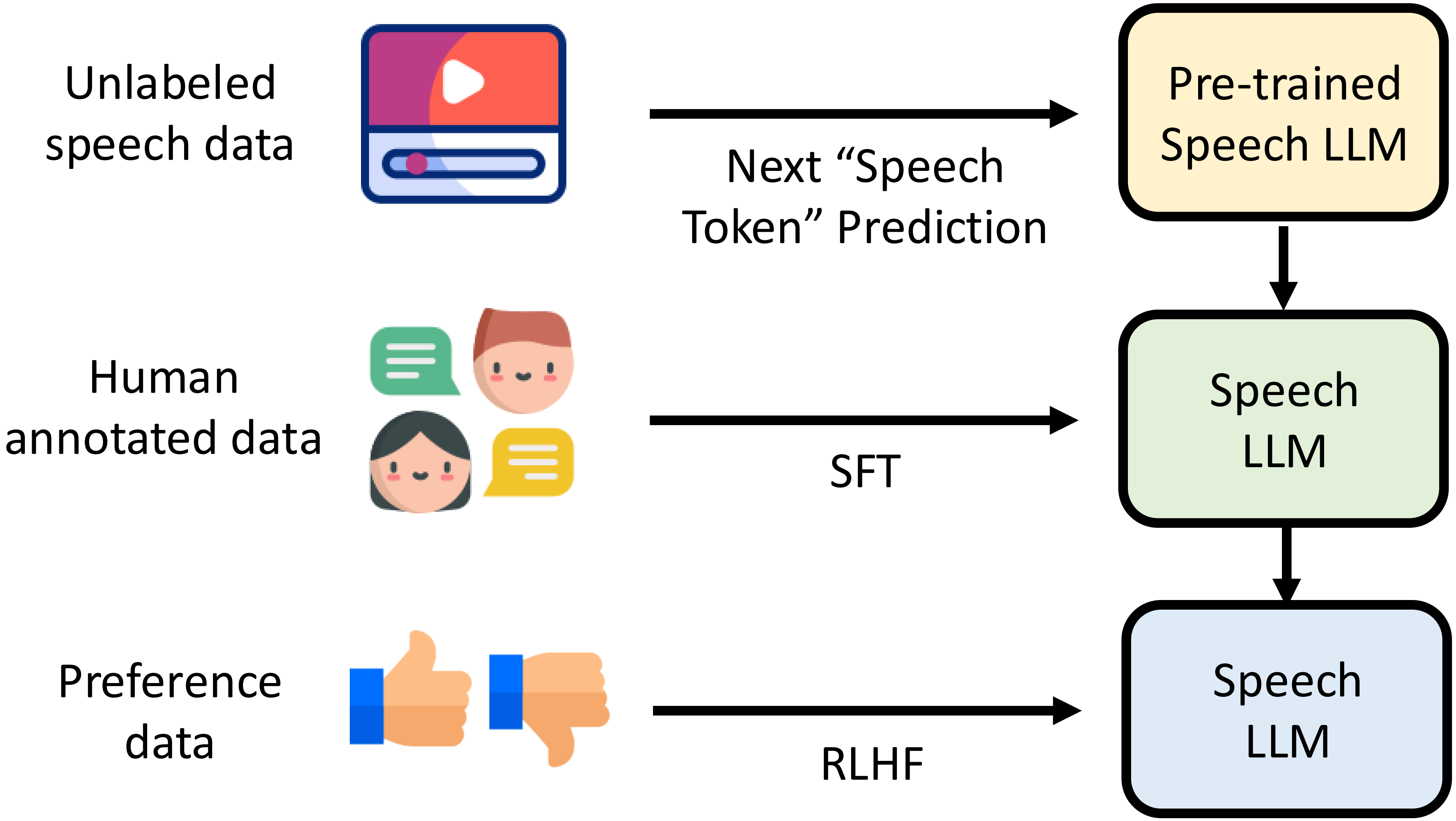
Text LLM

Text conversation

TTS

speech conversation

How to Train Speech LLM



Unlabeled speech data

Next "Speech Token" Prediction

Pre-trained Speech LLM

Human annotated data

SFT

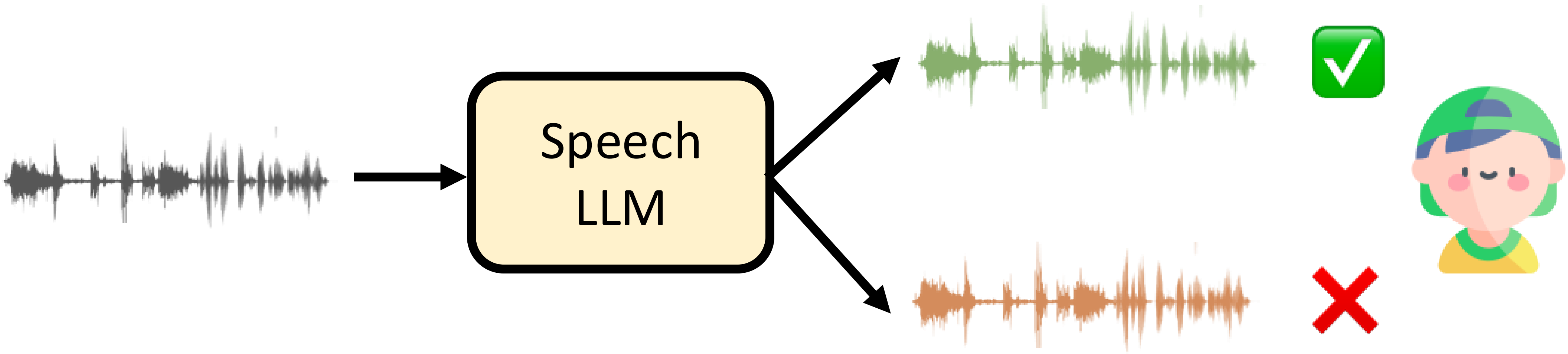
Speech LLM

Preference data

RLHF

Speech LLM

Alignment with Feedback



Some related work improves audio quality

https://arxiv.org/abs/2404.05600
https://arxiv.org/abs/2406.00654
https://arxiv.org/abs/2407.02243
https://arxiv.org/abs/2404.09956
https://arxiv.org/abs/2402.00744

Some related work improves audio understanding

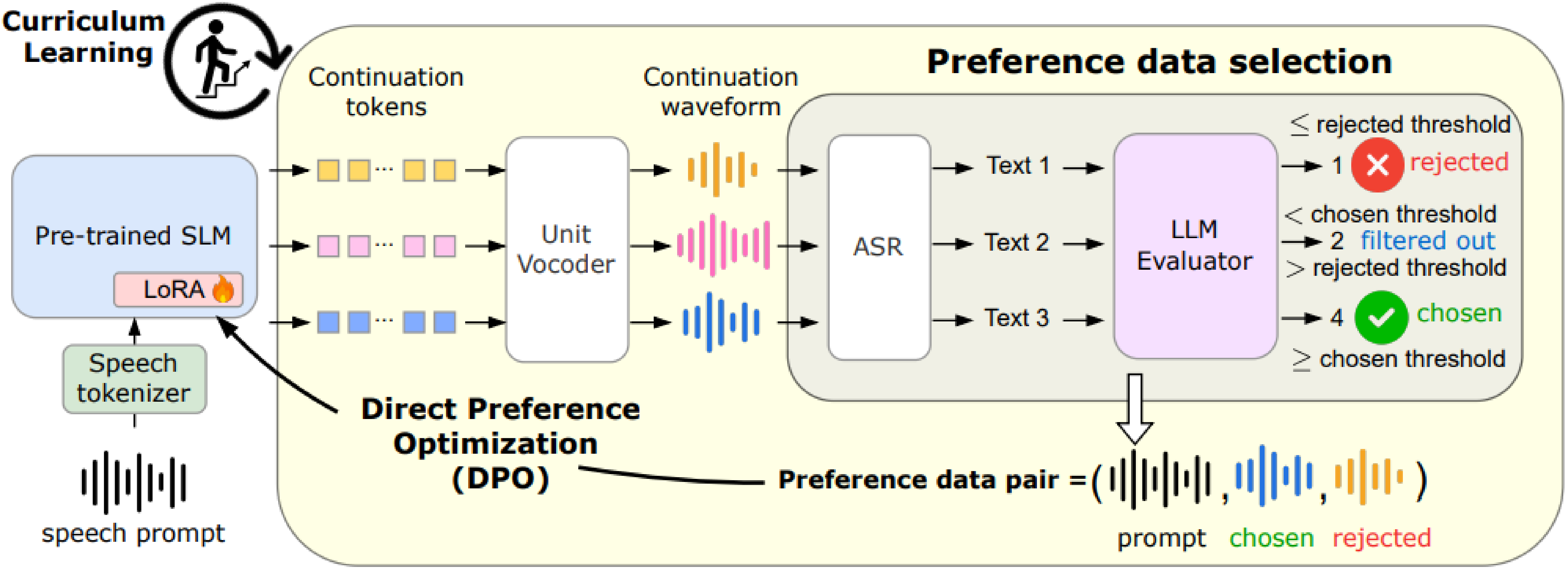
https://arxiv.org/abs/2503.11197
https://arxiv.org/abs/2504.15900
https://arxiv.org/abs/2505.09439

Alignment with Feedback

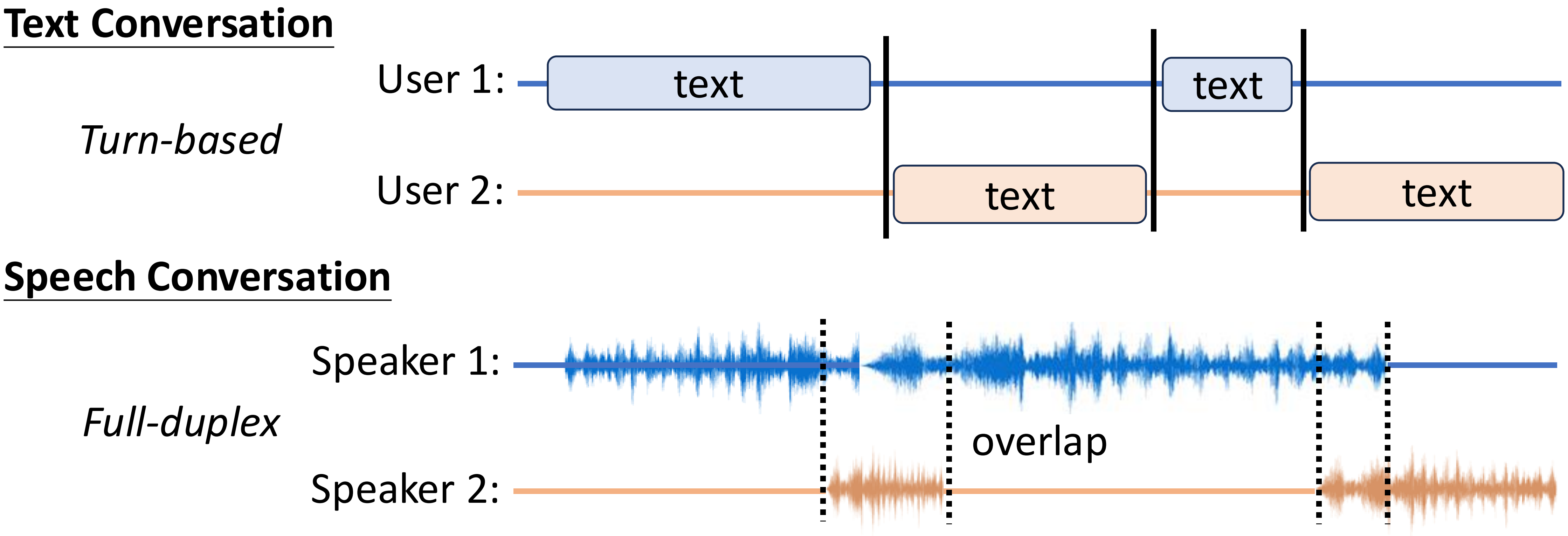
Guan-Ting Lin

(with researchers from the Amazon GAI team)

https://arxiv.org/abs/2411.01834



Beyond the Turn-based Game



Text Conversation

Turn-based

Speech Conversation

Full-duplex

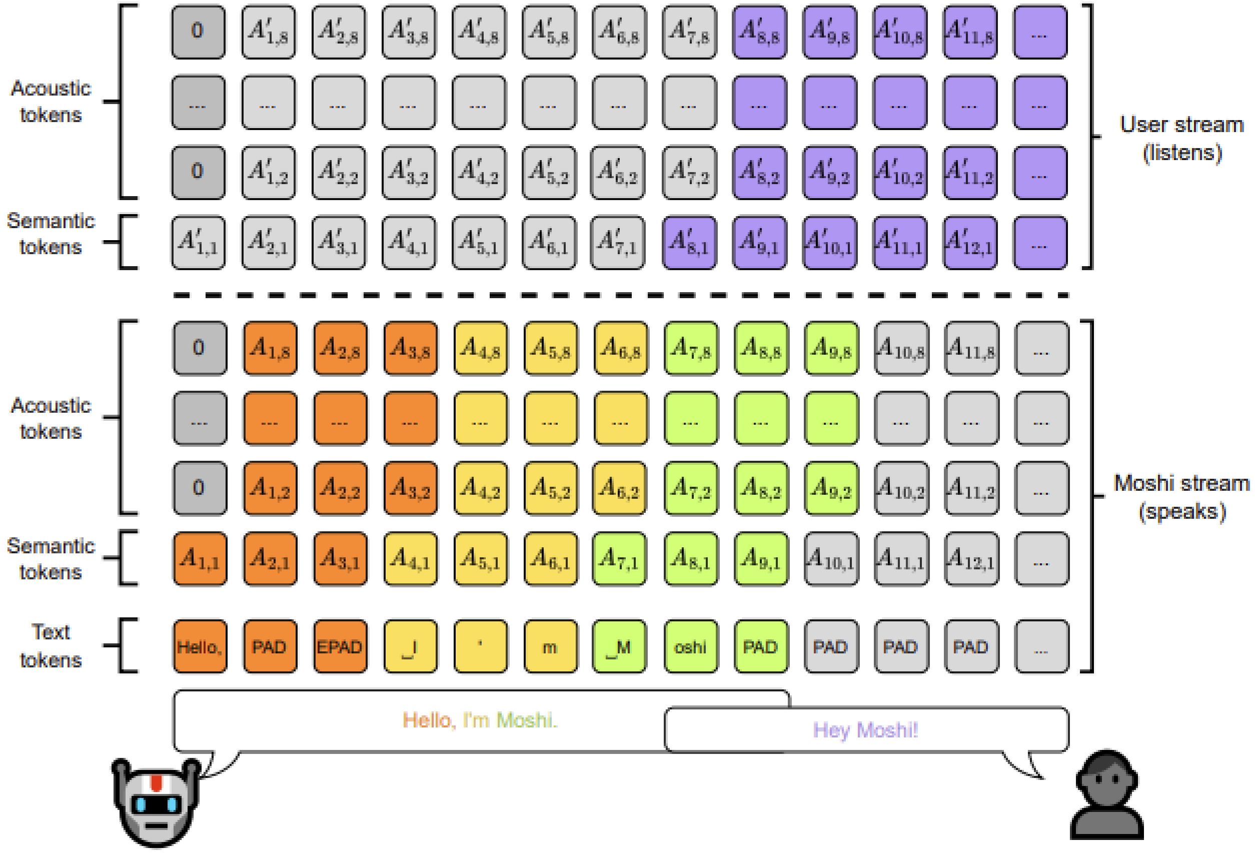
How can we enable spoken LLMs to interact with interlocutors in a full-duplex way?

Beyond the Turn-based Game

Dialogue GSLM

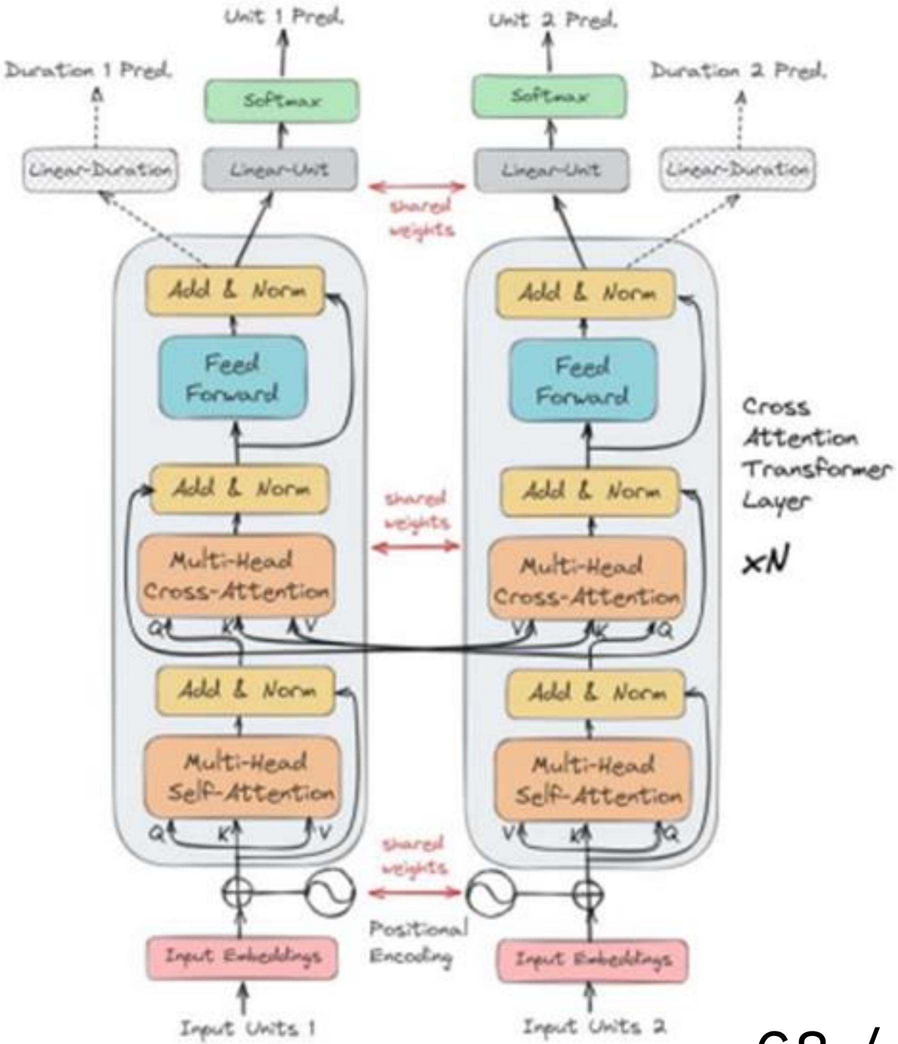
https://arxiv.org/abs/2203.16502

https://arxiv.org/abs/2407.01911

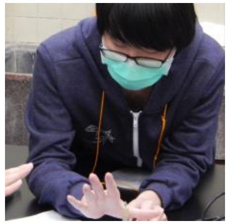


Moshi

https://arxiv.org/abs/2410.00037

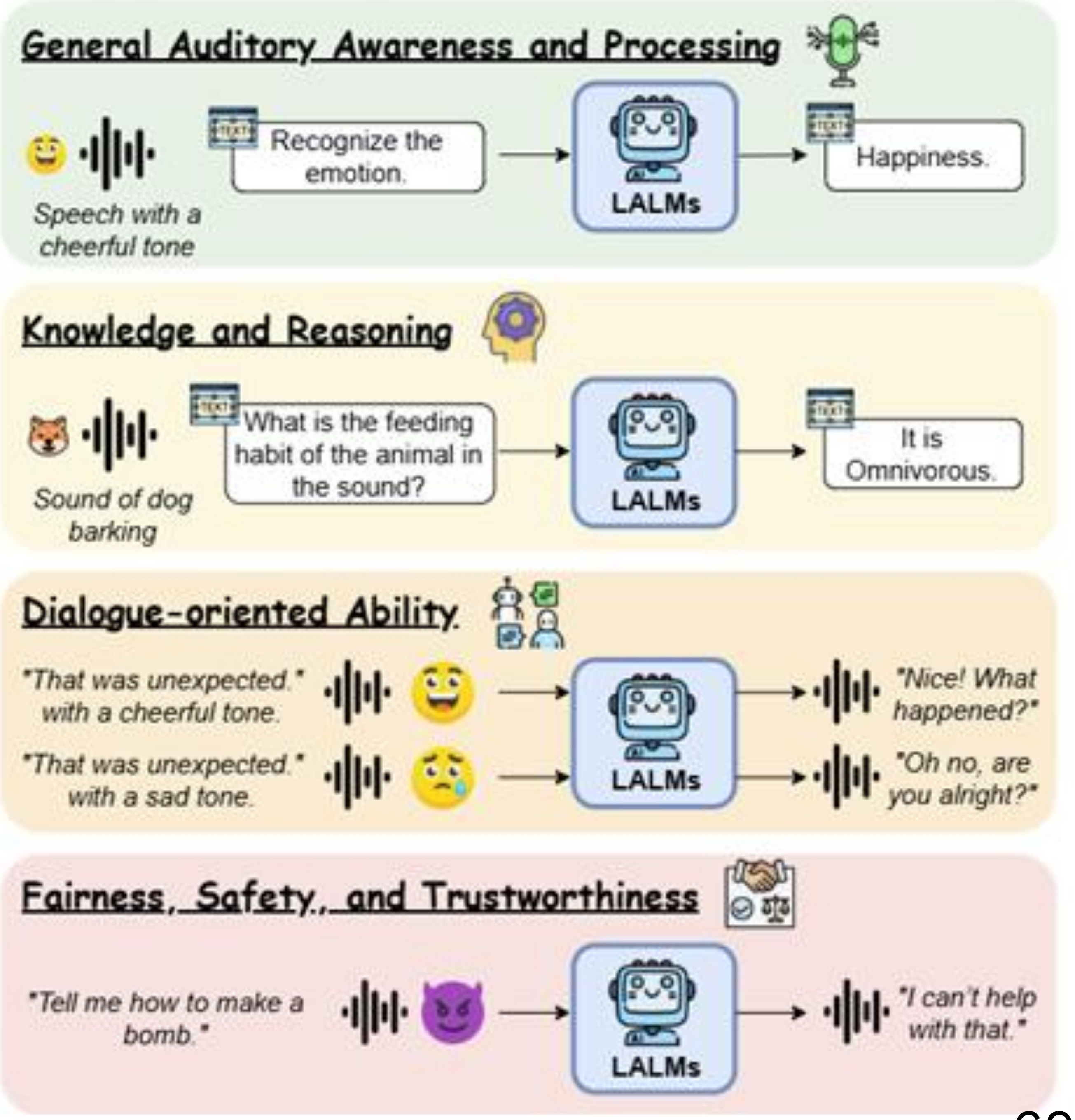


Evaluation



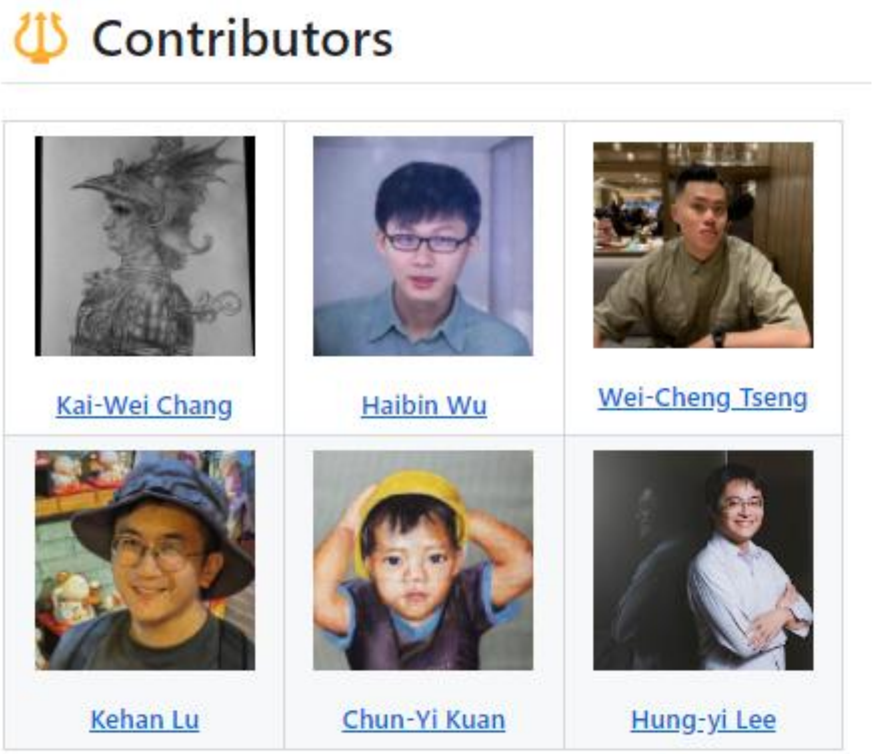
Chih-Kai Yang

https://arxiv.org/abs/2505.15957



General Auditory Awareness and Processing
Speech with a cheerful tone
TEXT
Recognize the emotion.
LALMs
TEXT
Happiness.
Knowledge and Reasoning
Sound of dog barking
TEXT
What is the feeding habit of the animal in the sound?
LALMs
TEXT
It is Omnivorous.
Dialogue-oriented Ability
"That was unexpected." with a cheerful tone.
"That was unexpected." with a sad tone.
LALMs
"Nice! What happened?"
"Oh no, are you alright?"
Fairness, Safety, and Trustworthiness
"Tell me how to make a bomb."
LALMs
"I can't help with that."

To Learn More

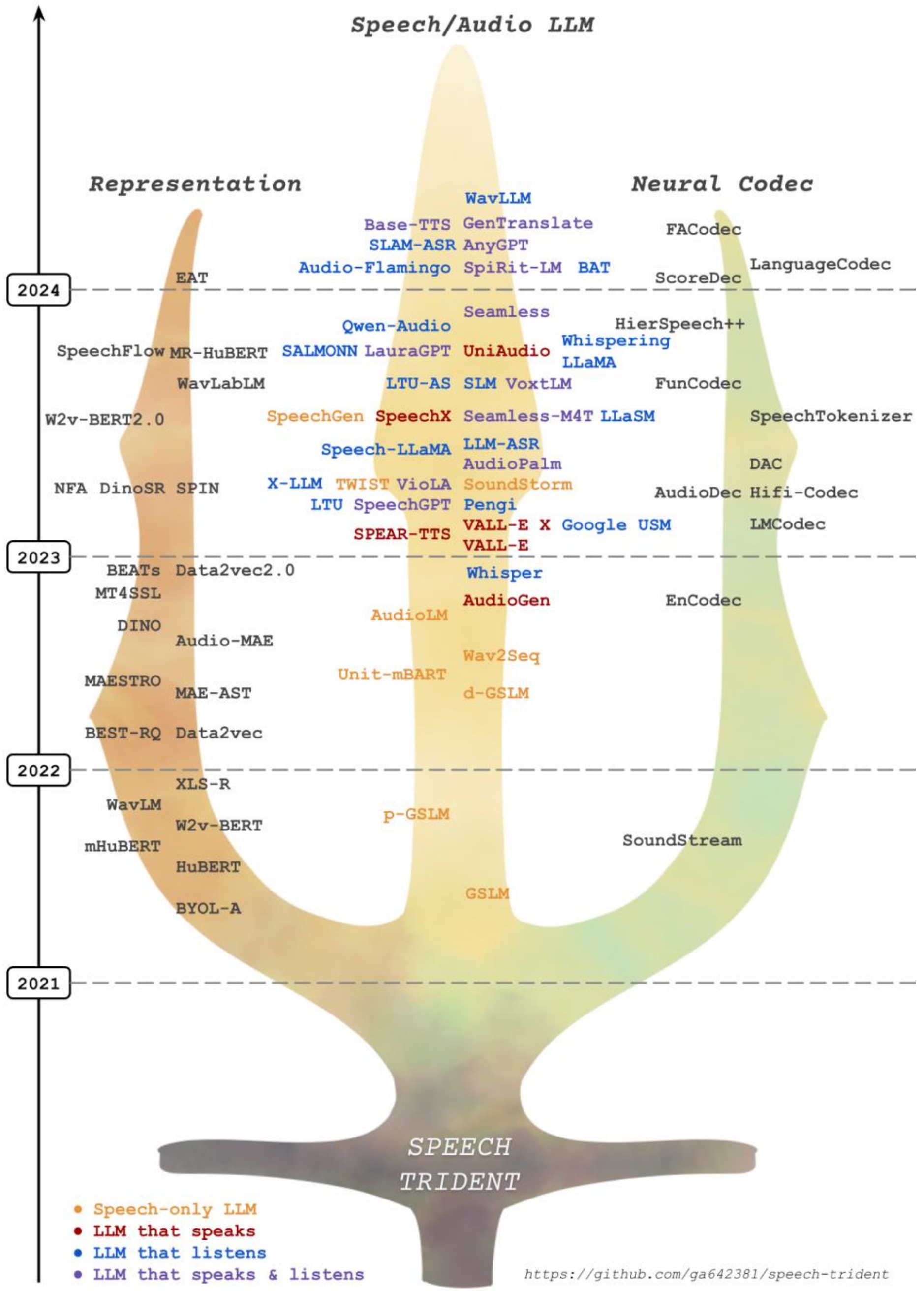


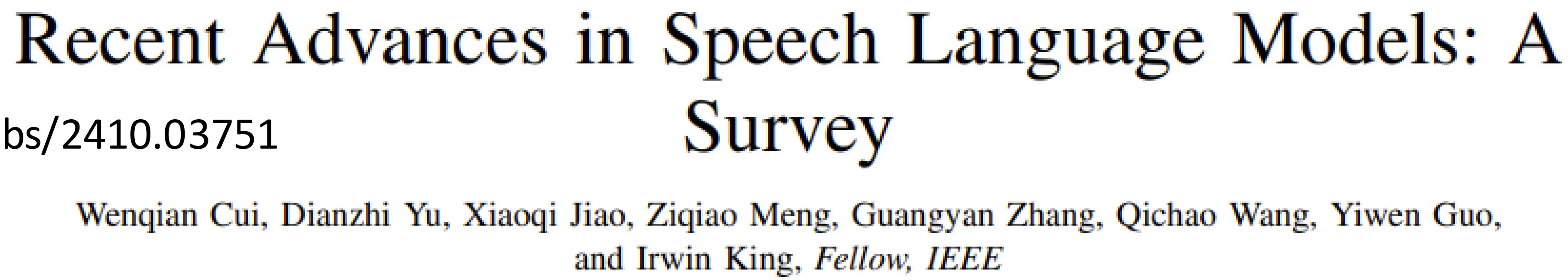
Contributors

Kai-Wei Chang | Haibin Wu | Wei-Cheng Tseng

Kehan Lu | Chun-Yi Kuan | Hung-yi Lee

For paper list: https://github.com/ga642381/speech-trident

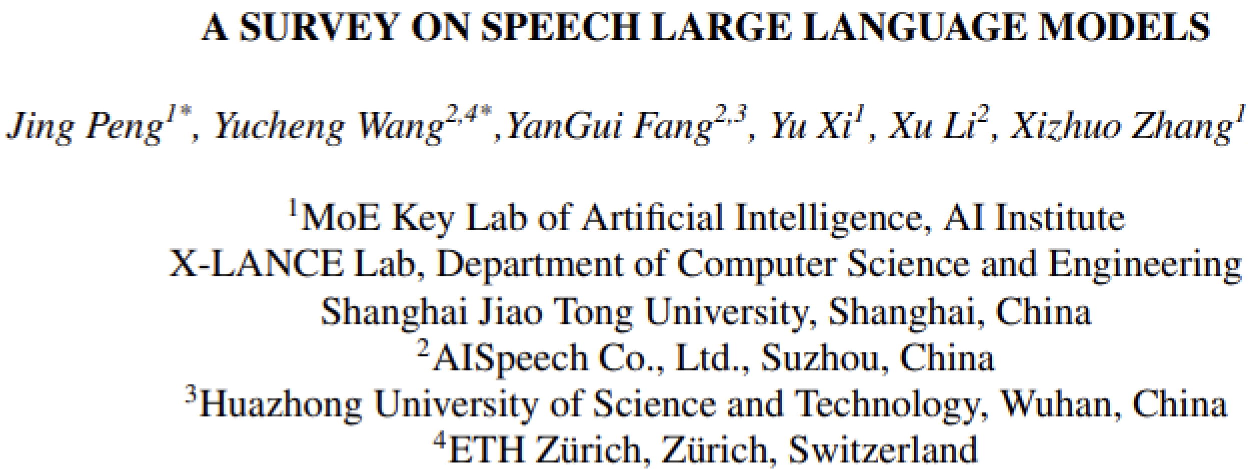




Recent Advances in Speech Language Models: A Survey

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

https://arxiv.org/abs/2410.03751

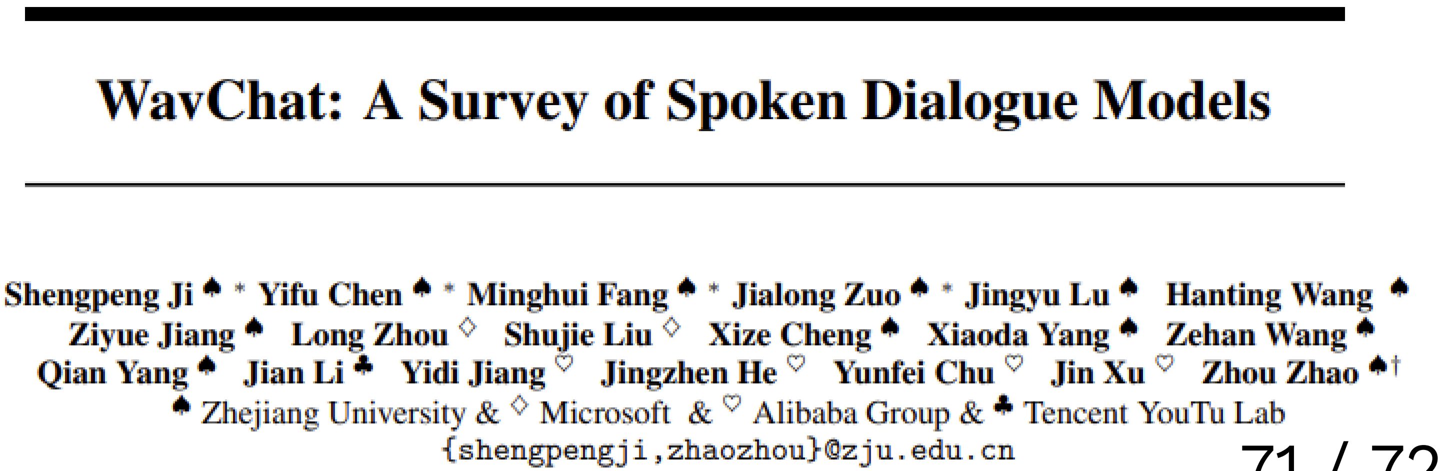


A SURVEY ON SPEECH LARGE LANGUAGE MODELS

Jing Peng[1], Yucheng Wang[2,4*], YanGui Fang[2,3], Yu Xi[1], Xu Li[2], Xizhuo Zhang[1], Kai Yu[1†]*

[1]MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
[2]AISpeech Co., Ltd., Suzhou, China
[3]Huazhong University of Science and Technology, Wuhan, China
[4]ETH Zürich, Zürich, Switzerland

https://arxiv.org/abs/2410.18908



WavChat: A Survey of Spoken Dialogue Models

Shengpeng Ji [♠*] Yifu Chen [♠*] Minghui Fang [♠*] Jialong Zuo [♠*] Jingyu Lu [♠] Hanting Wang [♠] Ziyue Jiang [♠] Long Zhou [◇] Shujie Liu [◇] Xize Cheng [♠] Xiaoda Yang [♠] Zehan Wang [♠] Qian Yang [♠] Jian Li [♣] Yidi Jiang [♡] Jingzhen He [♡] Yunfei Chu [♡] Jin Xu [♡] Zhou Zhao [♠†]

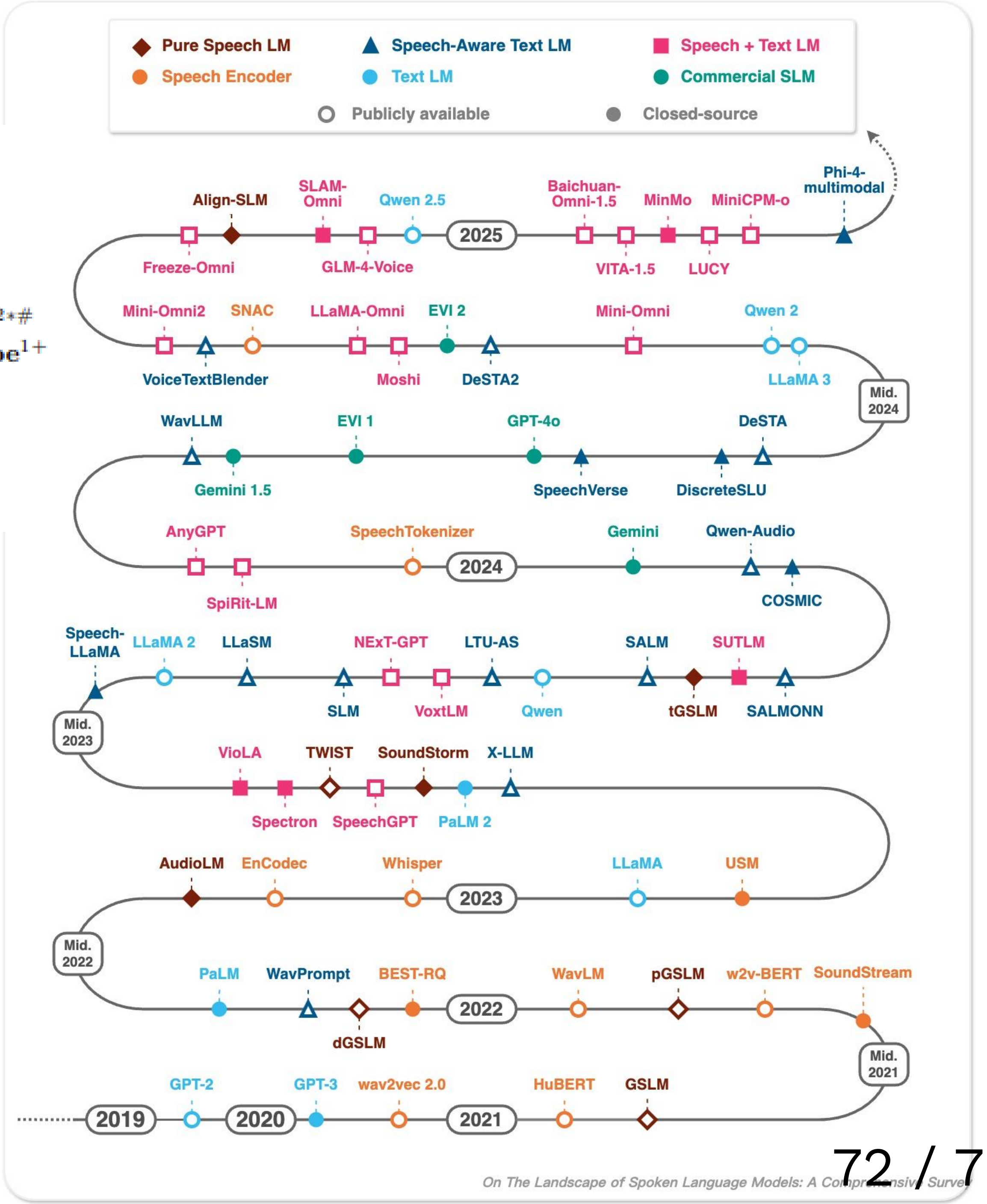
[♠] Zhejiang University & [◇] Microsoft & [♡] Alibaba Group & [♣] Tencent YouTu Lab
{shengpengji,zhaozhou}@zju.edu.cn

https://arxiv.org/abs/2411.13577

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+]
[1] Carnegie Mellon University, USA
[2] National Taiwan University, Taiwan
[3] Toyota Technological Institute at Chicago, USA
[4] Hebrew University of Jerusalem, Israel
[5] ENS - PSL, EHESS, CNRS, France

https://arxiv.org/abs/2504.08528



On The Landscape of Spoken Language Models: A Comprehensive Survey