Towards Better Dynamics in Full-Duplex Spoken Language Models

Speaker: Kai-Wei Chang

2025/11/18

About me

2020 - B.S., EE, NTU, Taiwan

2025 - Ph.D., EECS, NTU, Taiwan

About me

2020 - B.S., EE, NTU, Taiwan

2025 - Ph.D., EECS, NTU, Taiwan

Thesis - *Towards a Universal Speech Model: Prompting Speech Language Models for Diverse Speech Processing Tasks*

About me

2020 - B.S., EE, NTU, Taiwan

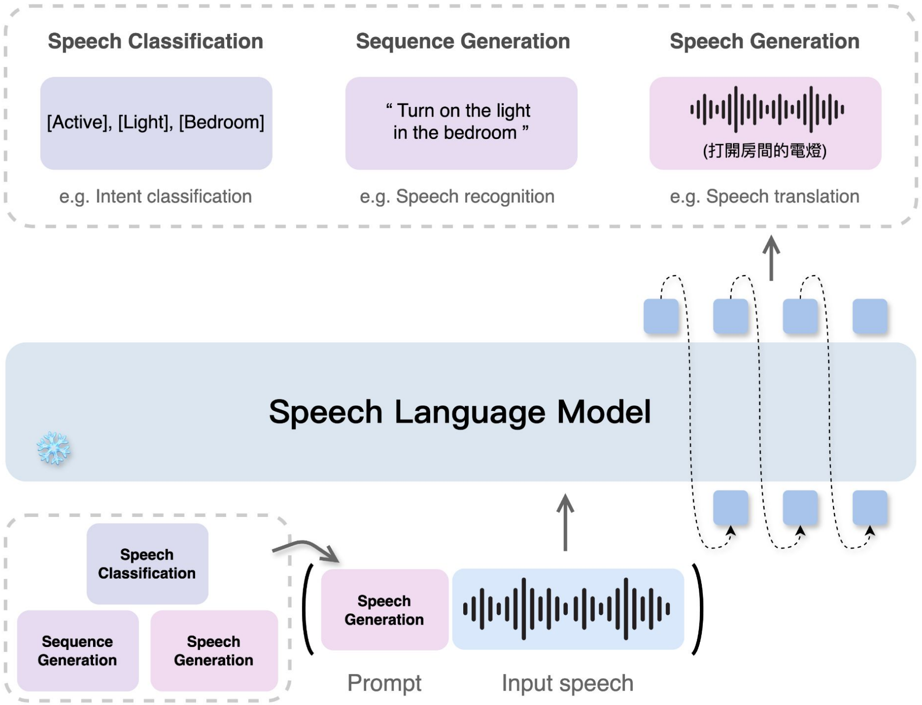
2025 - Ph.D., EECS, NTU, Taiwan

Thesis - *Towards a Universal Speech Model: Prompting Speech Language Models for Diverse Speech Processing Tasks*

SpeechPrompt: Prompting Speech Language Models for Speech Processing Tasks, TASLP 2024

國立臺灣大學
National Taiwan University

SpeechPrompt



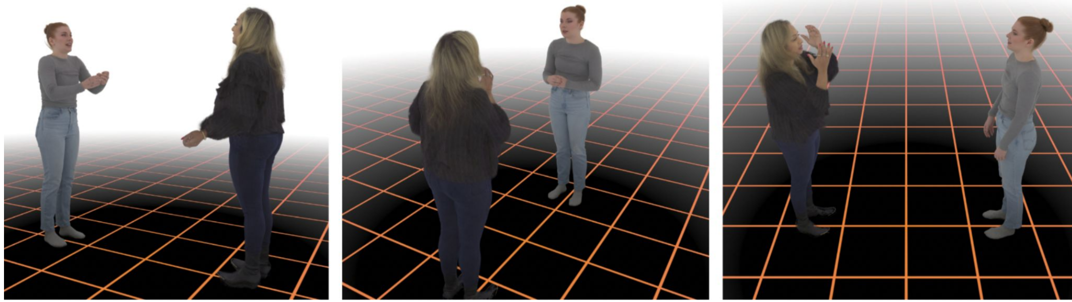
About me

2020 - B.S., EE, NTU, Taiwan

2023 - Meta Reality Labs, Pittsburgh

2025 - Ph.D., EECS, NTU, Taiwan

Codec Avatar



Seamless Interaction: Dyadic Audiovisual Motion Modeling and Large-Scale Dataset
https://arxiv.org/abs/2506.22554

About me

2020	-	B.S., EE, NTU, Taiwan
2023	-	Meta Reality Labs, Pittsburgh
2025	-	Ph.D., EECS, NTU, Taiwan
Now	-	Postdoc. at MIT CSAIL

About me

2020 - B.S., EE, NTU, Taiwan

2023 - Meta Reality Labs, Pittsburgh

2025 - Ph.D., EECS, NTU, Taiwan

Now - Postdoc. at MIT CSAIL

Research Interest

Spoken Language Model

Prompt / Steer for desired behavior

Human - AI / AI - AI Interaction

Outline

- Spoken Language Model (Survey)
  - Taxonomy
  - Full-duplex systems
  - The current benchmarks for full-duplex systems
- Full-duplex-Bench v2
- Game-Time Benchmark
- Activation Steering

On The Landscape of Spoken Language Models: A Comprehensive Survey

* Equal-contribution first authors

\+ Equal-contribution last authors

On The Landscape of Spoken Language Models: A Comprehensive Survey - TMLR 2025 https://arxiv.org/abs/2504.08528

On The Landscape of Spoken Language Models: A Comprehensive Survey

What is a spoken language model?

No rigorous definition,
but what this survey paper focus on...

On The Landscape of Spoken Language Models: A Comprehensive Survey - TMLR 2025 https://arxiv.org/abs/2504.08528

On The Landscape of Spoken Language Models: A Comprehensive Survey

What is a spoken language model?

- Takes speech as input optionally text input and ouput

On The Landscape of Spoken Language Models: A Comprehensive Survey - TMLR 2025 https://arxiv.org/abs/2504.08528

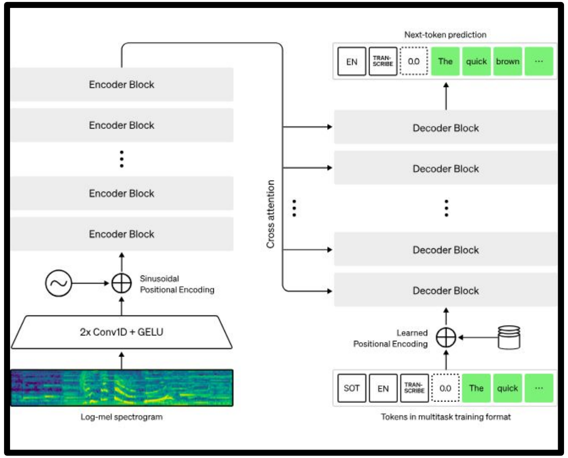
On The Landscape of Spoken Language Models: A Comprehensive Survey

What is a spoken language model?

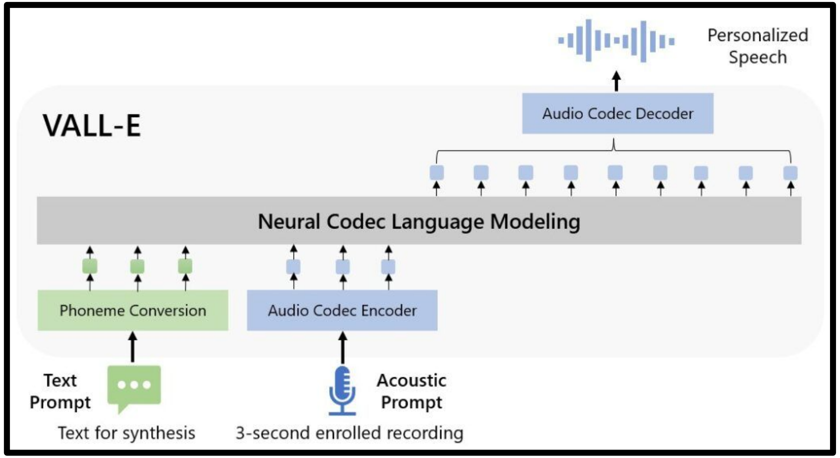
- Takes speech as input optionally text input and ouput
- *Intended* to be “universal” : instruction tuning / dialouge system

On The Landscape of Spoken Language Models: A Comprehensive Survey - TMLR 2025 https://arxiv.org/abs/2504.08528

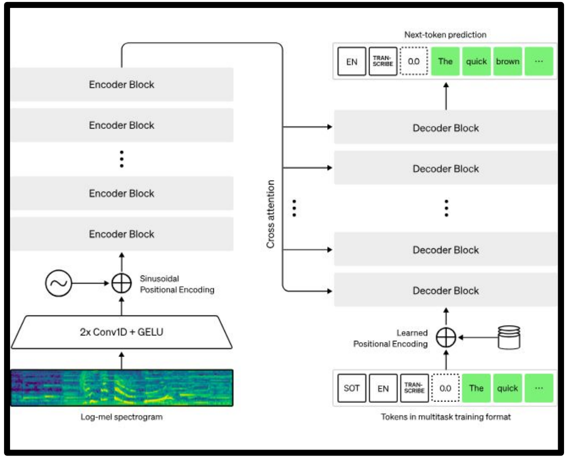
Whisper



VALL-E

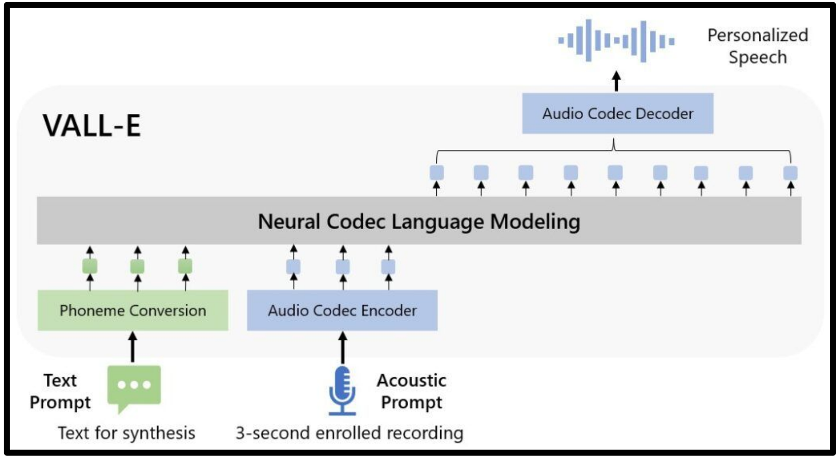


Whisper



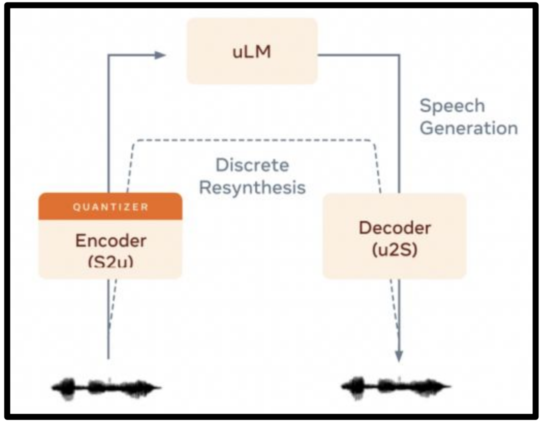
Task-specific:
Speech recognition

VALL-E



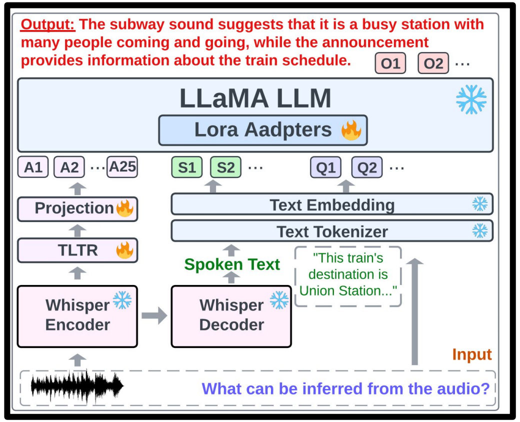
Task-specific:
Text-to-speech

GSLM



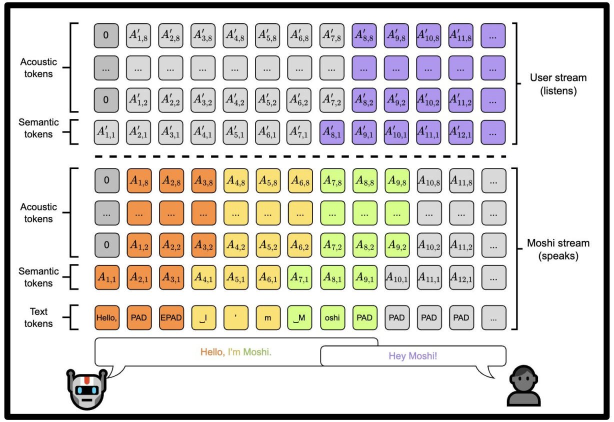
pre-trained on
speech-only data

LTU-AS



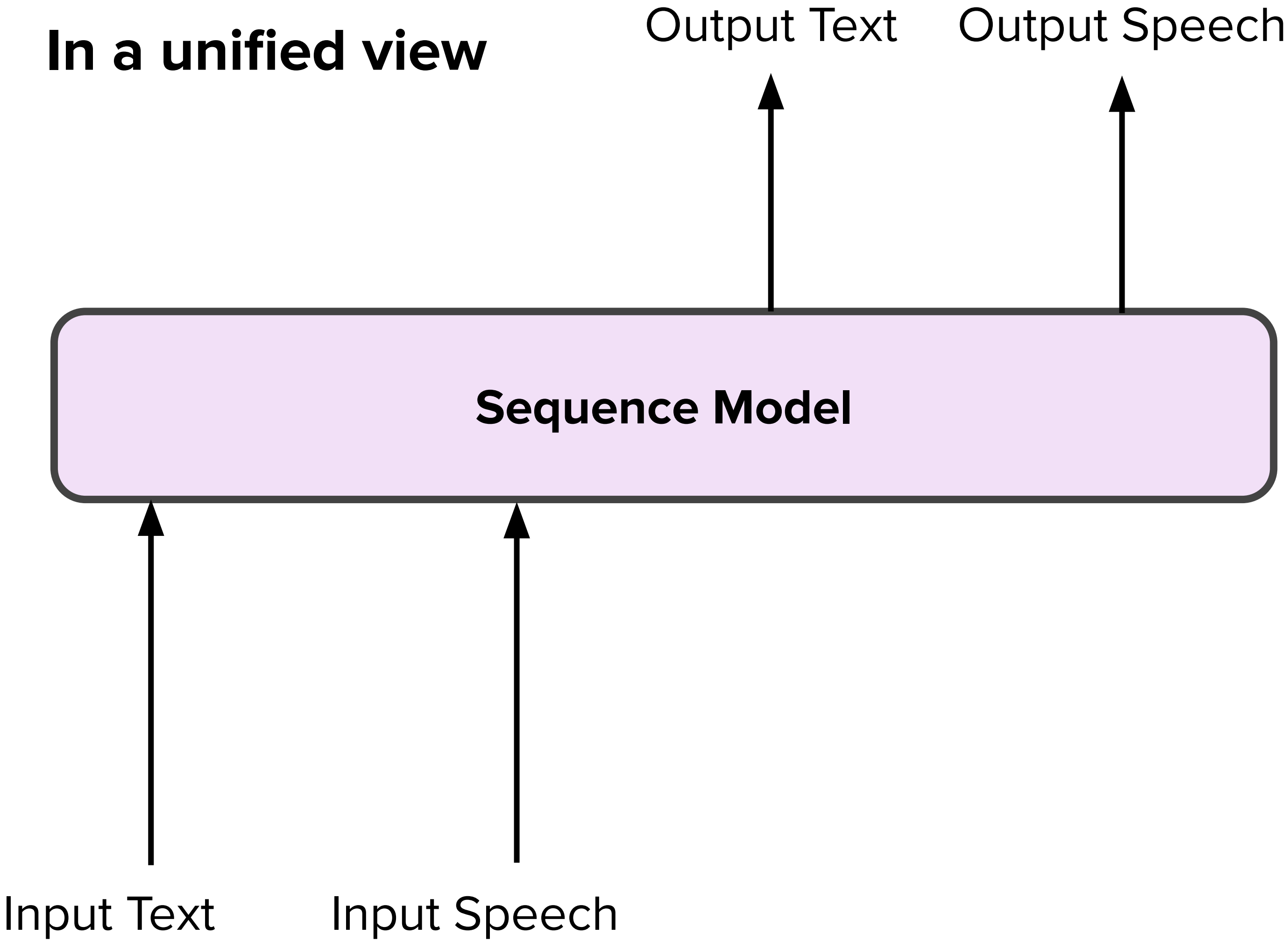
instruction tuned
speech understanding

Moshi

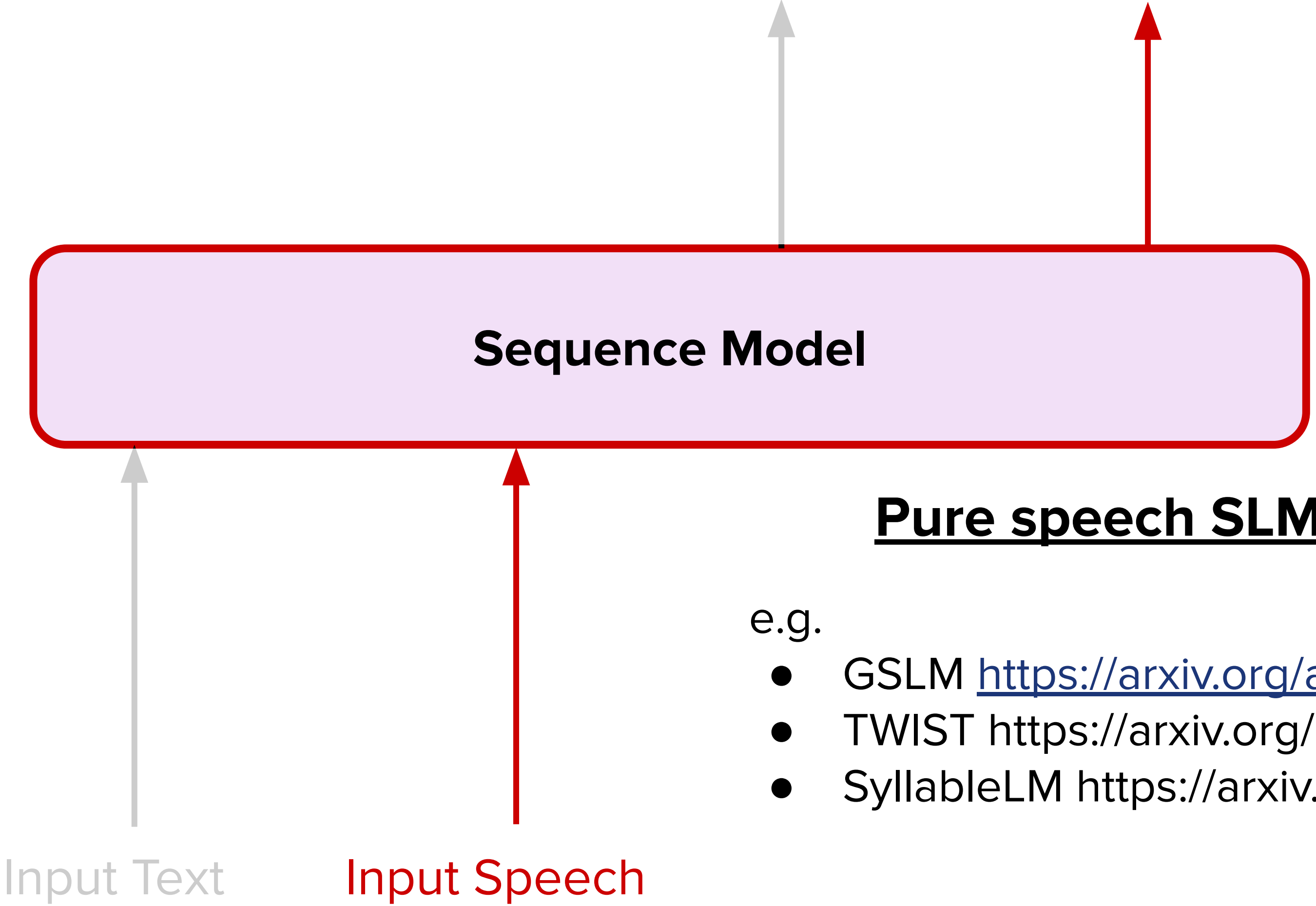


dialouge system

In a unified view



In a unified view

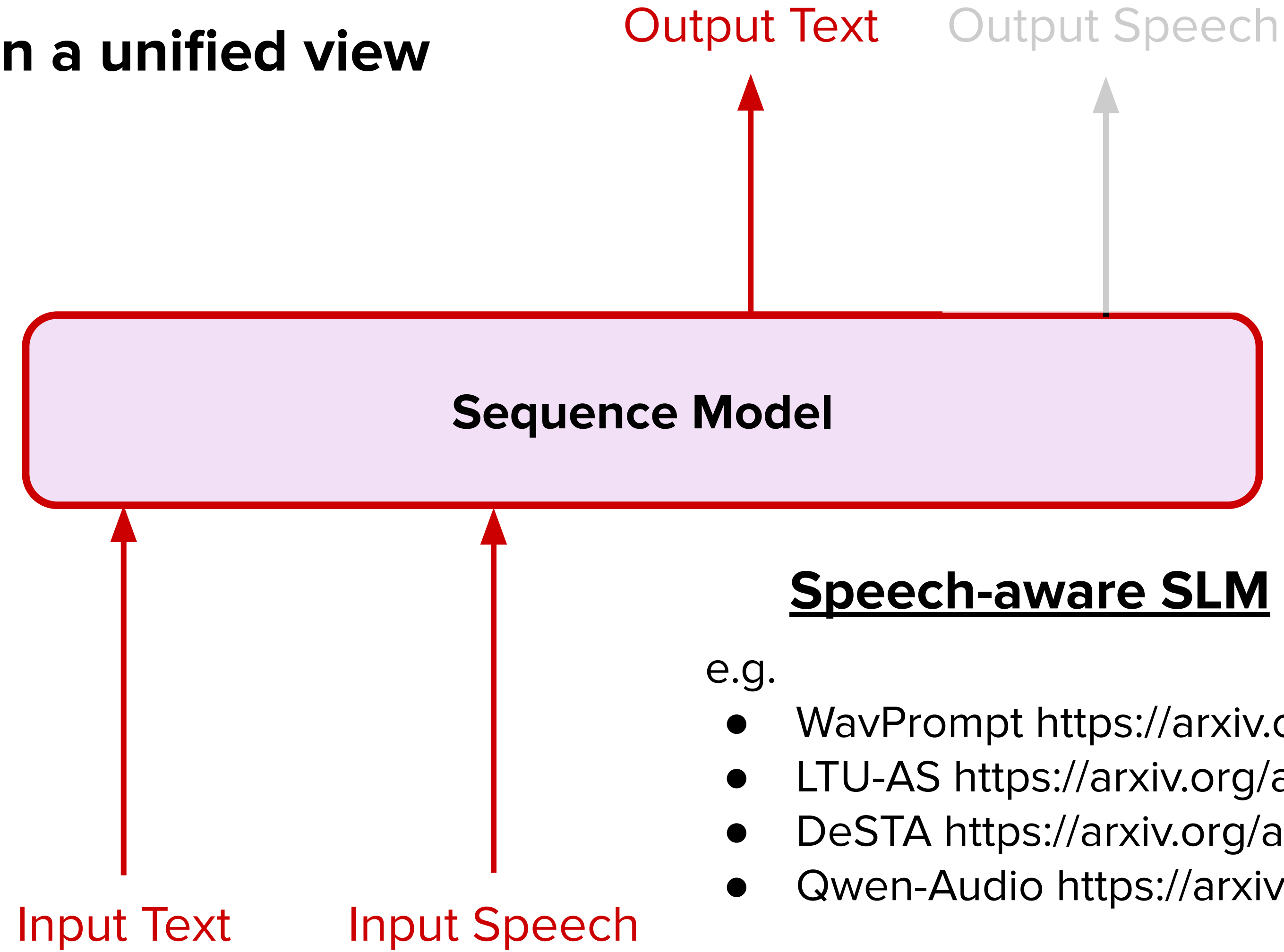


<u>Pure speech SLM</u>

e.g.

- GSLM https://arxiv.org/abs/2102.01192
- TWIST https://arxiv.org/abs/2305.13009
- SyllableLM https://arxiv.org/abs/2410.04029

In a unified view

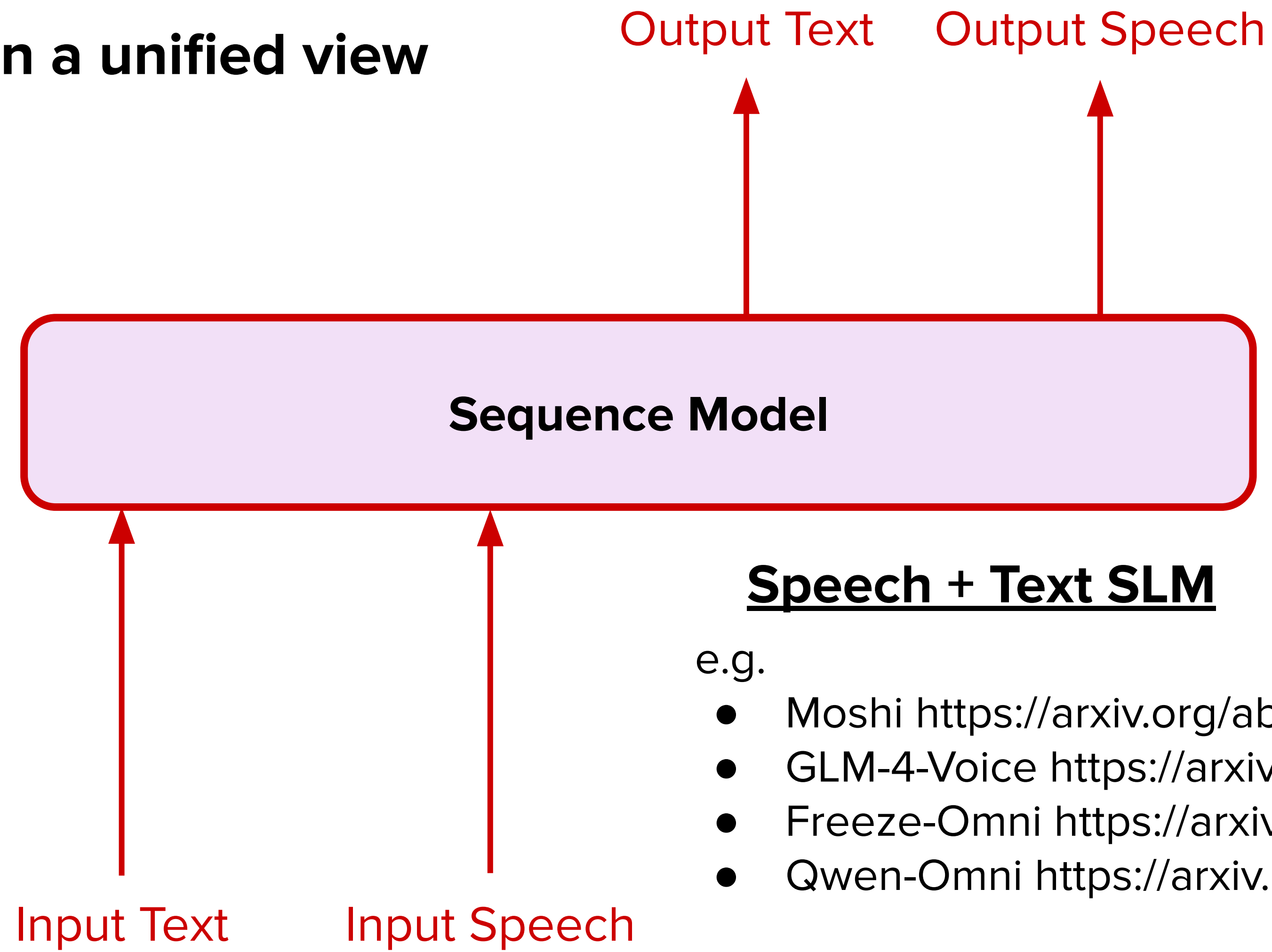


<u>Speech-aware SLM</u>

e.g.

- WavPrompt https://arxiv.org/abs/2203.15863
- LTU-AS https://arxiv.org/abs/2309.14405
- DeSTA https://arxiv.org/abs/2406.18871
- Qwen-Audio https://arxiv.org/abs/2311.07919

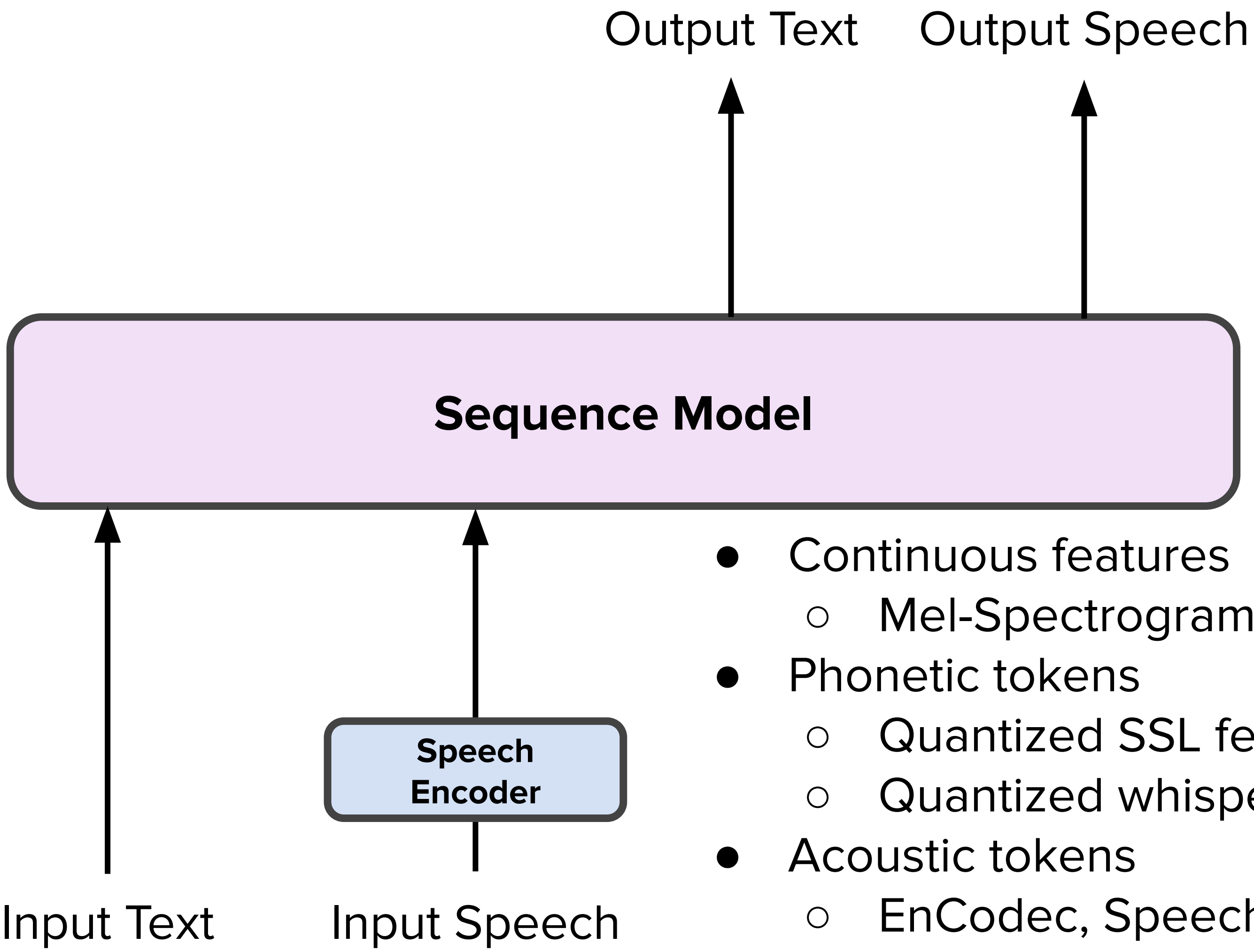
In a unified view



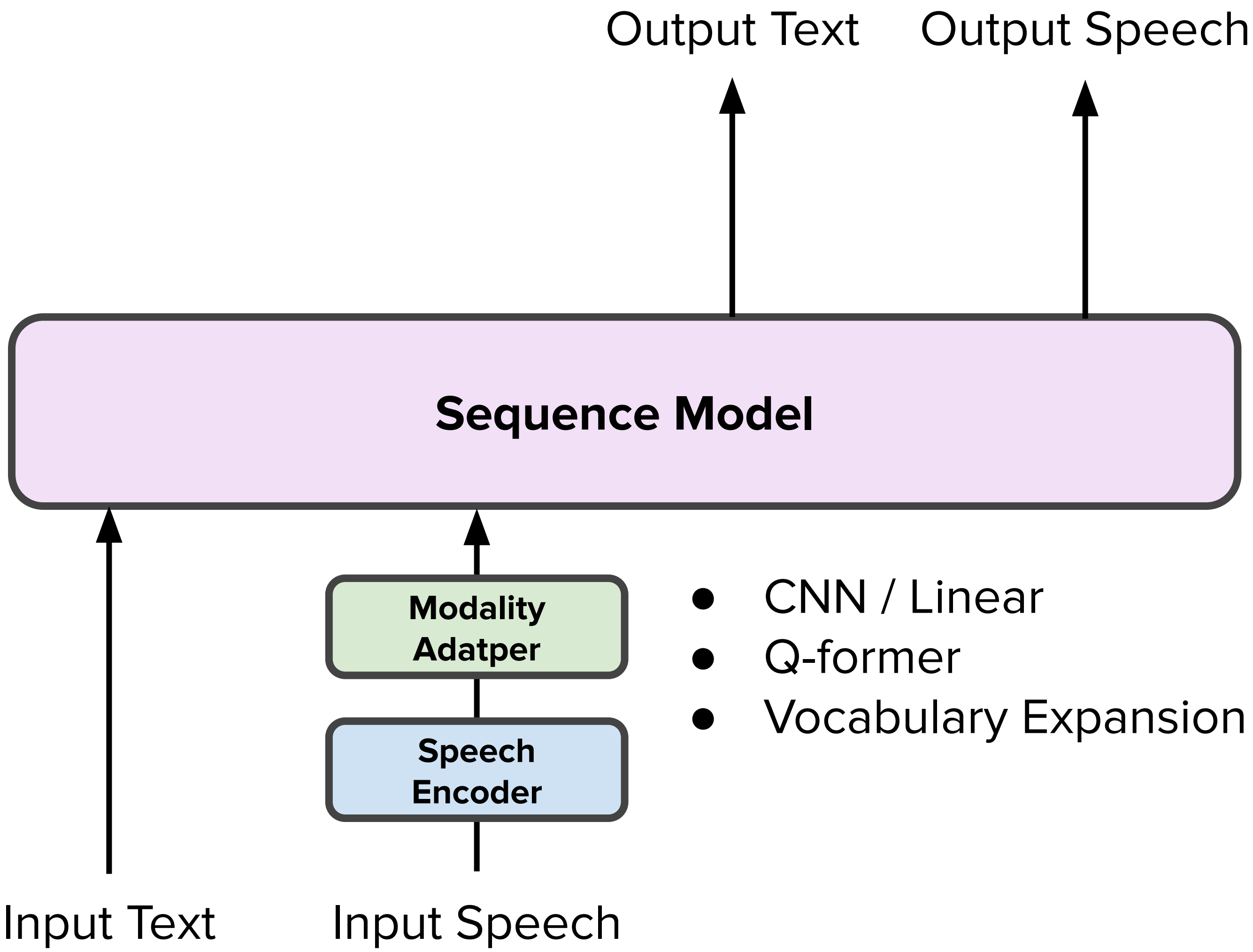
Speech + Text SLM

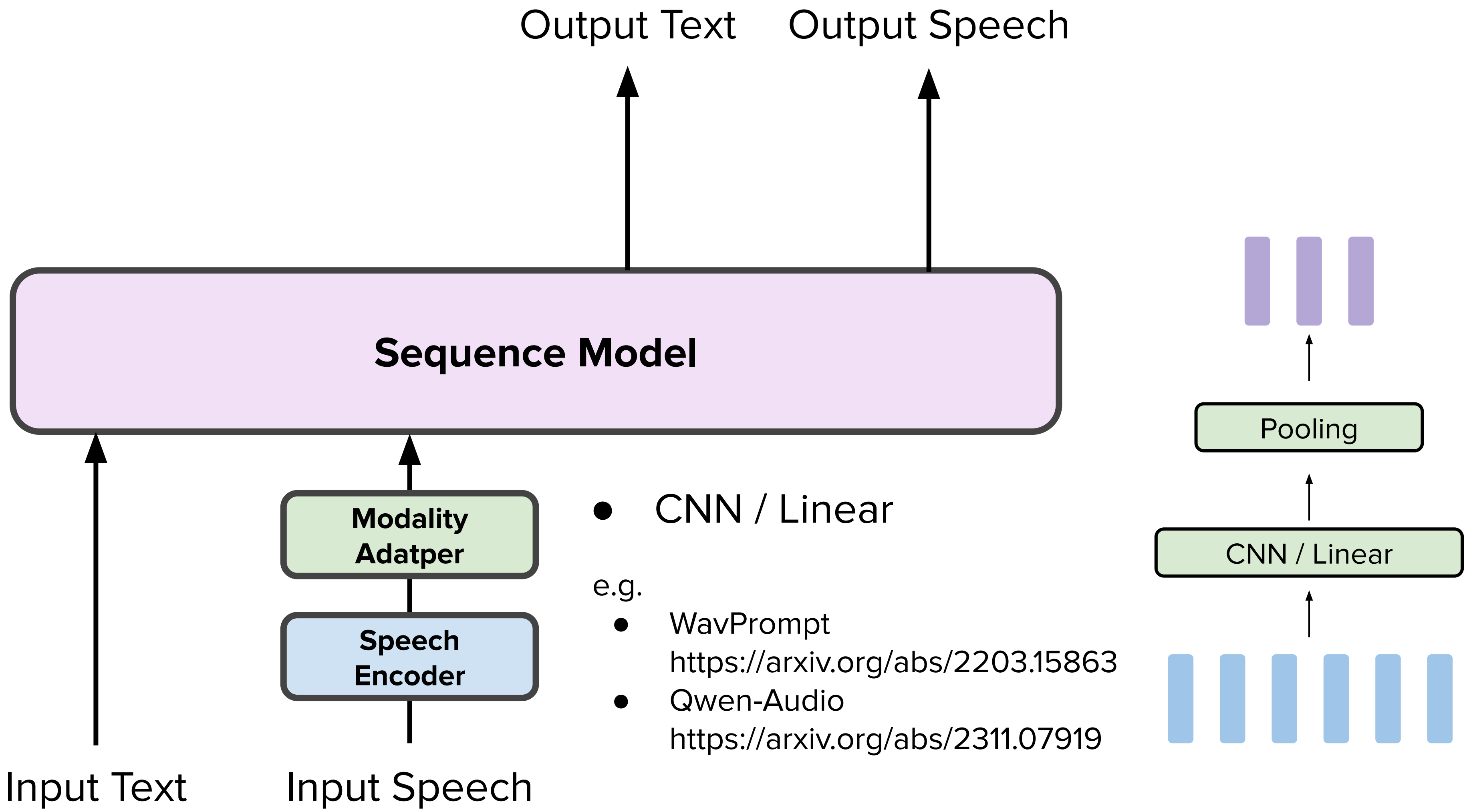
e.g.

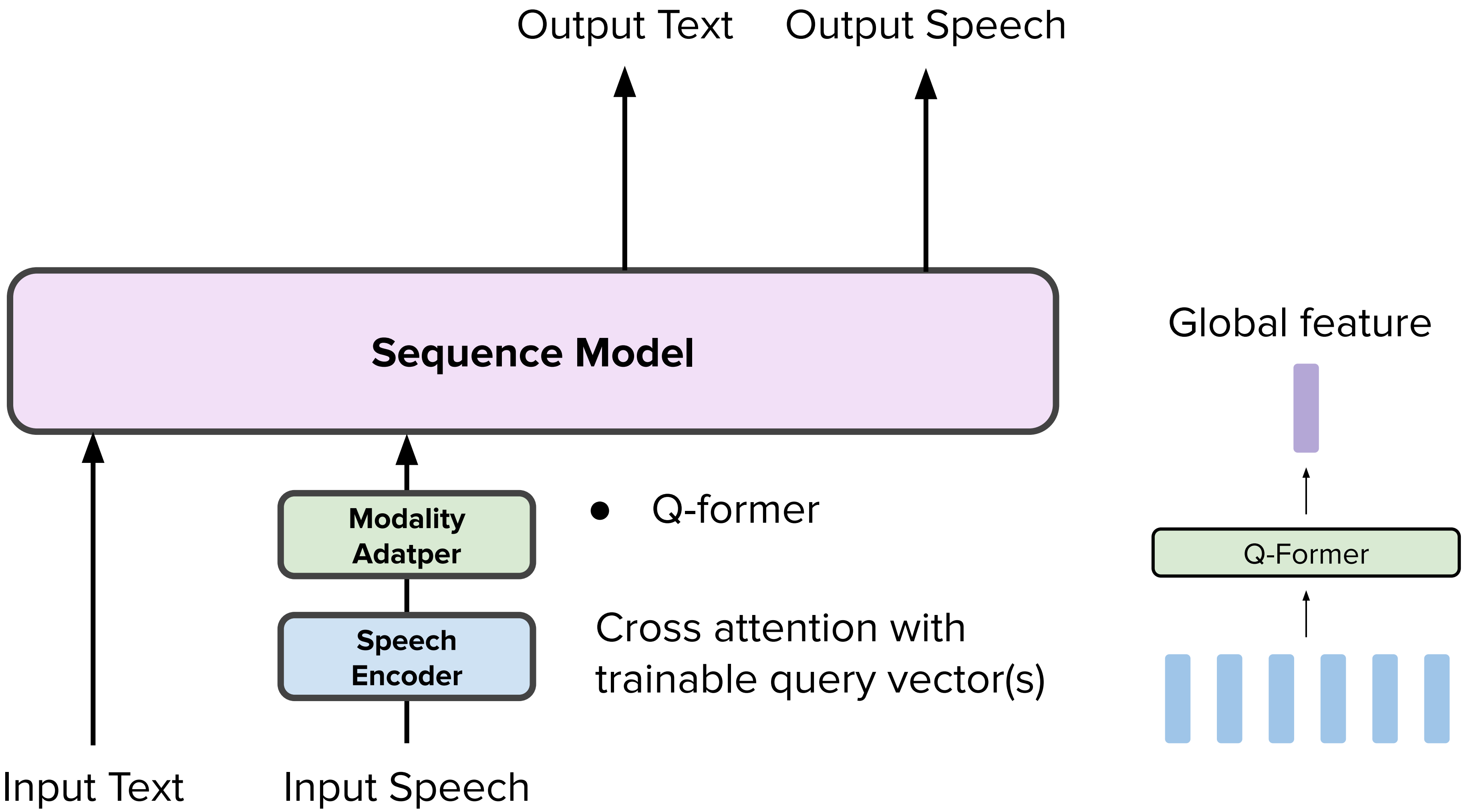
- Moshi https://arxiv.org/abs/2410.00037
- GLM-4-Voice https://arxiv.org/abs/2412.02612
- Freeze-Omni https://arxiv.org/abs/2411.00774
- Qwen-Omni https://arxiv.org/abs/2503.20215

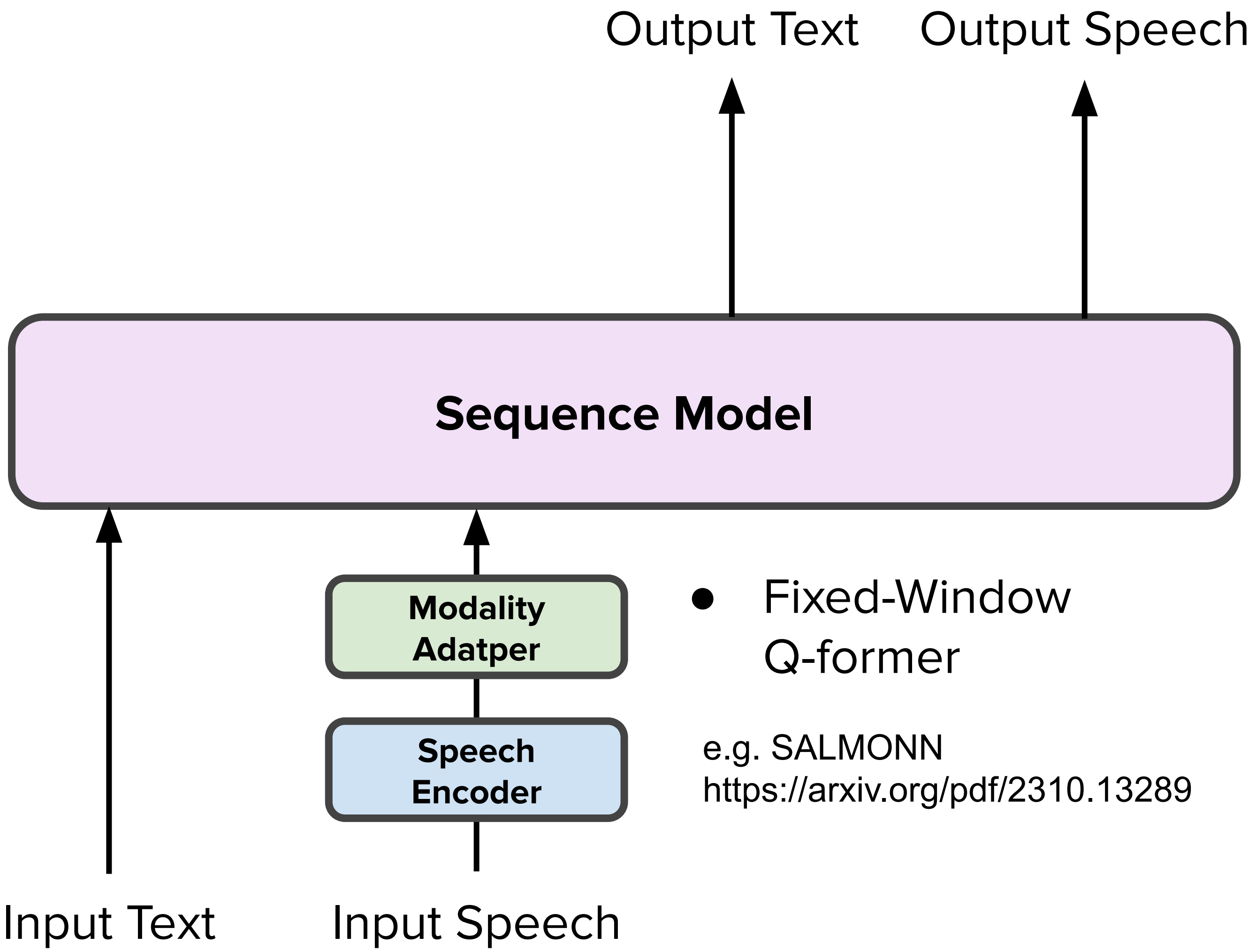


- Continuous features
  - Mel-Spectrogram, SSL feature
- Phonetic tokens
  - Quantized SSL feature
  - Quantized whisper feature
- Acoustic tokens
  - EnCodec, SpeechTokenizer, ...


Output Text
Output Speech
Sequence Model
Modality Adatper
Speech Encoder
Input Text
Input Speech
CNN / Linear
Q-former
Vocabulary Expansion

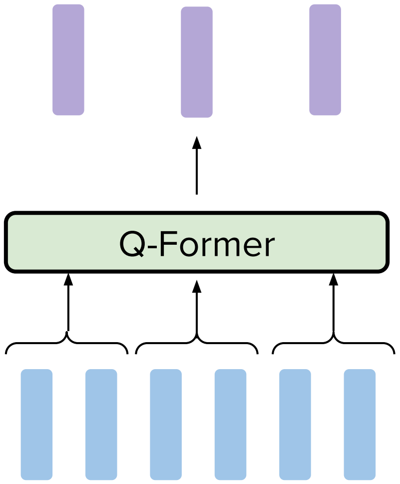

Output Text
Output Speech
Sequence Model
Pooling
Modality Adatper
CNN / Linear
Speech Encoder
Input Text
Input Speech
• CNN / Linear
e.g.
• WavPrompt https://arxiv.org/abs/2203.15863
• Qwen-Audio https://arxiv.org/abs/2311.07919

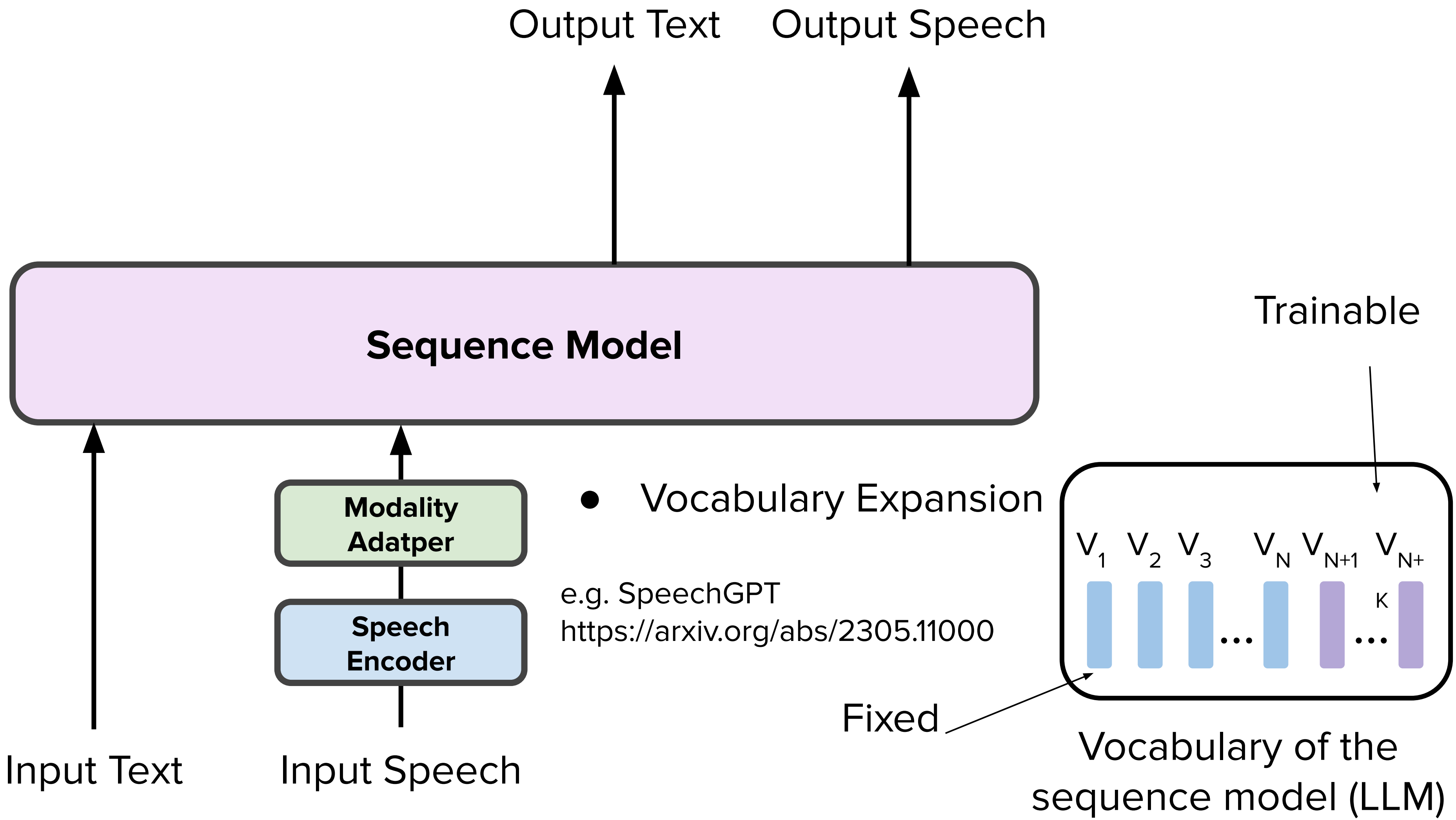

Output Text
Output Speech
Sequence Model
Global feature
Modality Adatper
Speech Encoder
• Q-former
Cross attention with trainable query vector(s)
Q-Former
Input Text
Input Speech

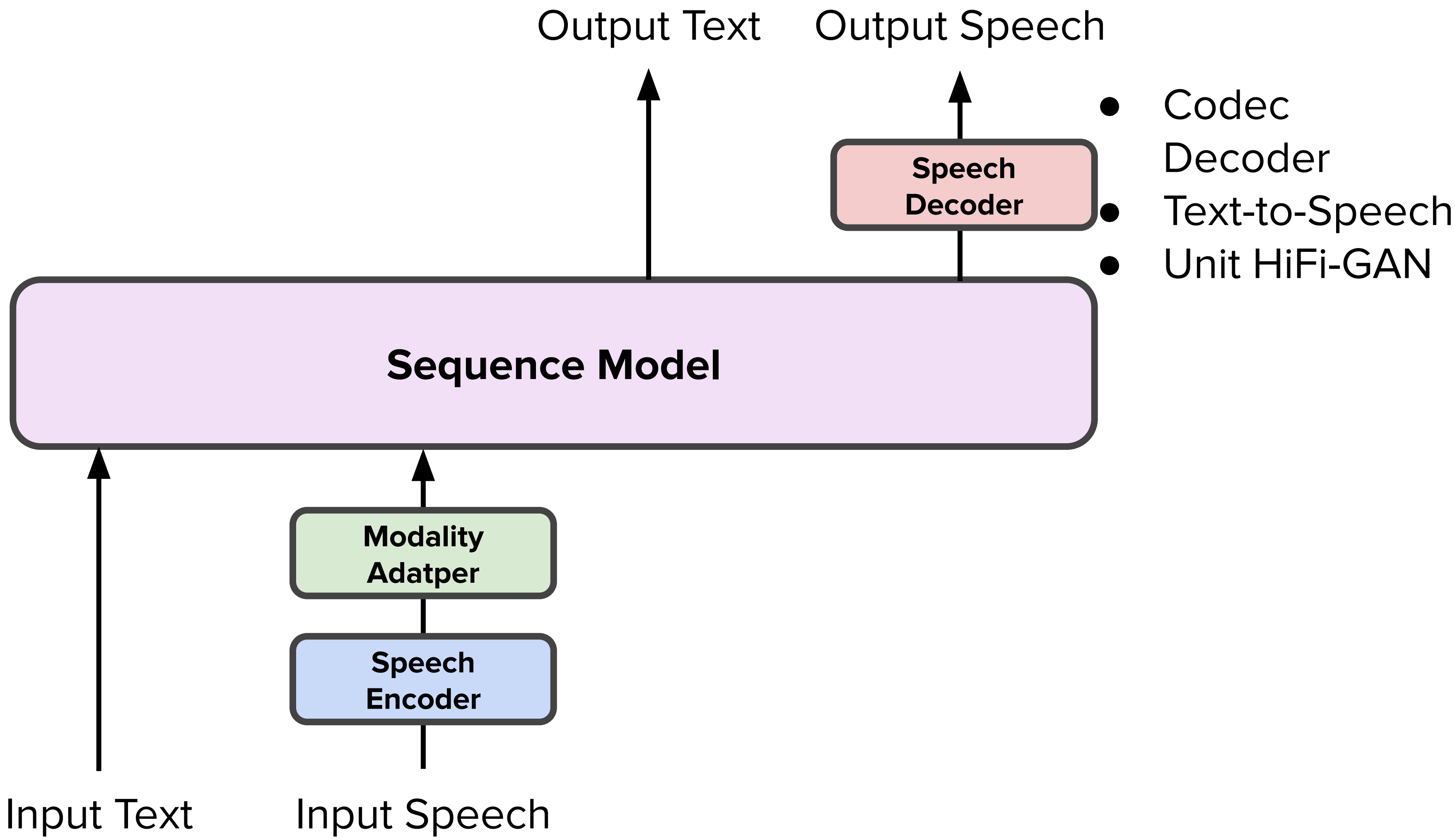


- Fixed-Window Q-former

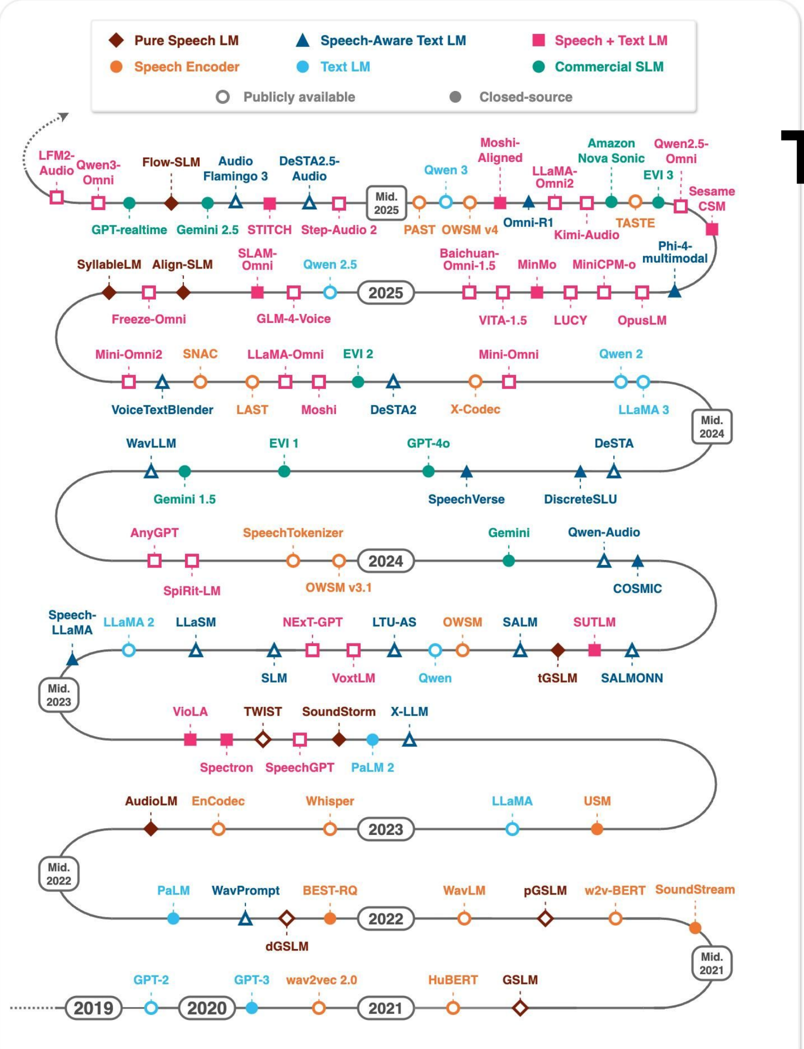
e.g. SALMONN
https://arxiv.org/pdf/2310.13289




Output Text
Output Speech
Sequence Model
Trainable
Modality Adatper
Speech Encoder
• Vocabulary Expansion
e.g. SpeechGPT
https://arxiv.org/abs/2305.11000
V_1 V_2 V_3 … V_N V_{N+1} … V_{N+K}
Fixed
Vocabulary of the sequence model (LLM)
Input Text
Input Speech


Output Text
Output Speech
Codec Decoder
Text-to-Speech
Unit HiFi-GAN
Speech Decoder
Sequence Model
Modality Adatper
Speech Encoder
Input Text
Input Speech

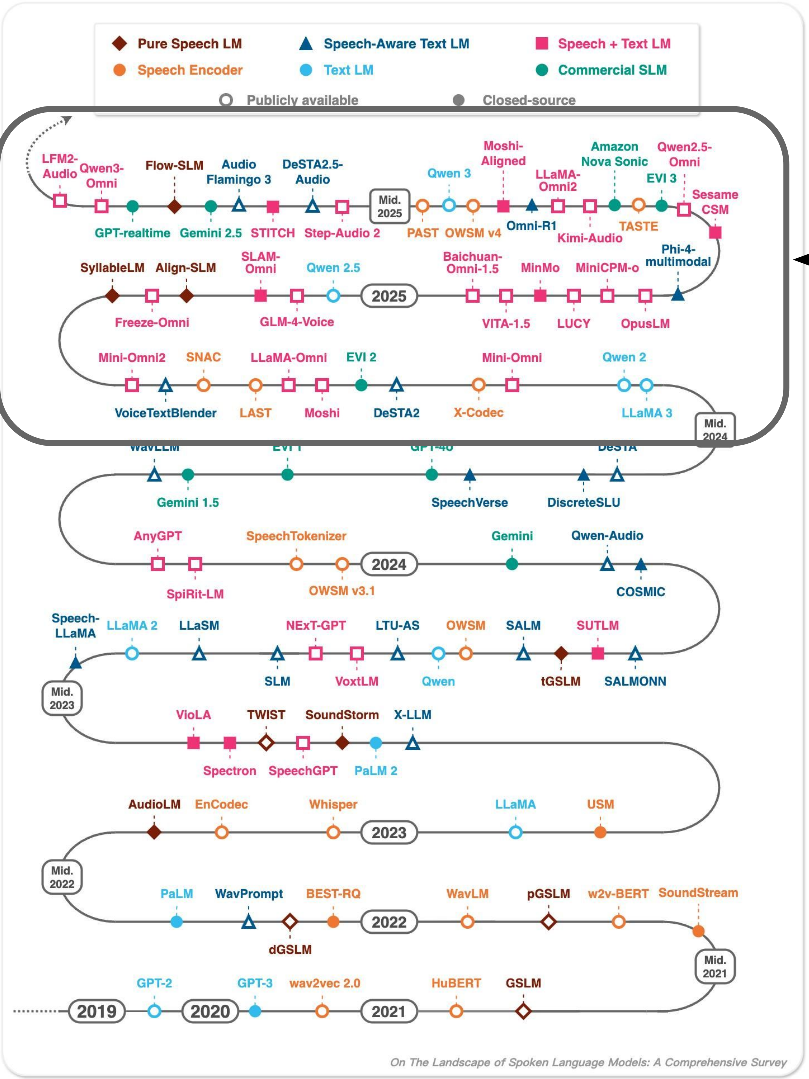
Survey of representative SLMs



On The Landscape of Spoken Language Models: A Comprehensive Survey

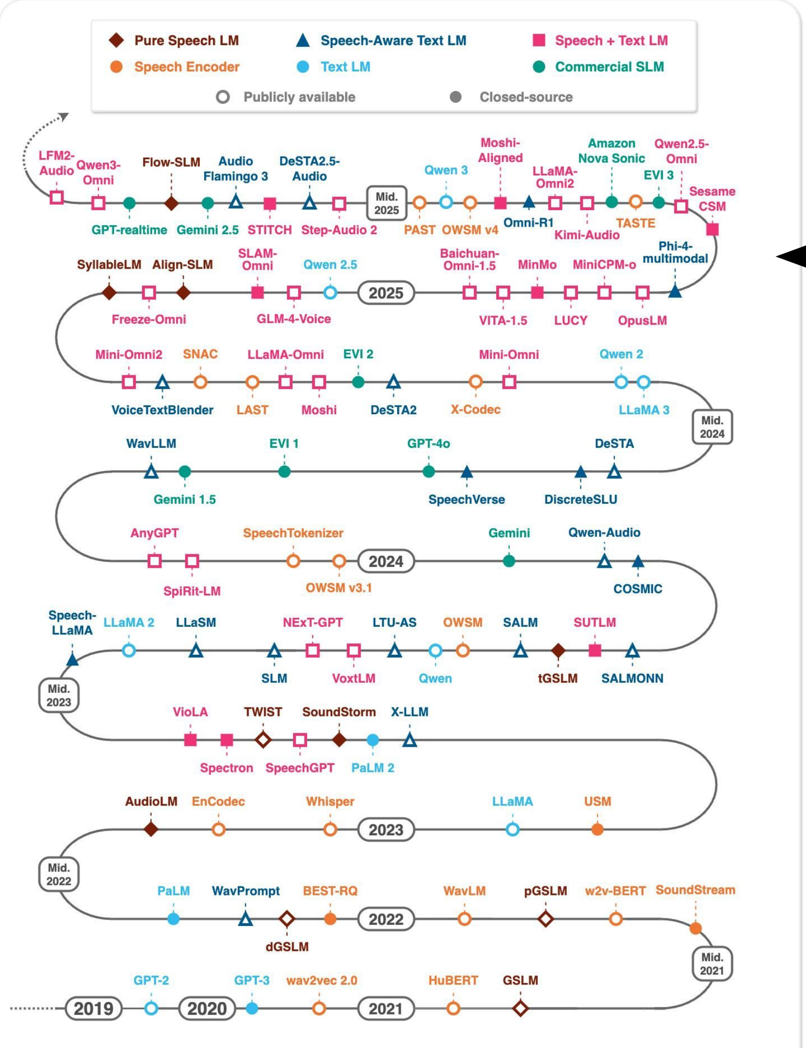
Today

Survey of representative SLMs



On The Landscape of Spoken Language Models: A Comprehensive Survey

a growing trend of Speech + Text LM



On The Landscape of Spoken Language Models: A Comprehensive Survey

Survey of representative SLMs

a growing trend of
Speech + Text LM

Commercial LLMs integrate various modalities

GPT-4o (voice mode)
(May 2024 / Sep 2024)*

Gemini-Live
(Aug 2024)*

*rolling release

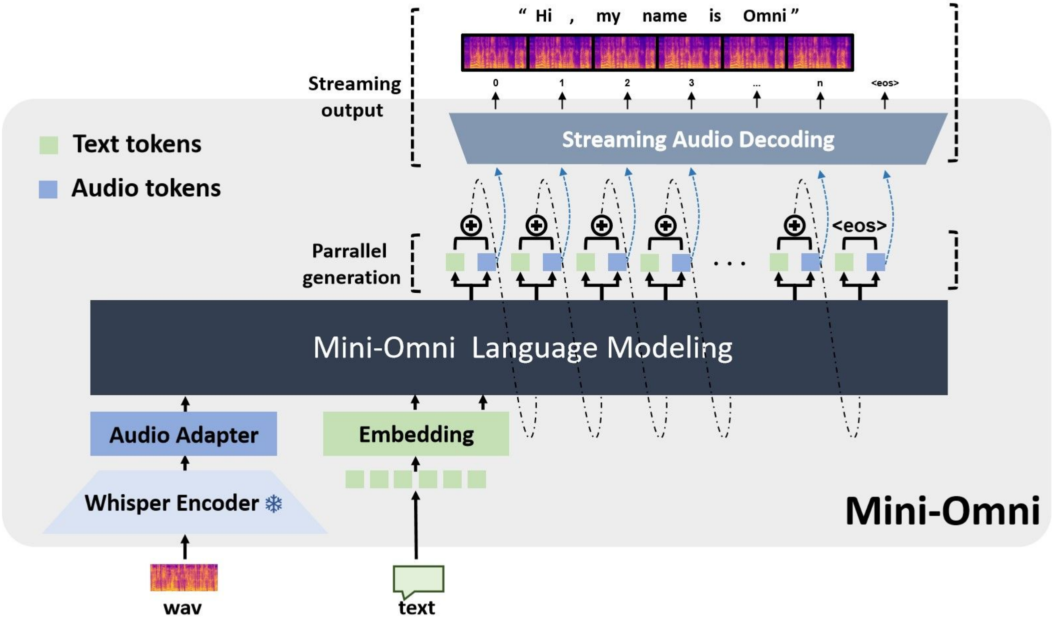
Open source models

Mini-Omni
(Aug 2024)

Moshi
(Sep 2024)

Freeze-Omni
(Nov 2024)

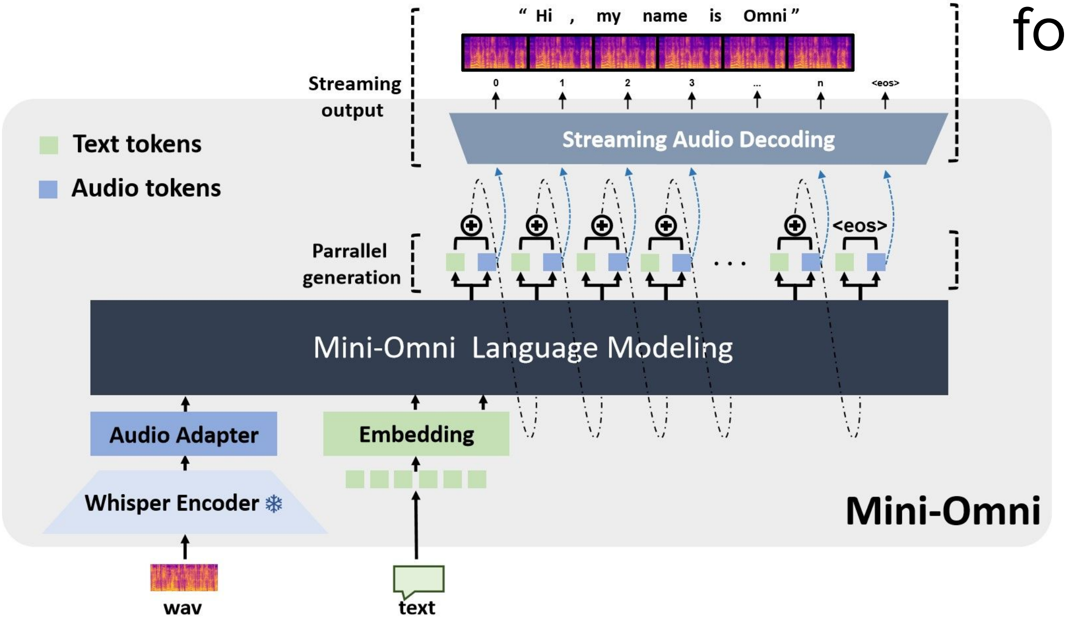
Omni models
focus on understanding multiple modalities



Speech Text

Omni models
focus on understanding multiple modalities

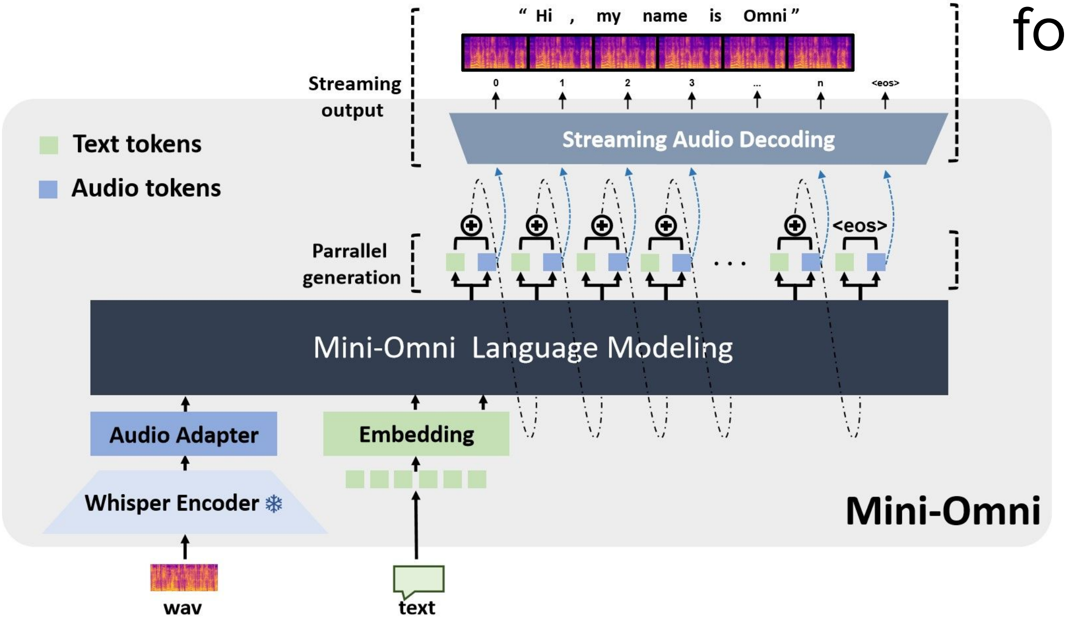
Streaming decoder for low latency



Speech Text

Omni models
focus on understanding multiple modalities

Streaming decoder
for low latency

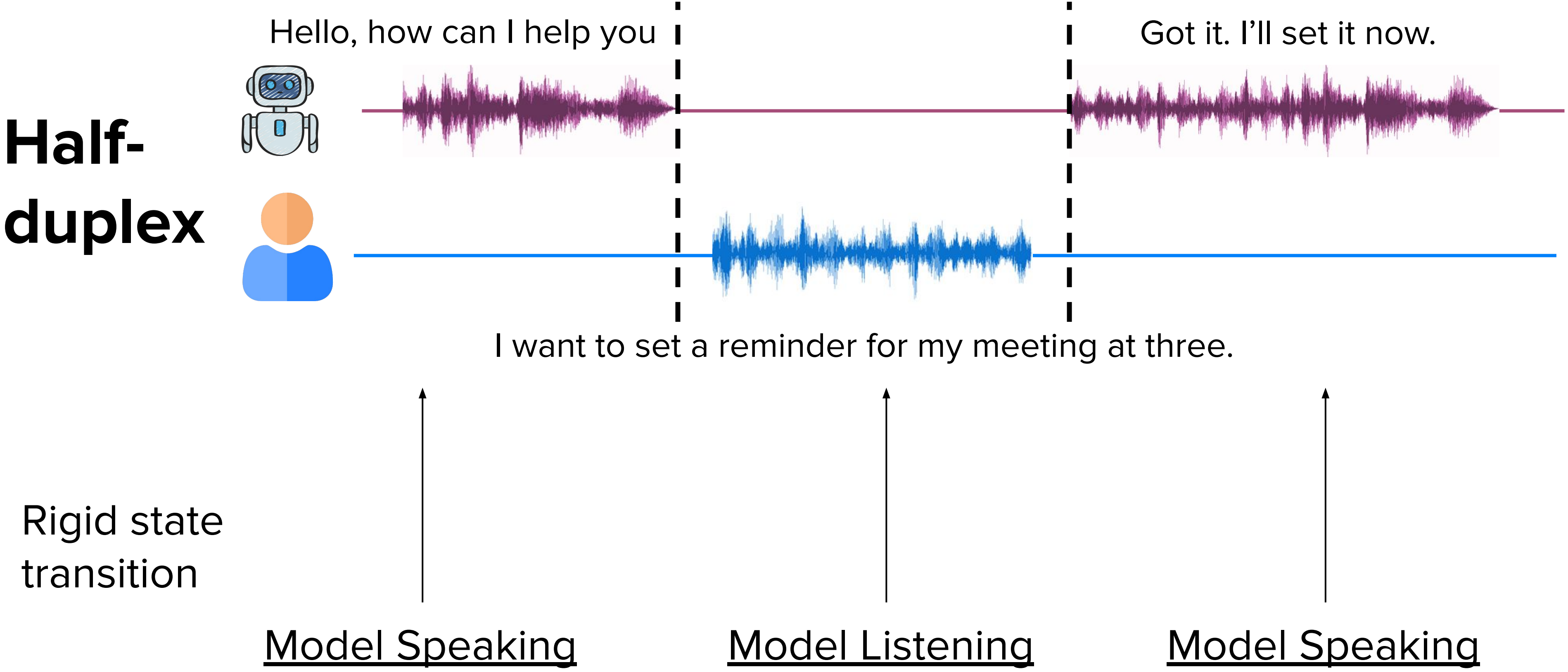


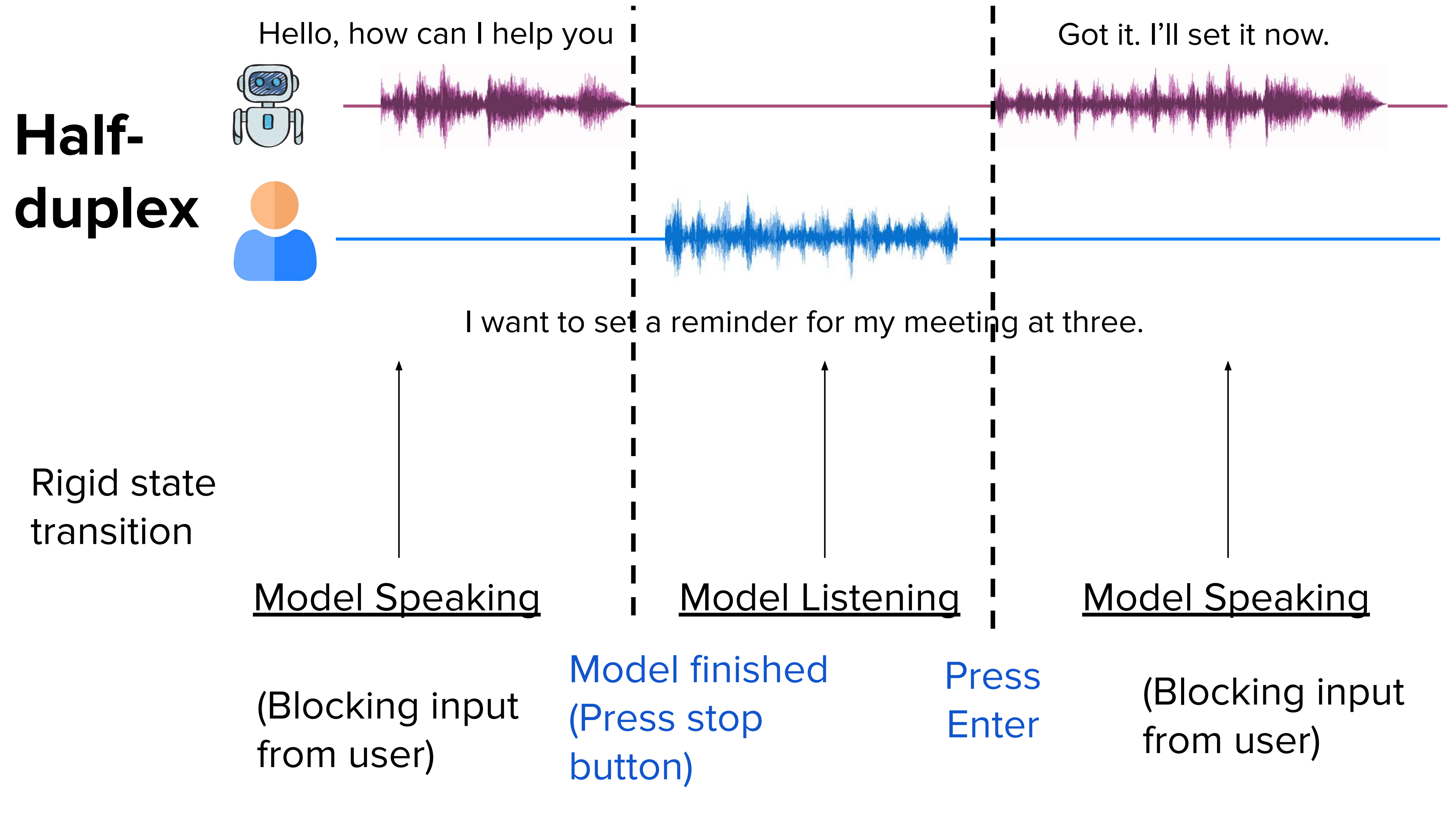
Speech

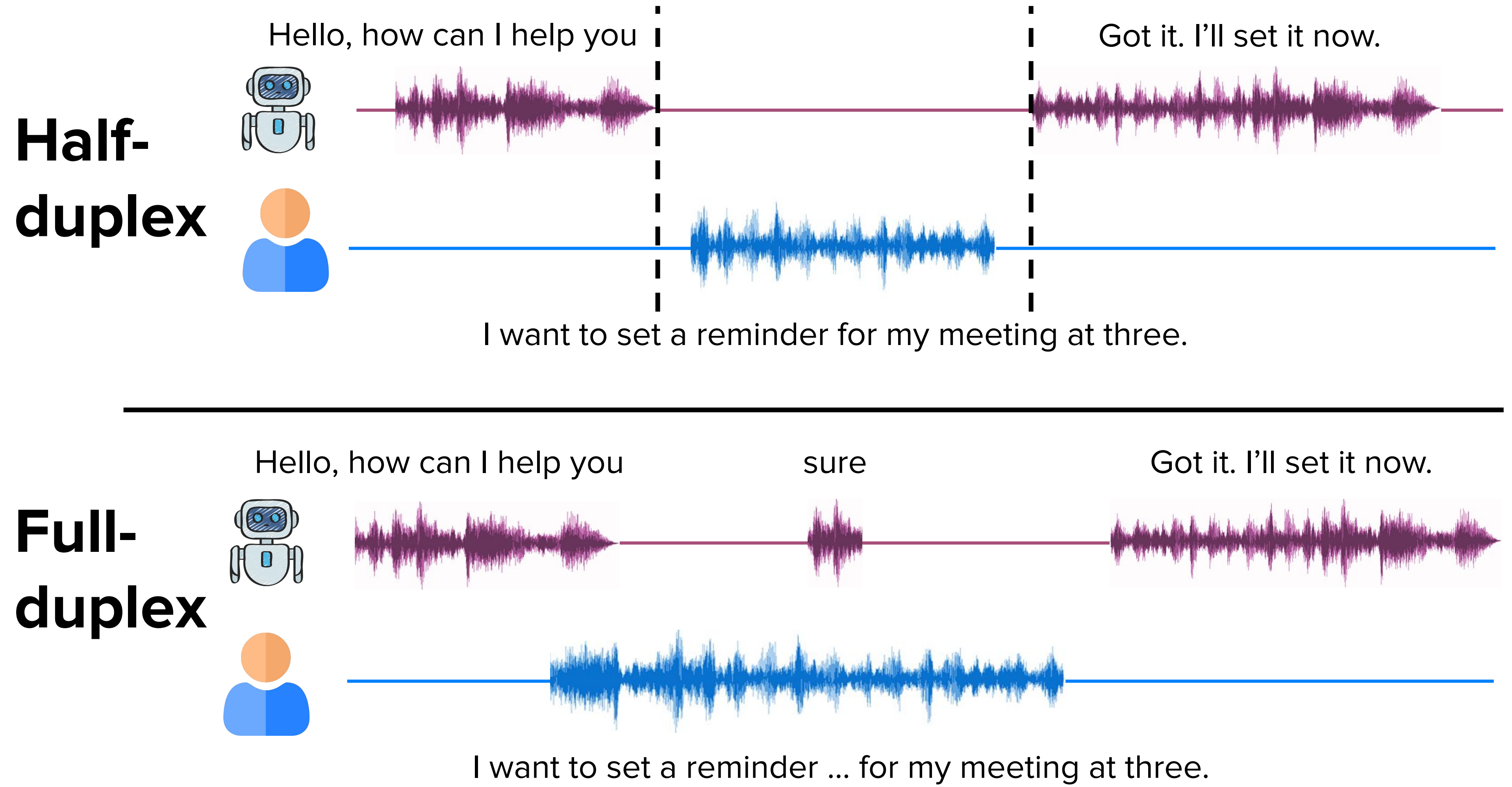
Text

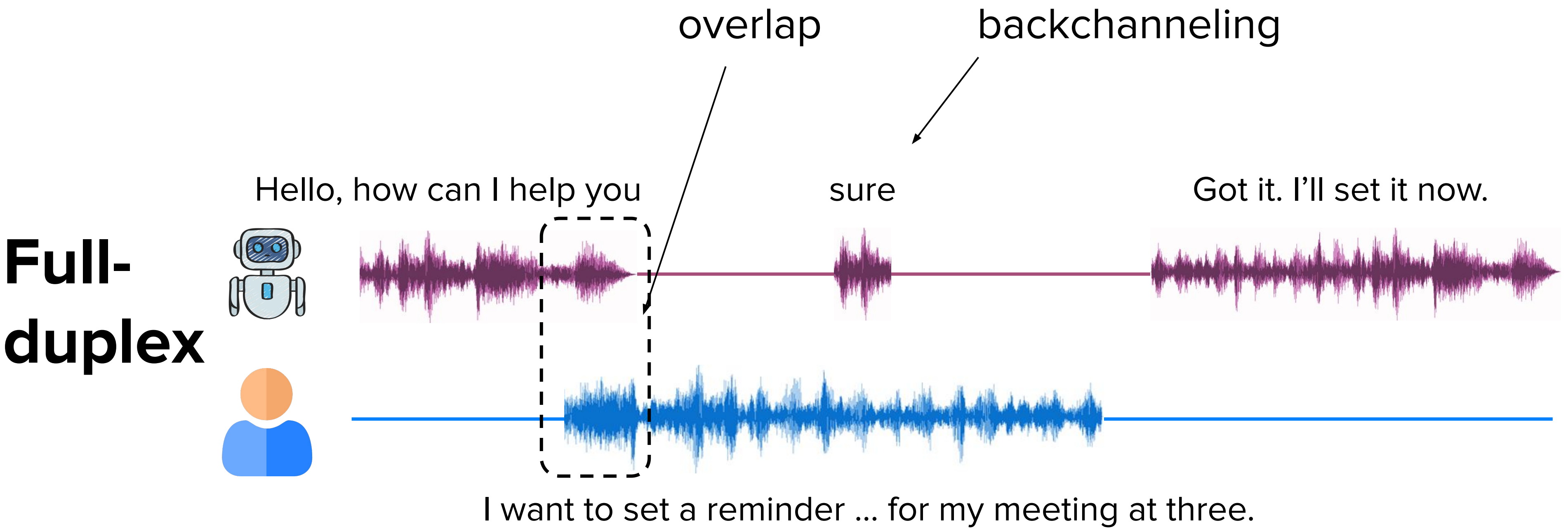
Turn-by-turn
(Half-duplex)


Half-duplex
Hello, how can I help you
Got it. I’ll set it now.
I want to set a reminder for my meeting at three.

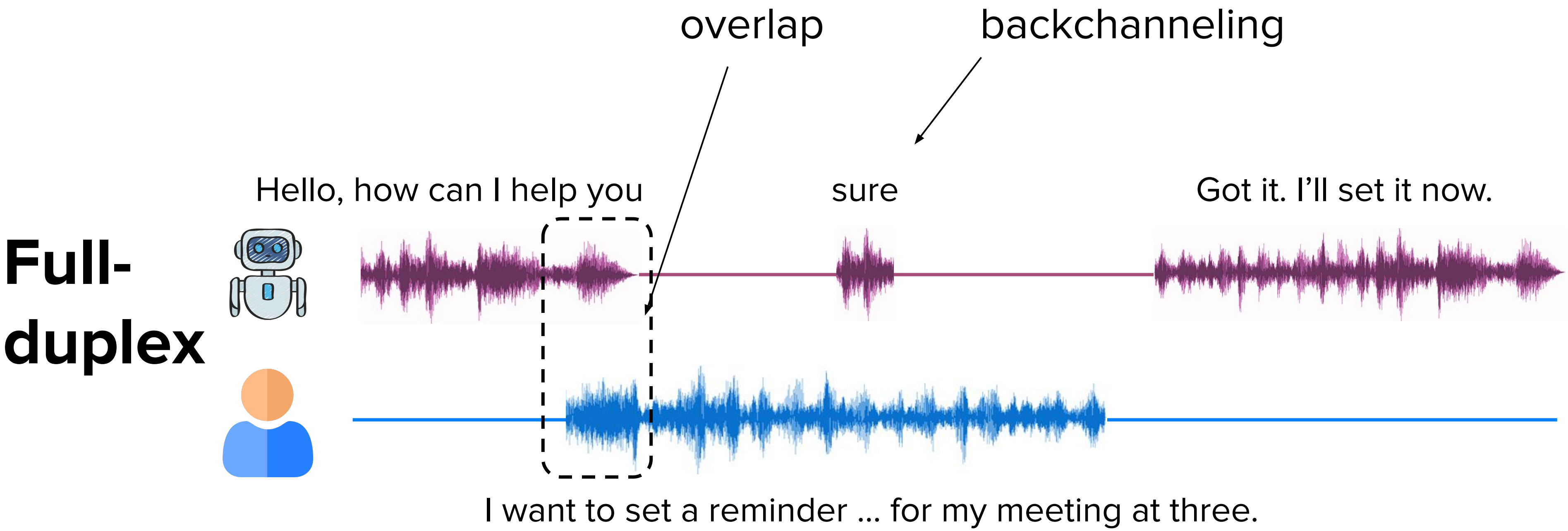

Half-duplex
Hello, how can I help you
Got it. I’ll set it now.
I want to set a reminder for my meeting at three.
Rigid state transition
Model Speaking
Model Listening
Model Speaking


Hello, how can I help you
Got it. I'll set it now.
Half-duplex
I want to set a reminder for my meeting at three.
Rigid state transition
Model Speaking
Model Listening
Model Speaking
(Blocking input from user)
Model finished (Press stop button)
Press Enter
(Blocking input from user)


Half-duplex
Hello, how can I help you
Got it. I'll set it now.
I want to set a reminder for my meeting at three.
Full-duplex
Hello, how can I help you
sure
Got it. I'll set it now.
I want to set a reminder ... for my meeting at three.


Full-duplex
overlap
backchanneling
Hello, how can I help you
sure
Got it. I'll set it now.
I want to set a reminder ... for my meeting at three.

Full-duplex models are capable of listening and speaking at the same time



Full-duplex models are capable of listening and speaking at the same time

overlap

backchanneling

Hello,

it. I'll set it now.

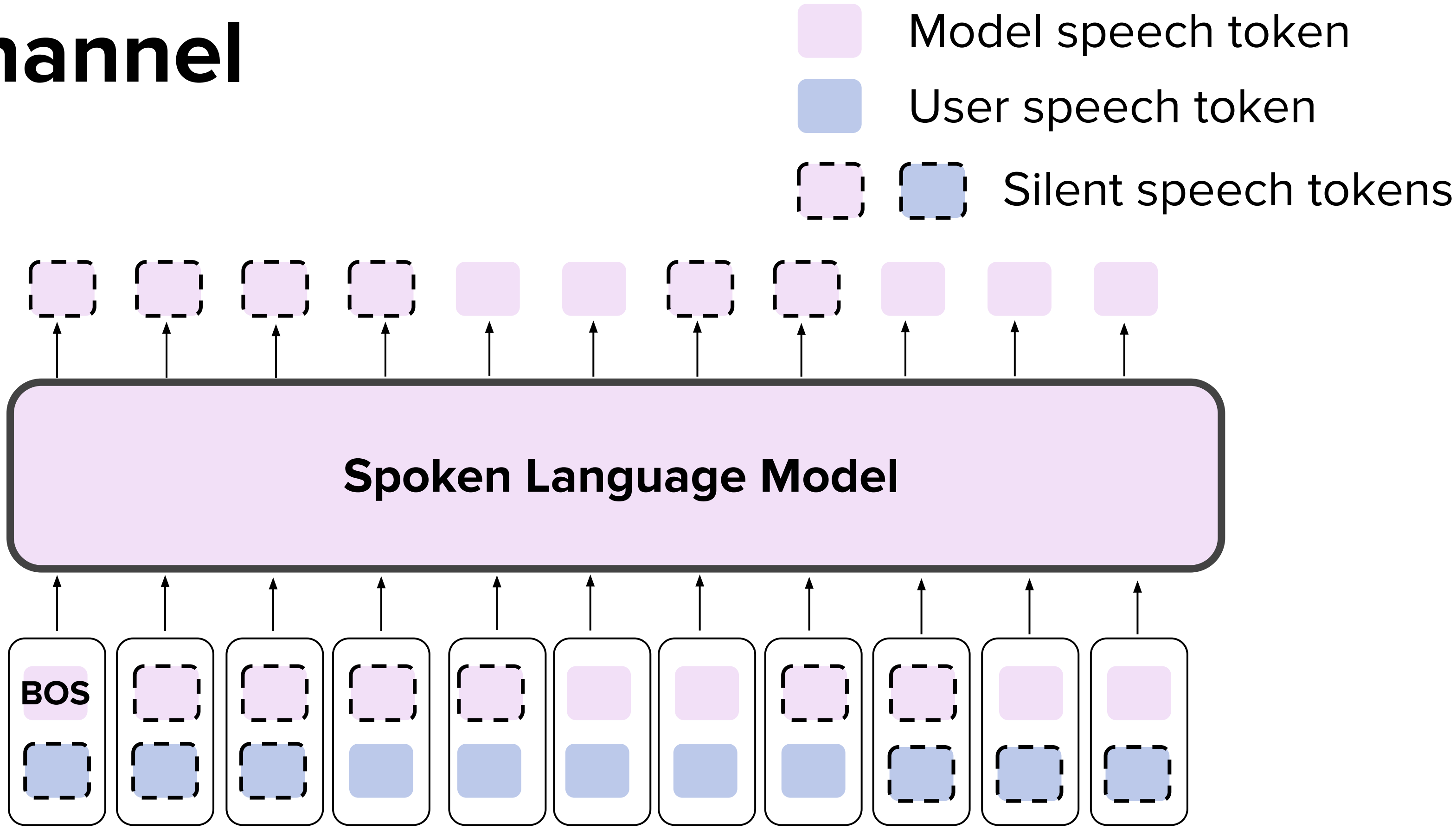
Full-duplex

How to achieve full-duplex?

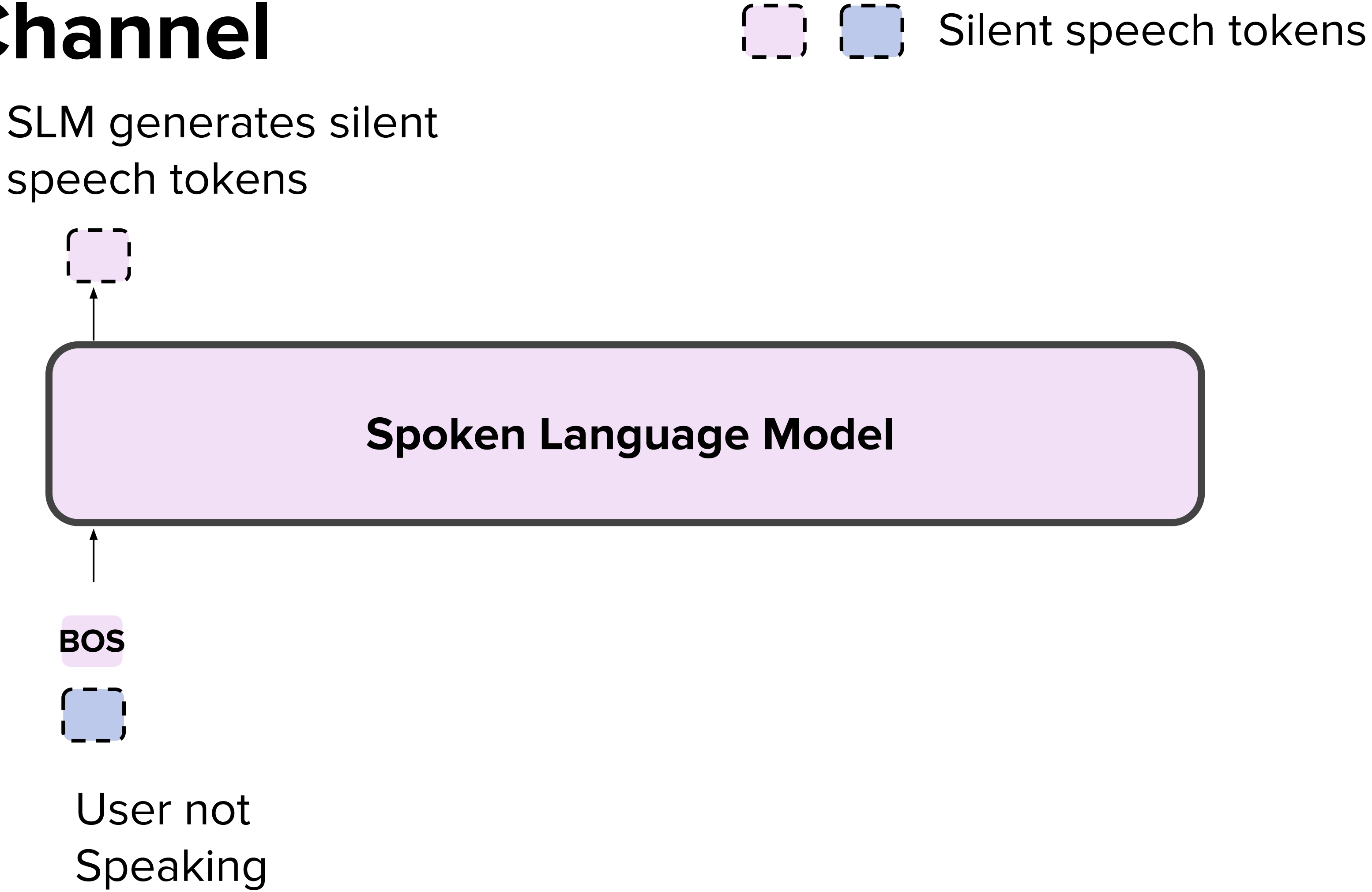
(1) Dual-Channel Modeling

(2) Time-Multiplexing

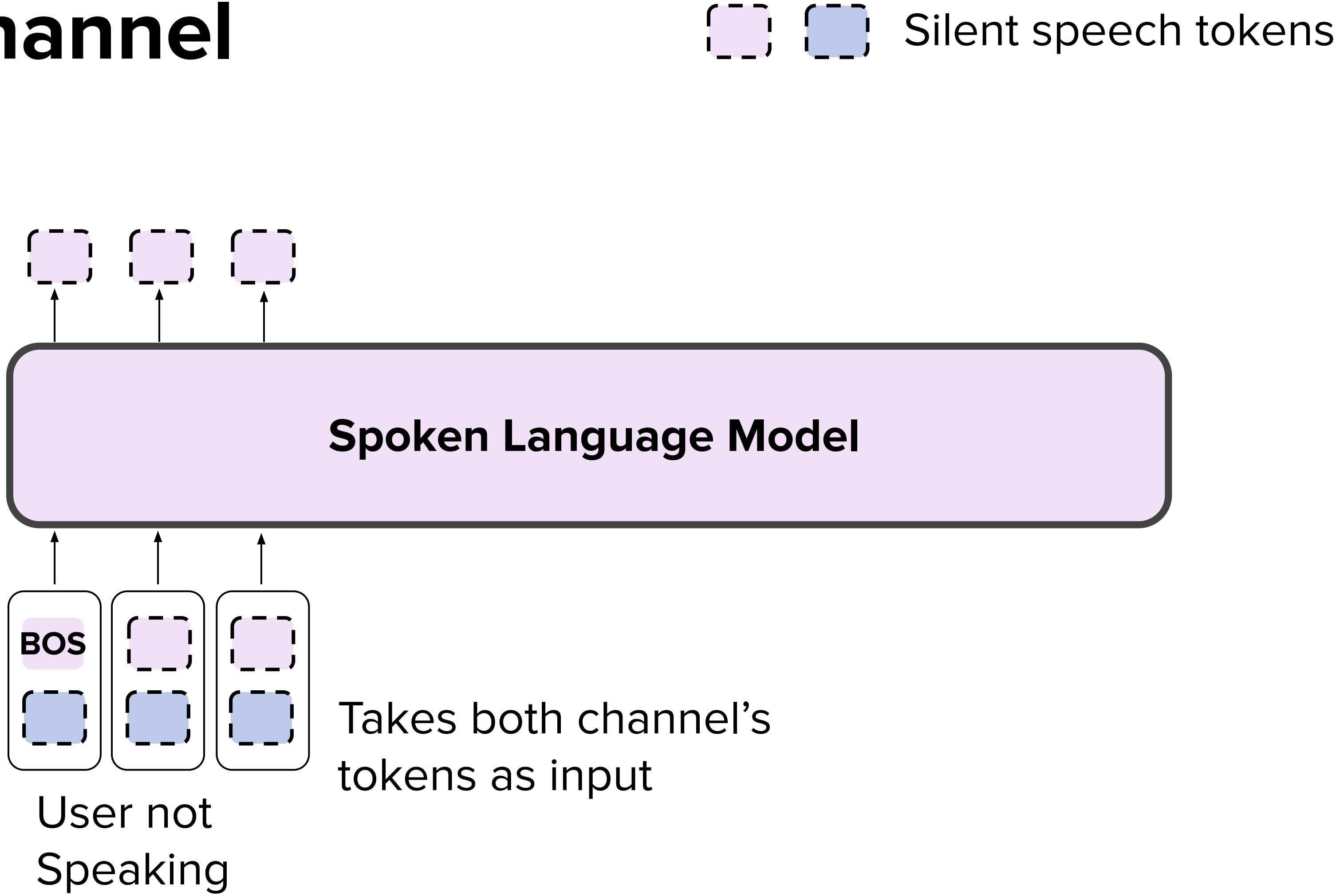
Dual-Channel



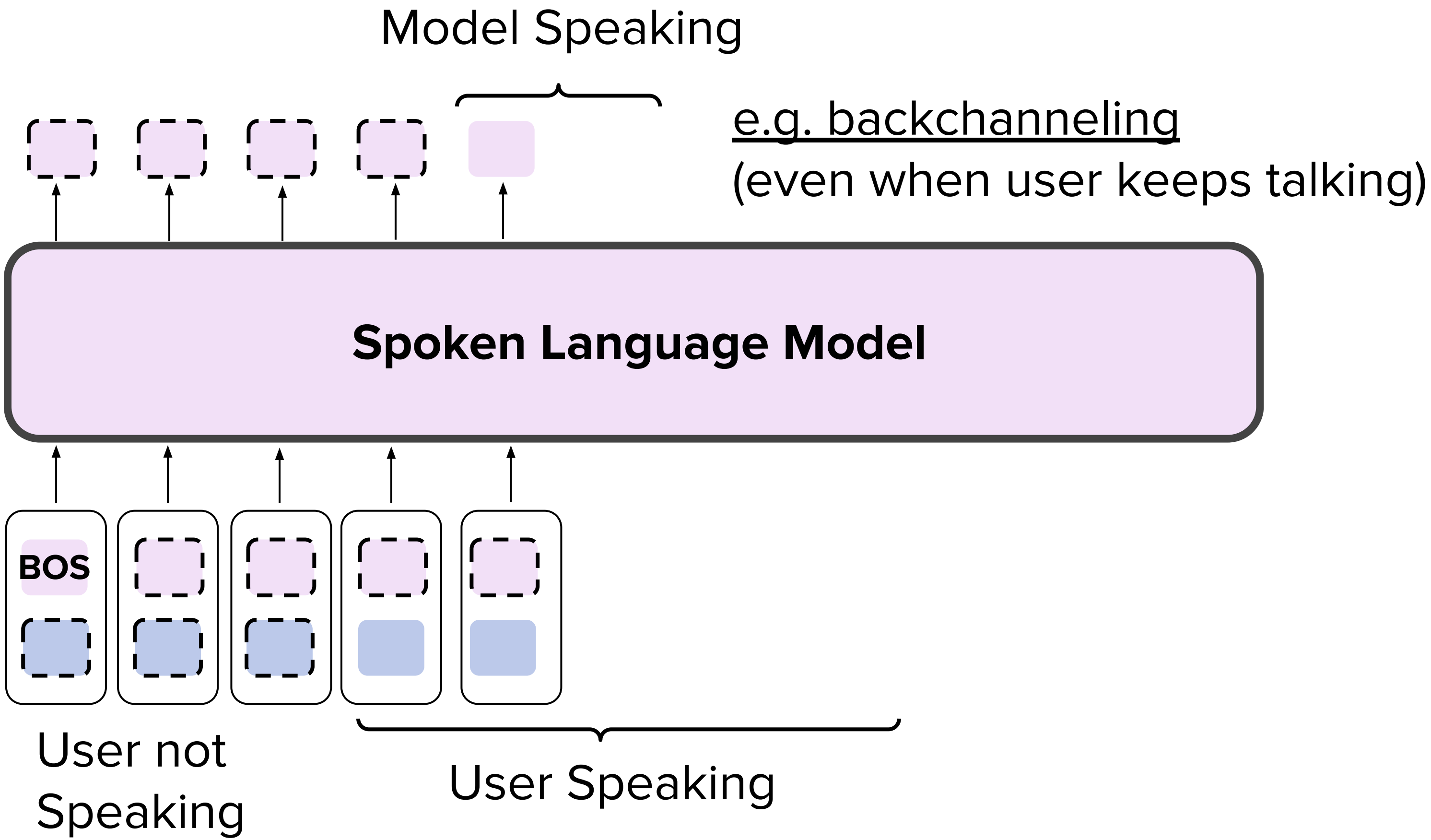
Dual-Channel



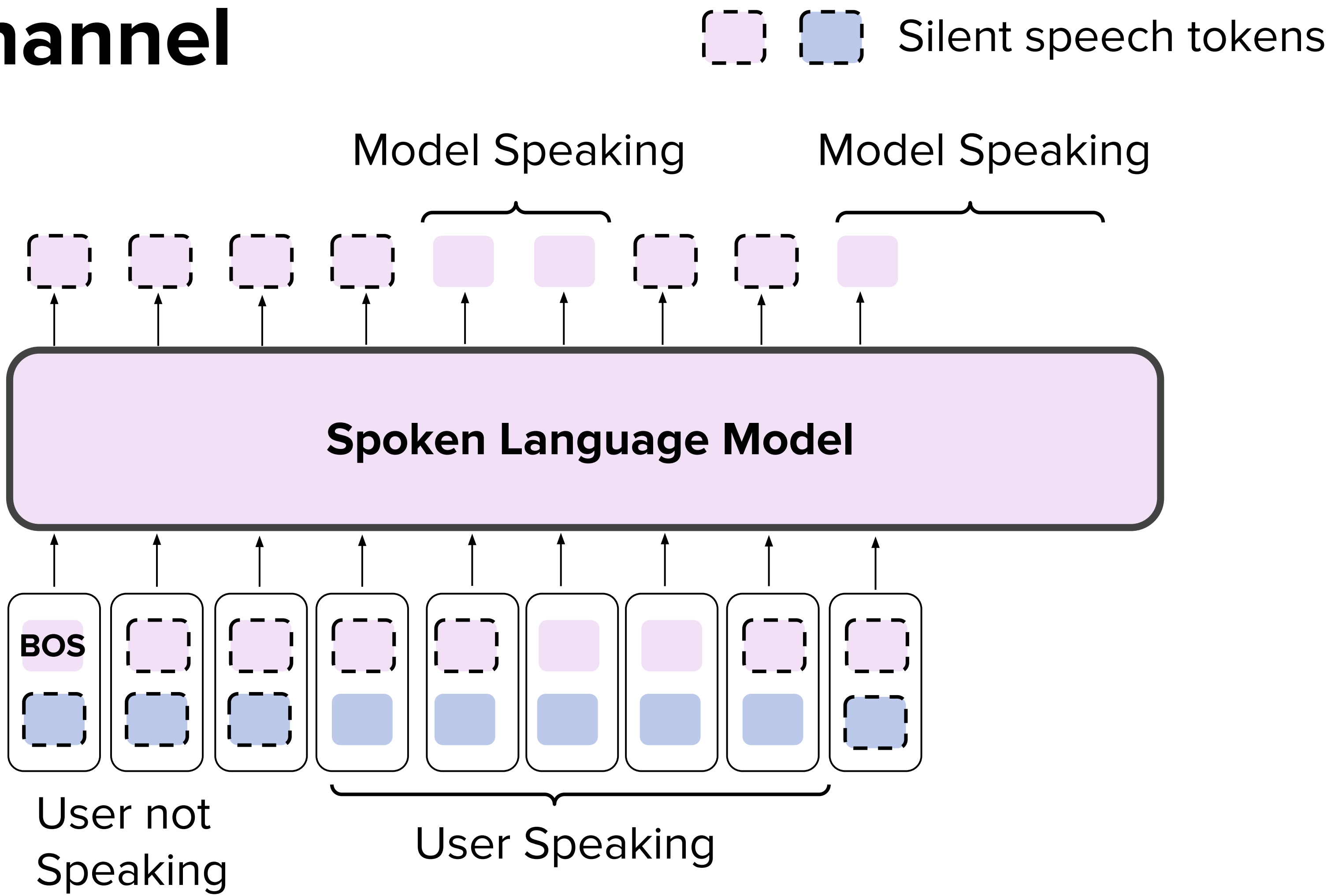
Dual-Channel



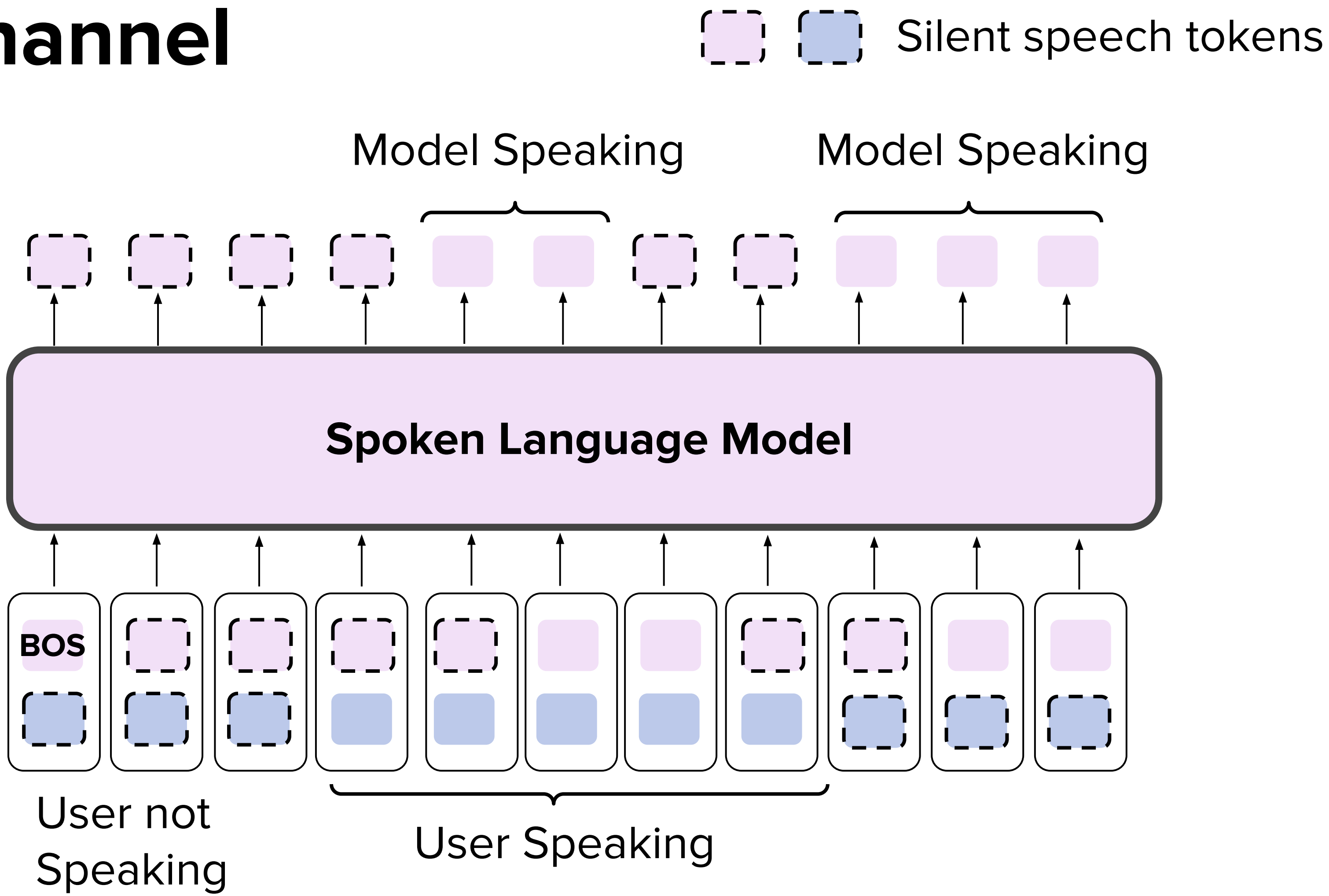
Dual-Channel

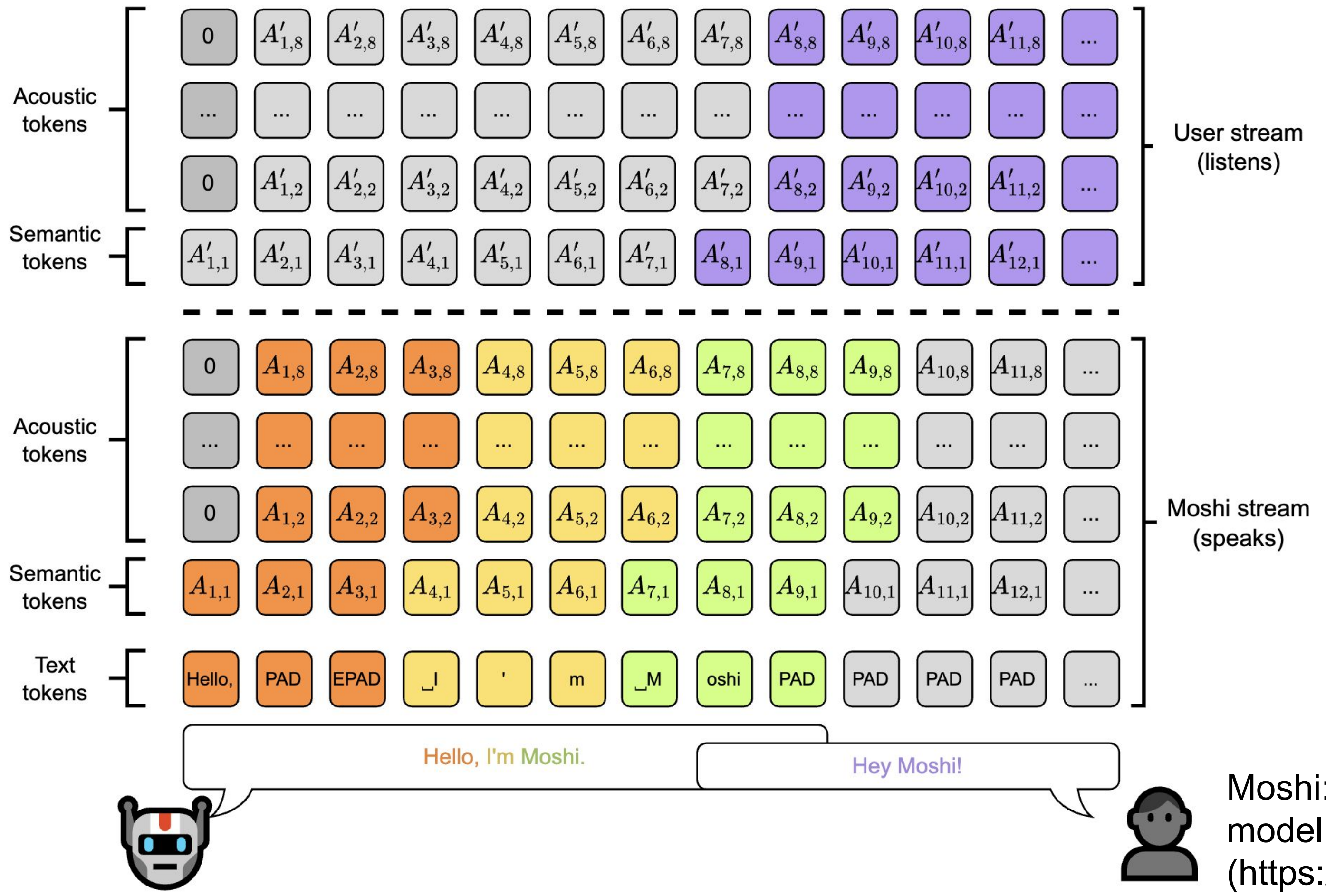


Dual-Channel



Dual-Channel



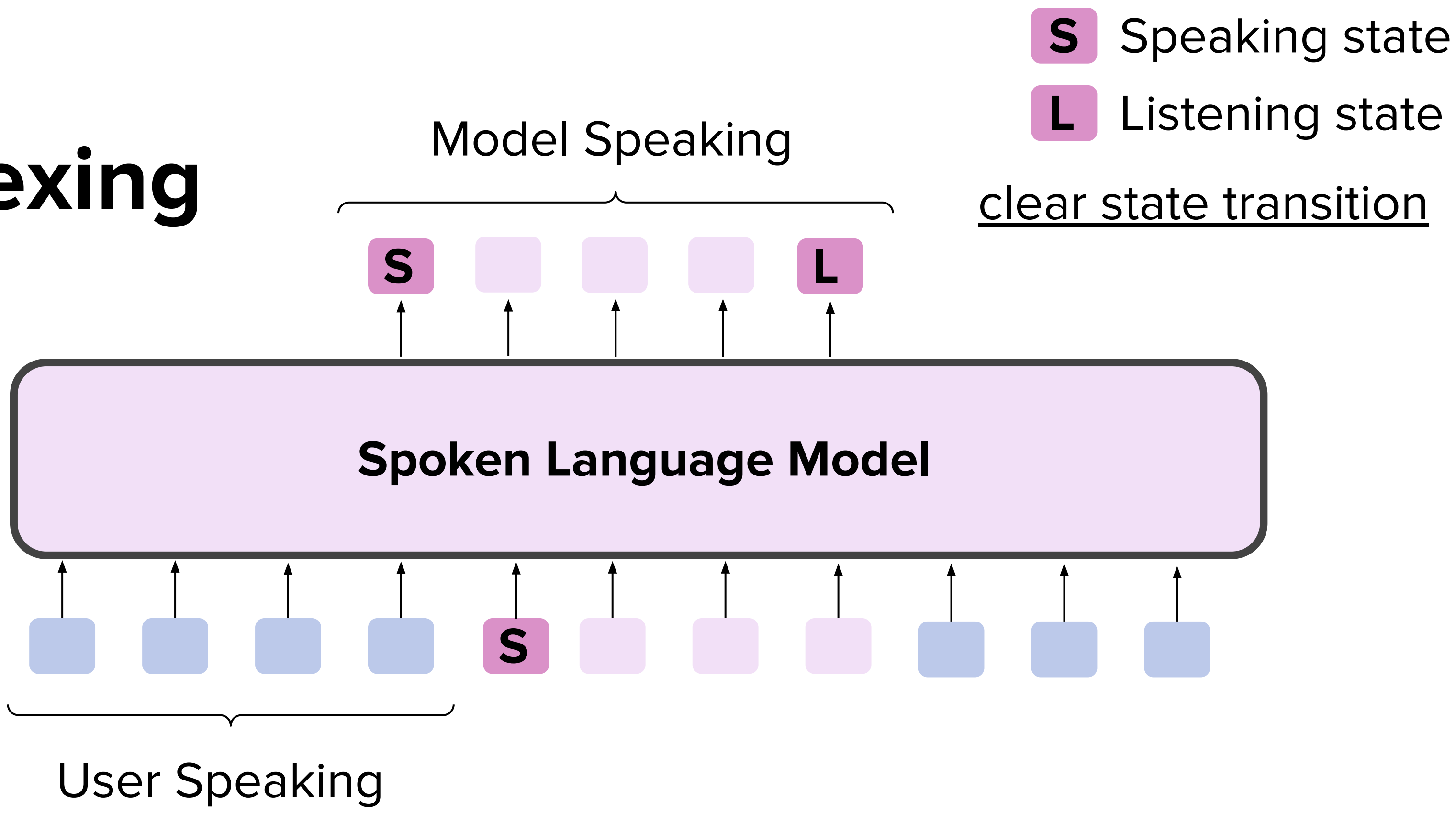


User Channel
8 audio codes

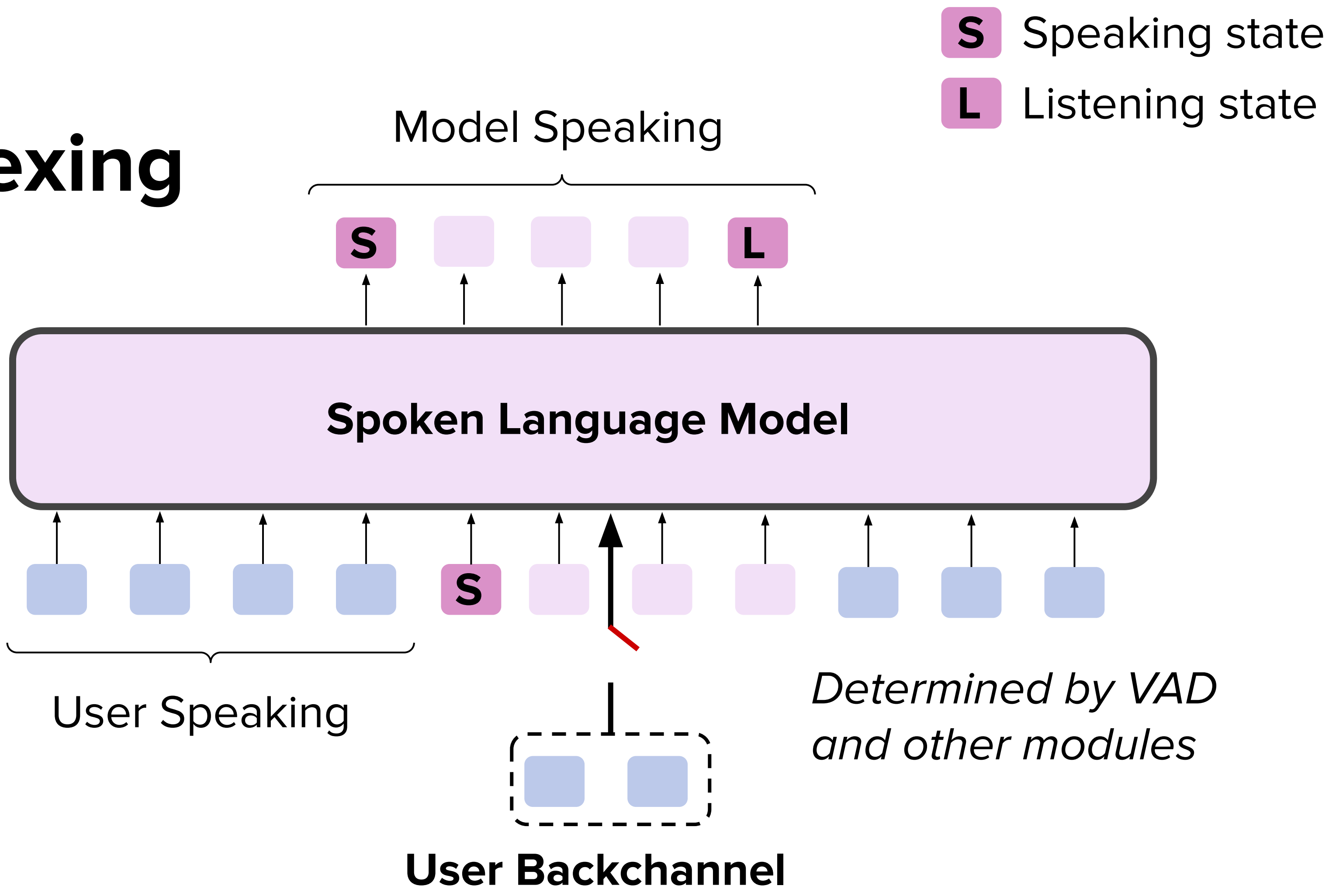
Moshi Channel
8 audio codes
+ 1 text token

Moshi: a speech-text foundation model for real-time dialogue (https://arxiv.org/abs/2410.00037)

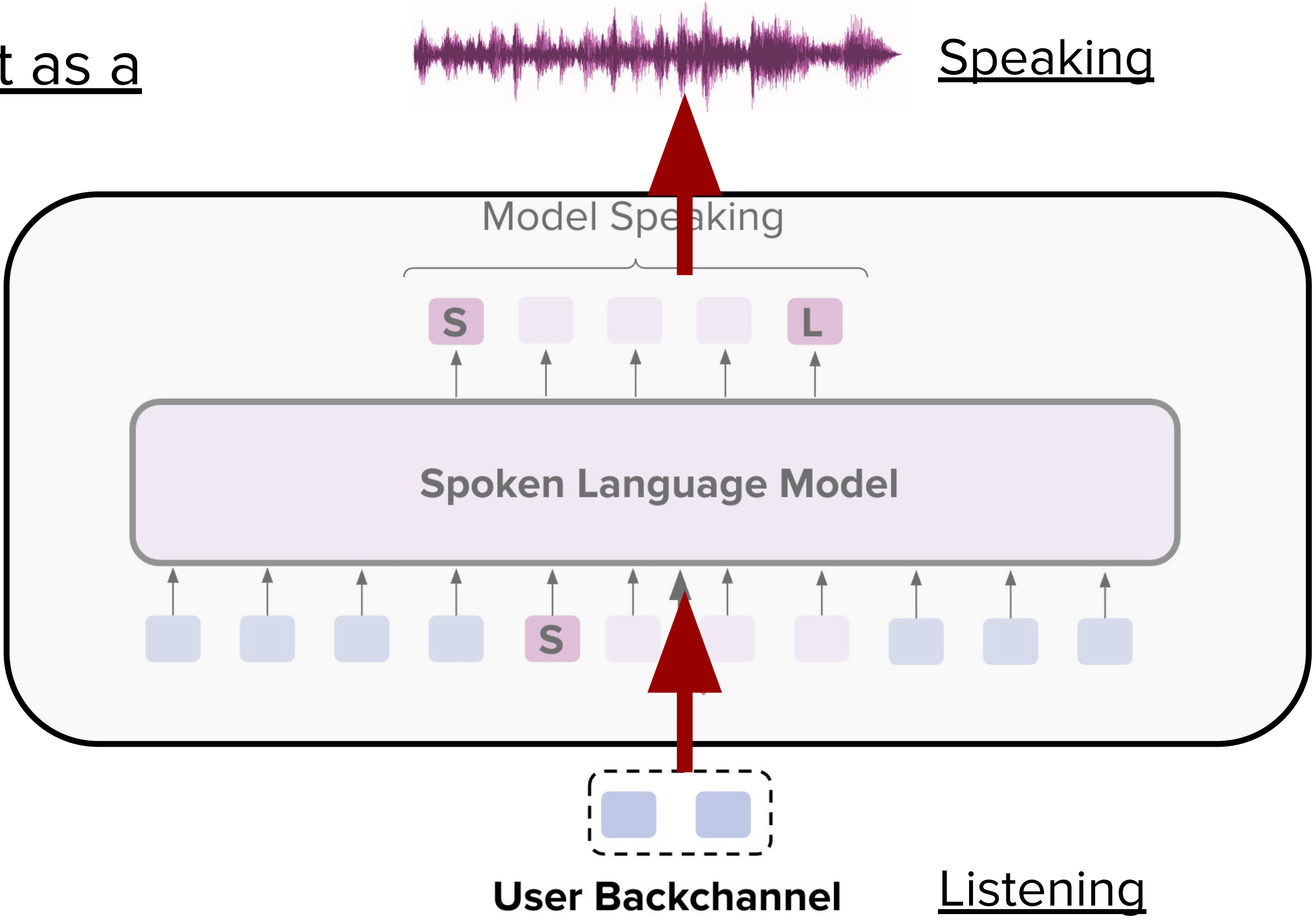
Time-Multiplexing



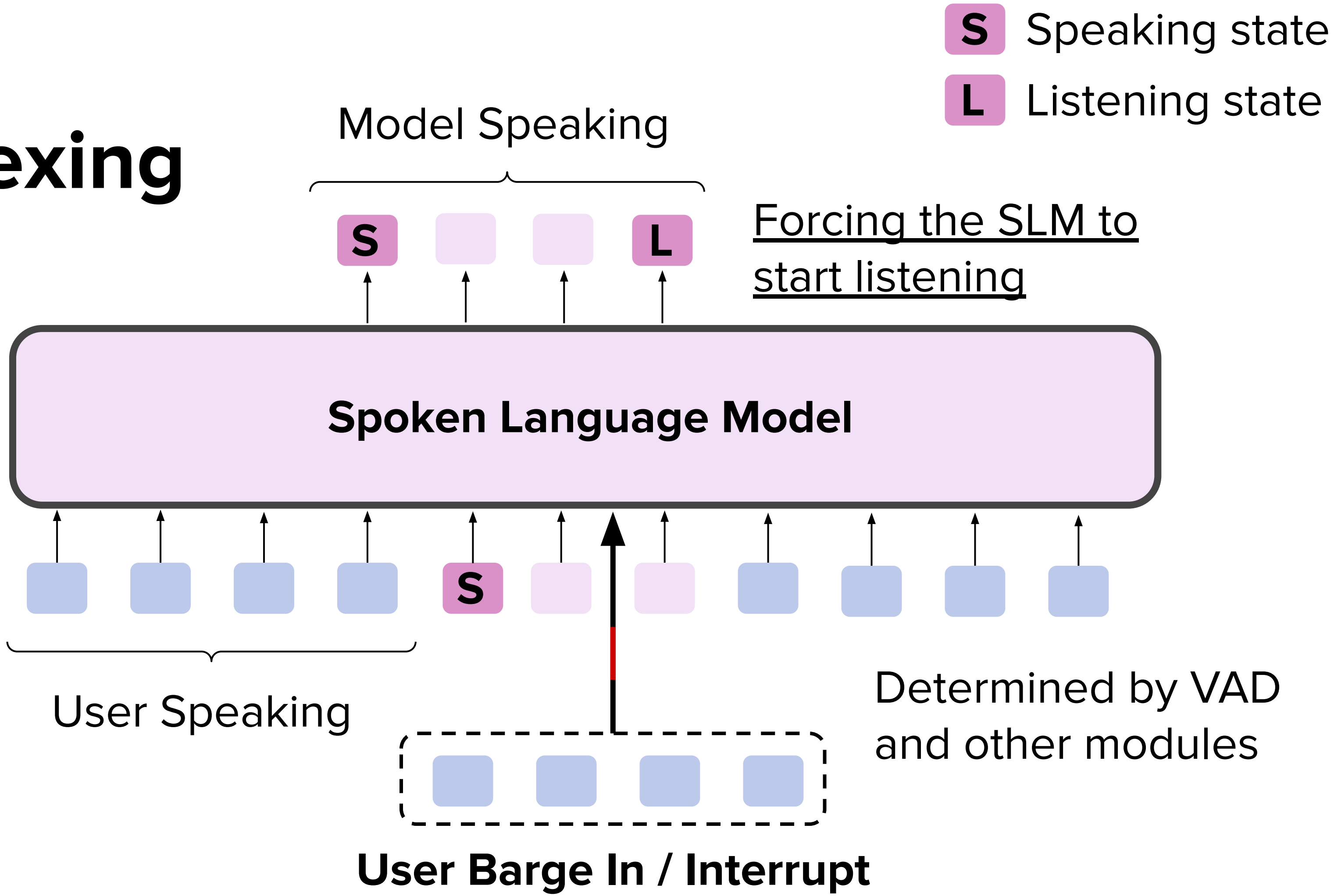
Time-Multiplexing



If we see it as a system



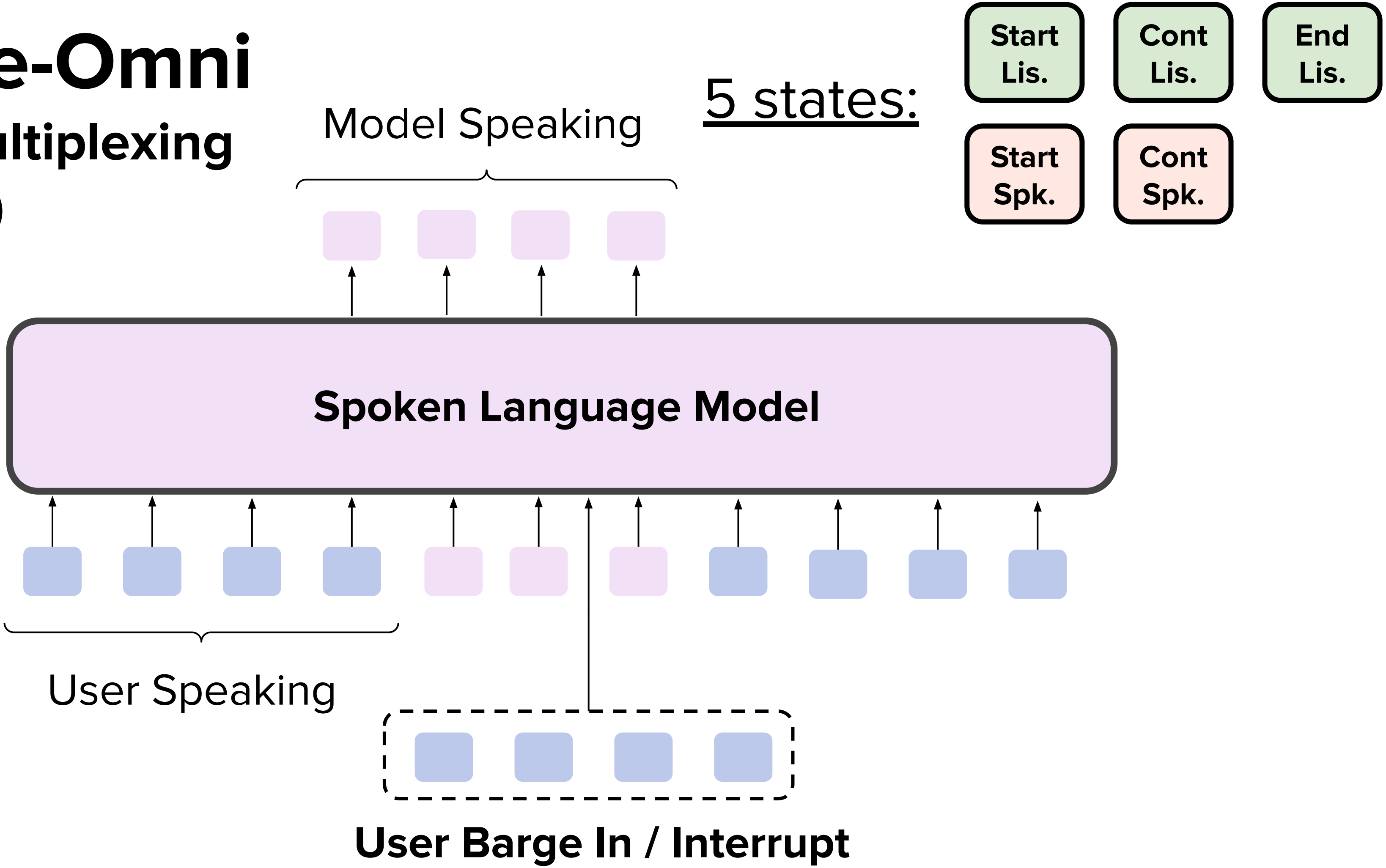
Time-Multiplexing



Freeze-Omni

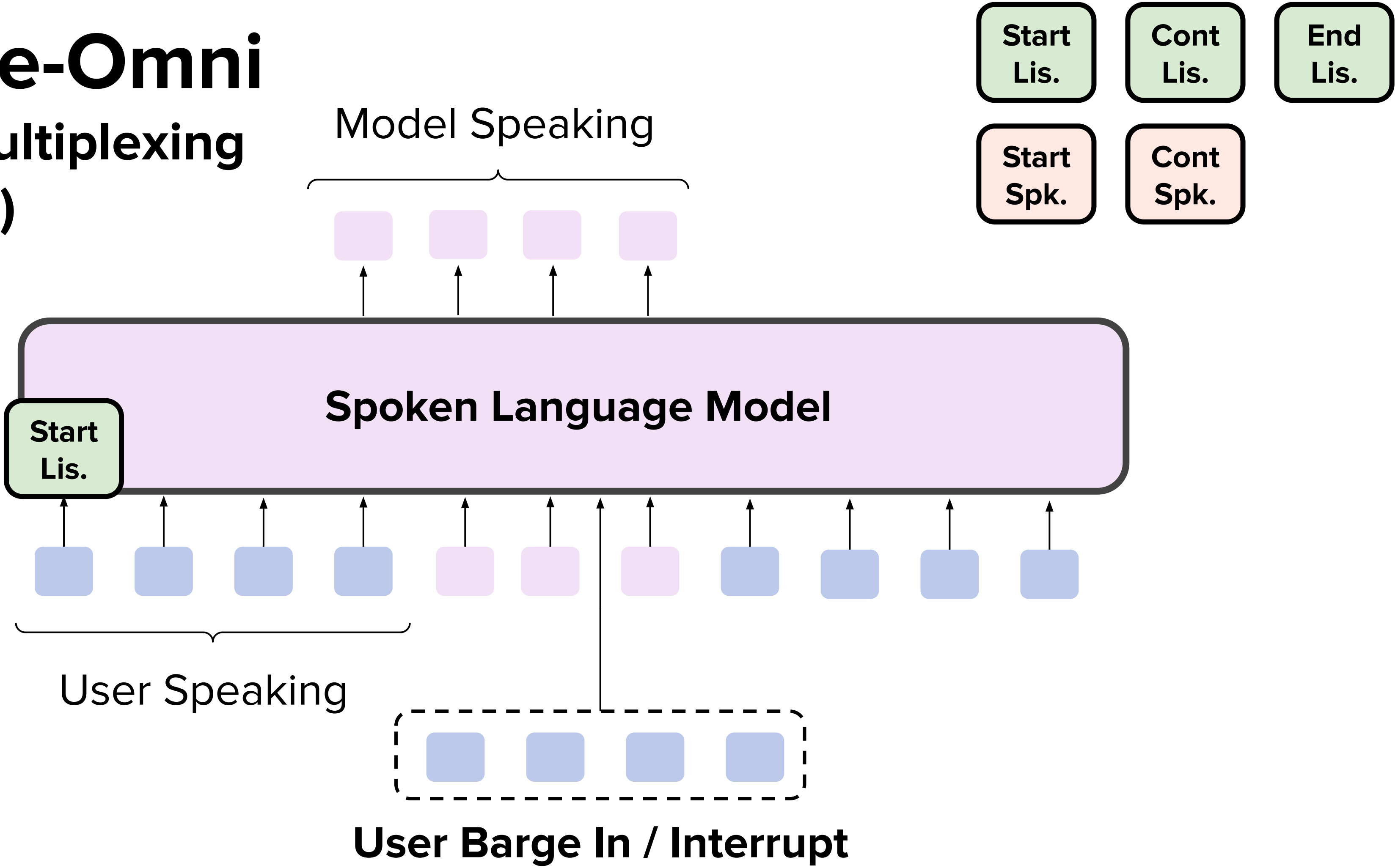
(Time Multiplexing example)

5 states:



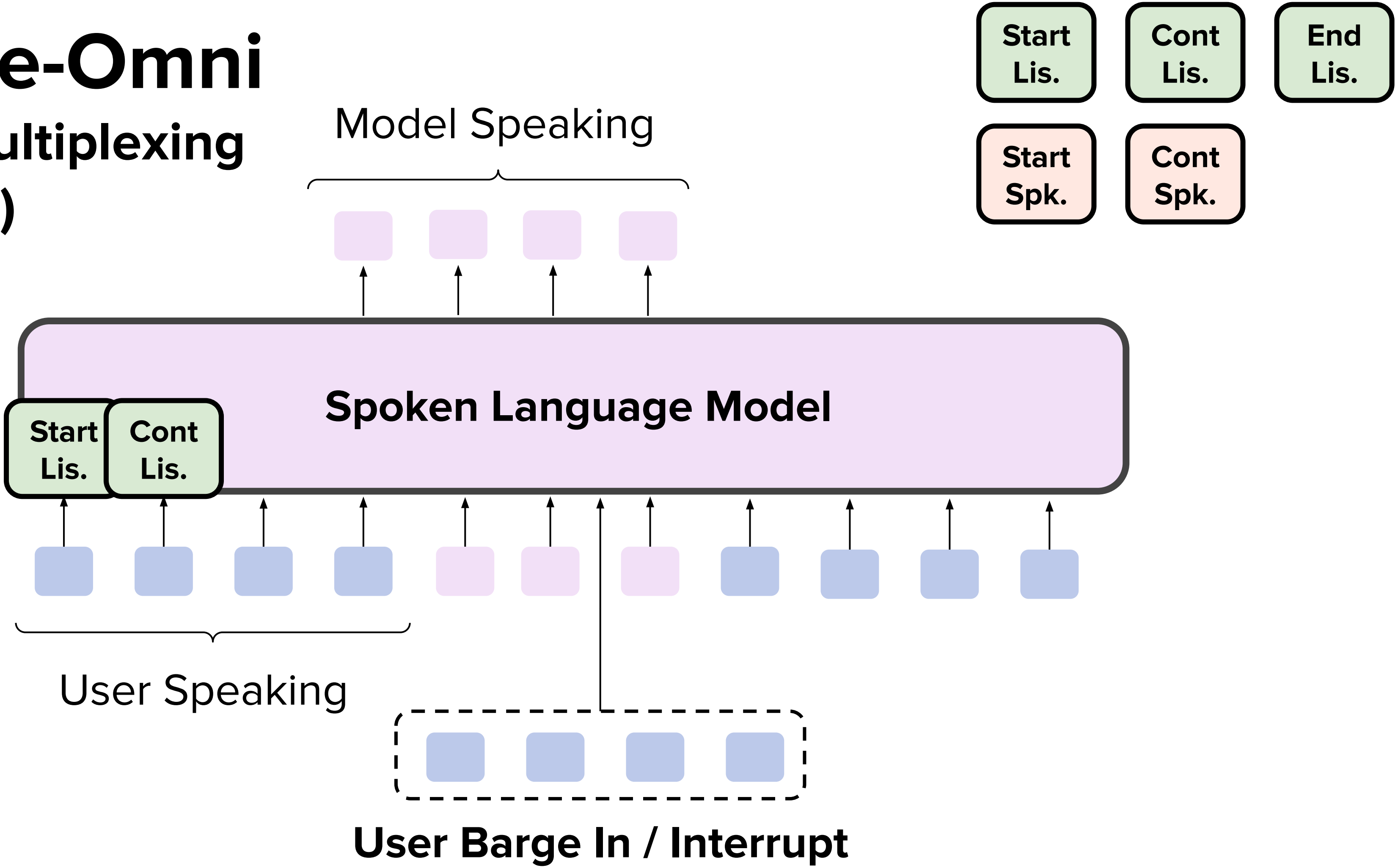
Freeze-Omni

(Time Multiplexing example)



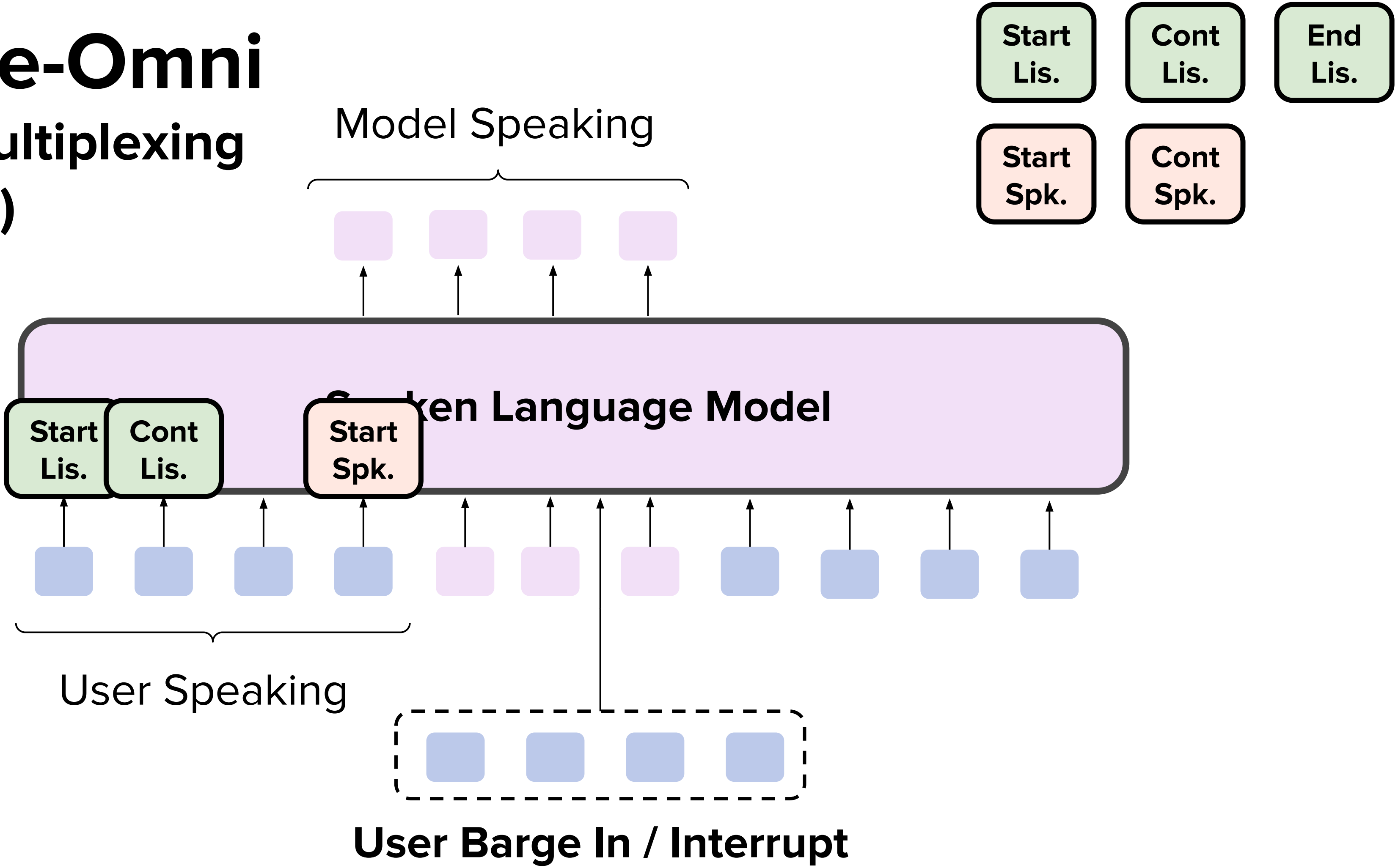
Freeze-Omni

(Time Multiplexing example)

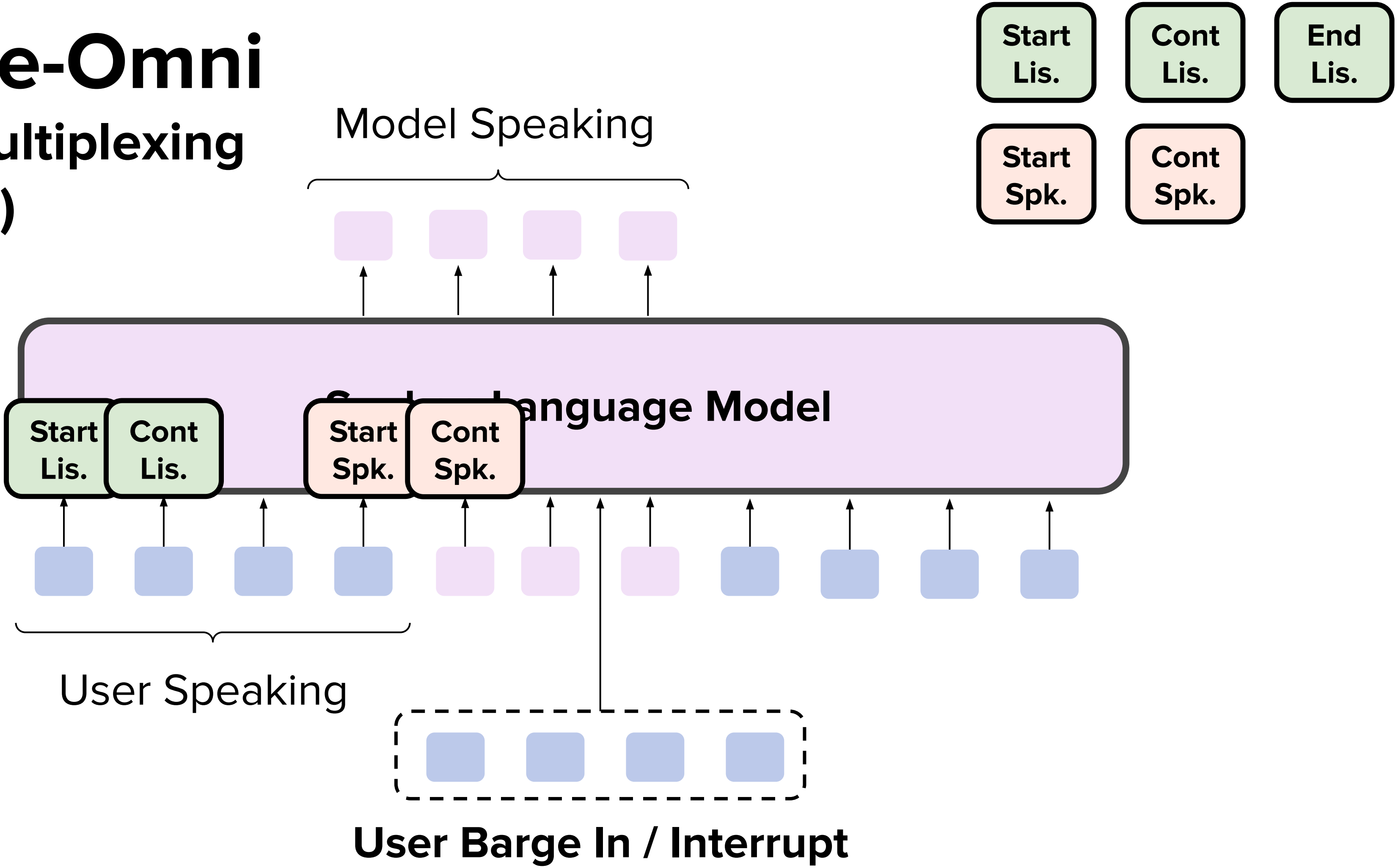


Freeze-Omni

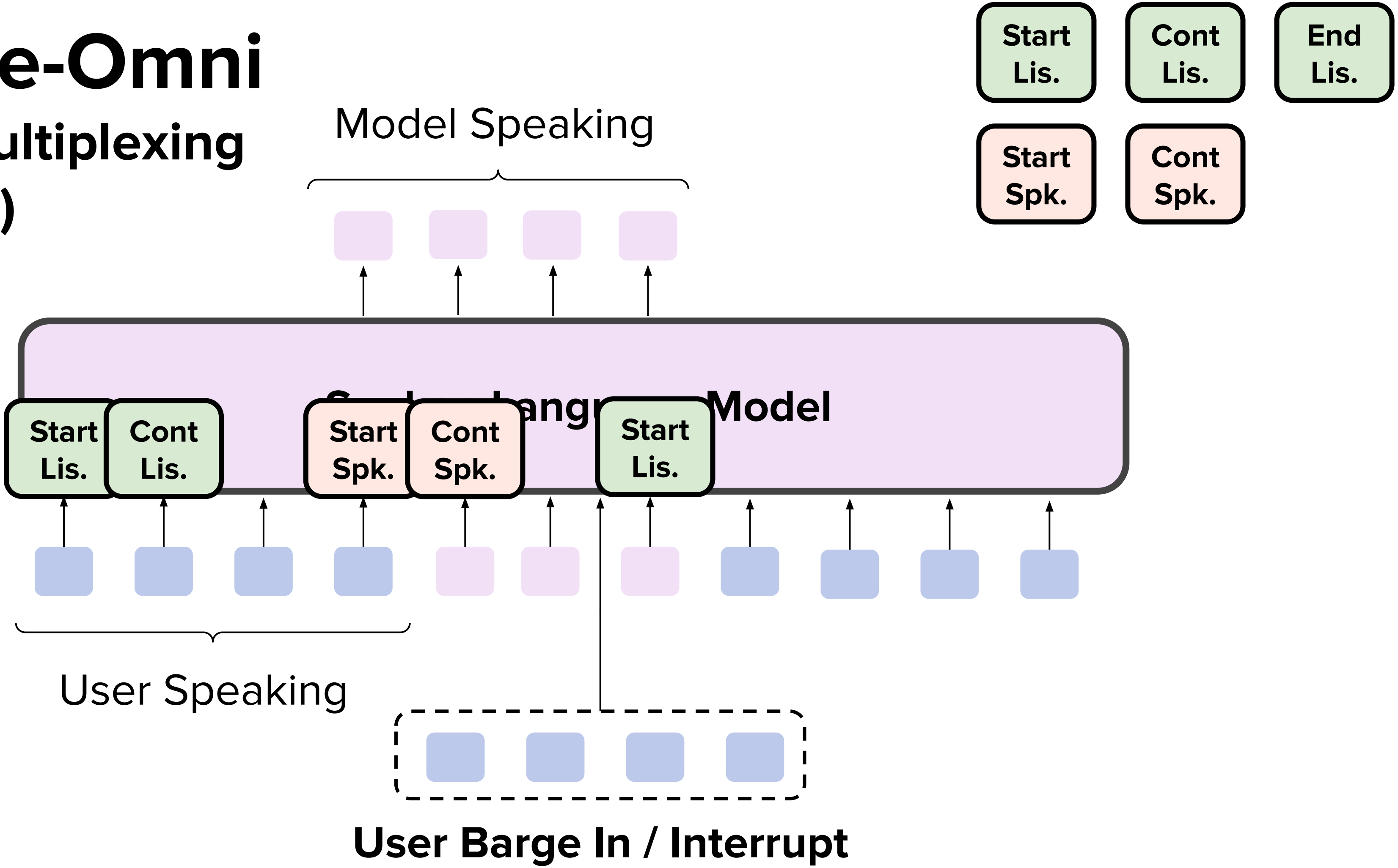
(Time Multiplexing example)



Freeze-Omni
(Time Multiplexing example)

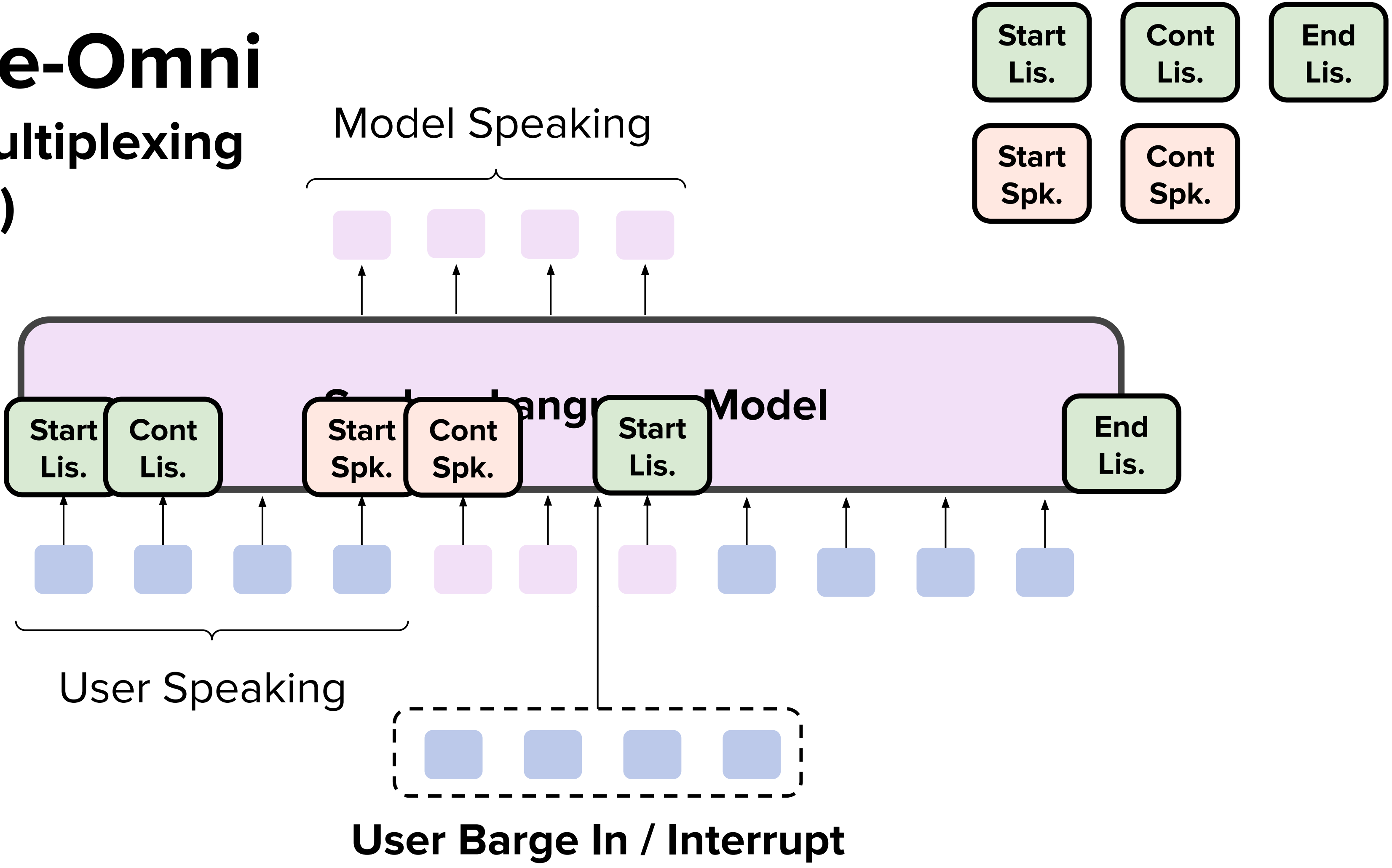


Freeze-Omni
(Time Multiplexing example)

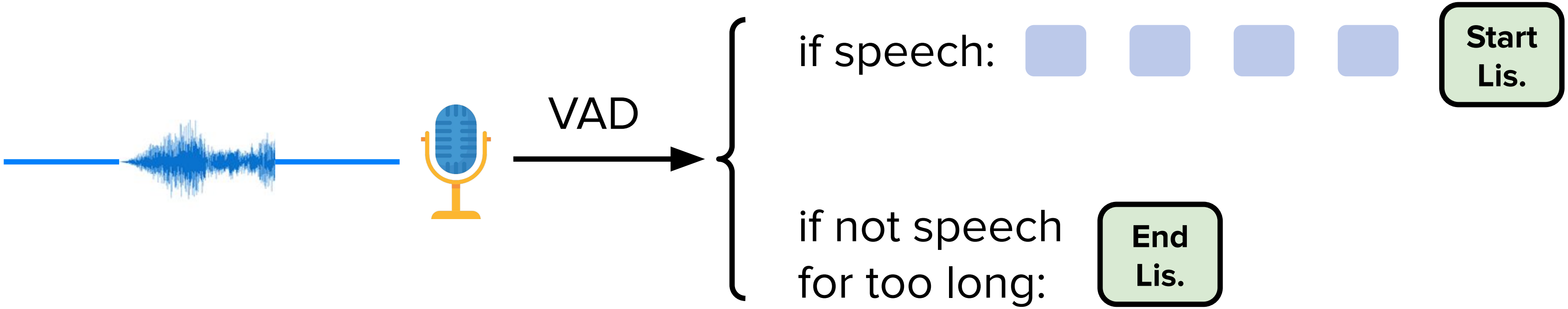


Freeze-Omni

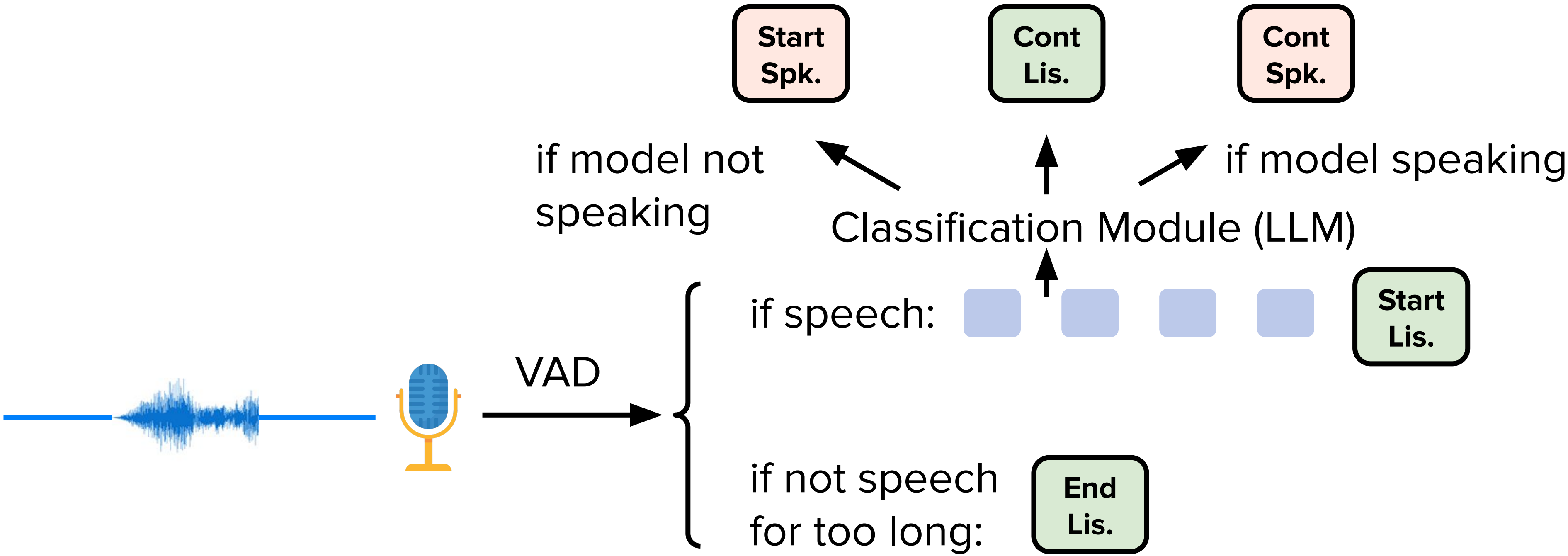
(Time Multiplexing example)

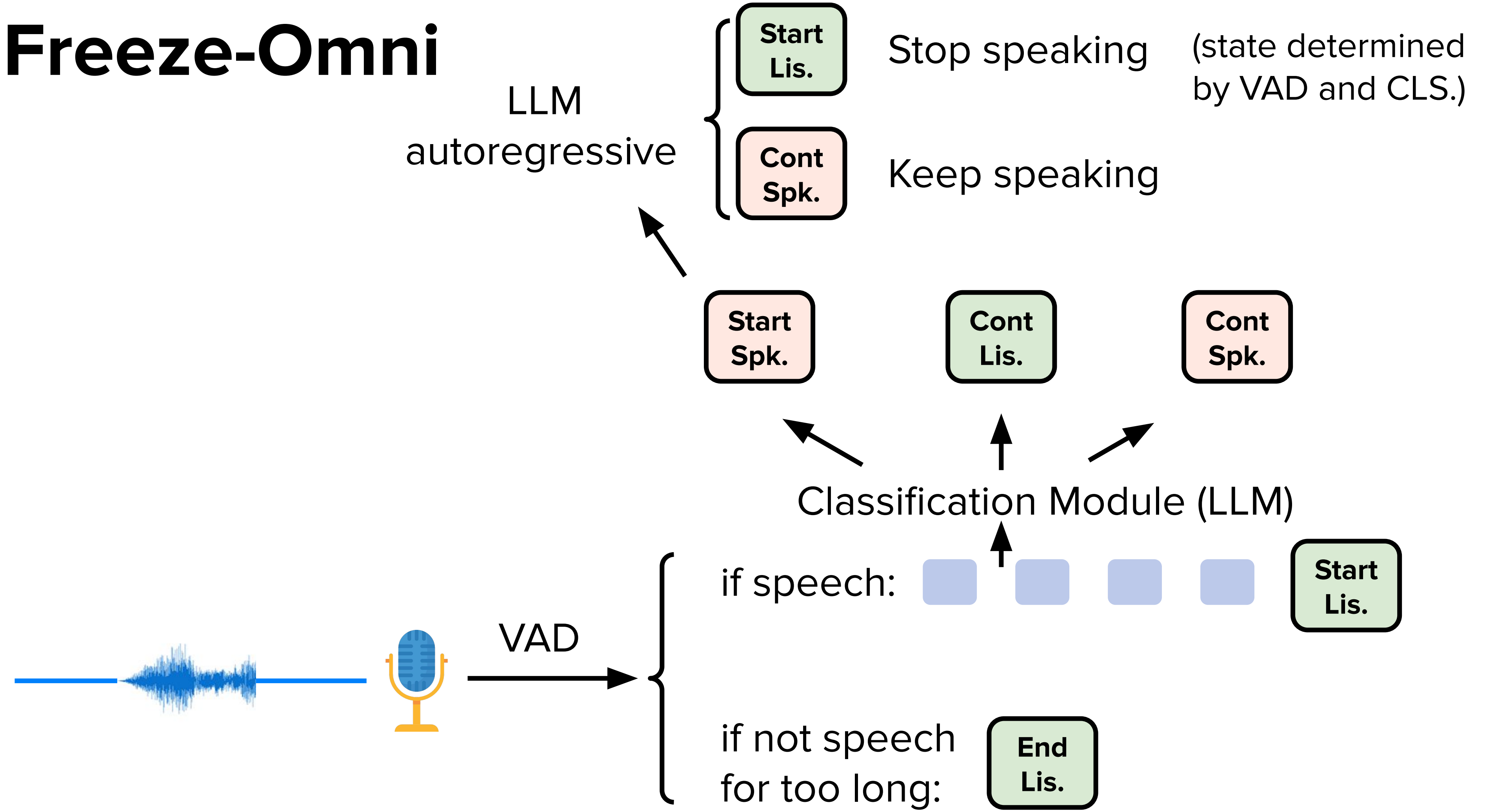


Freeze-Omni

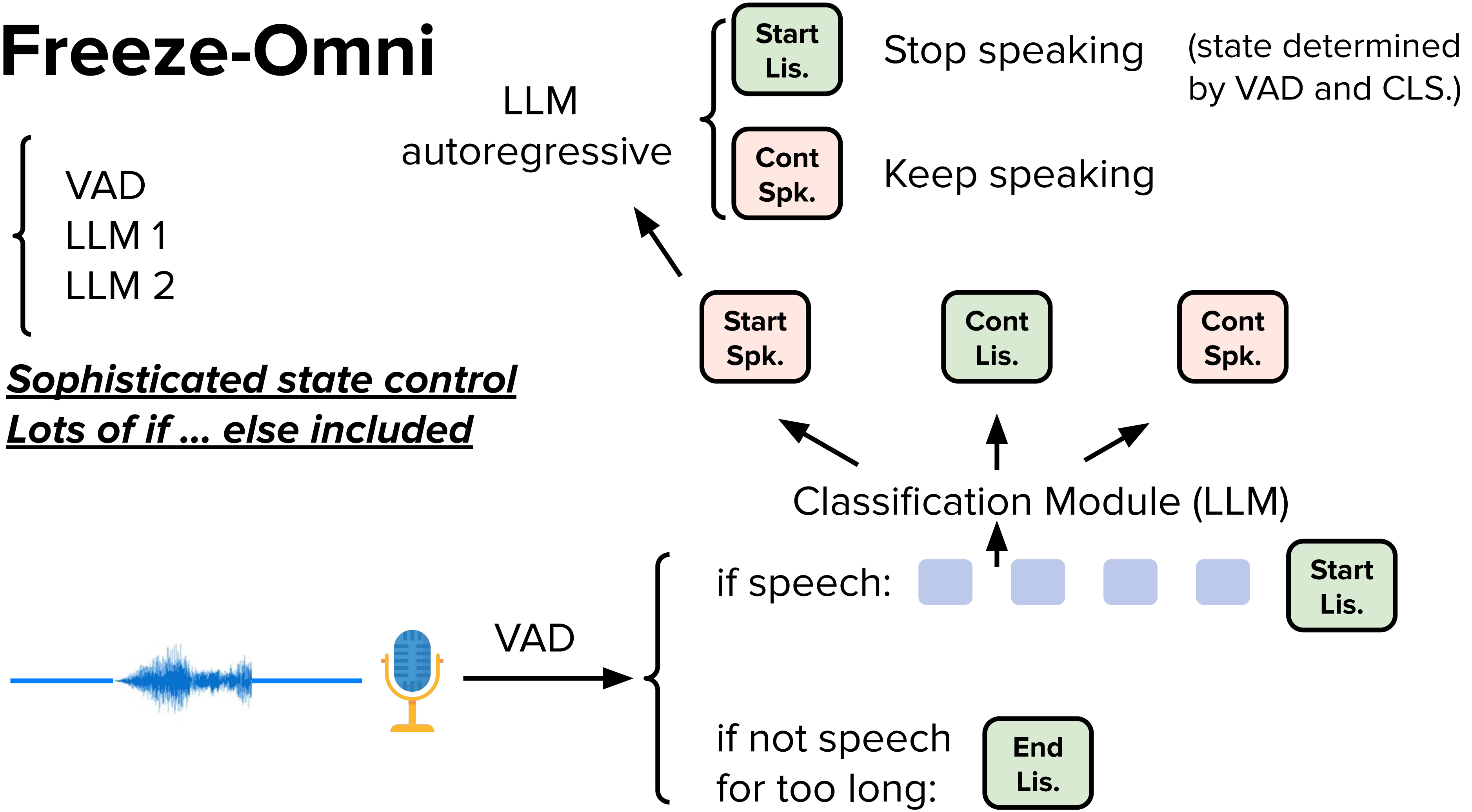


Freeze-Omni

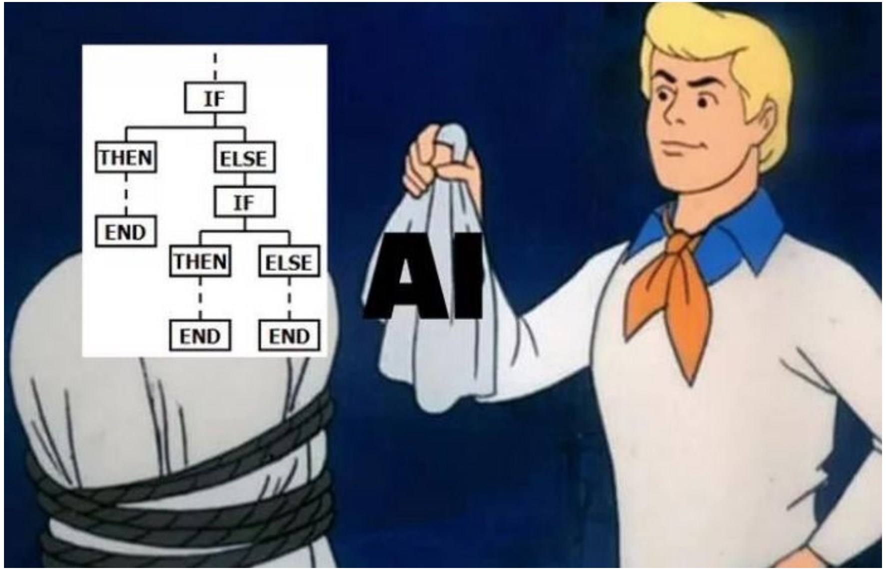

Start
Spk.
Cont
Lis.
Cont
Spk.
if model not
speaking
if model speaking
Classification Module (LLM)
if speech:
Start
Lis.
VAD
if not speech
for too long:
End
Lis.

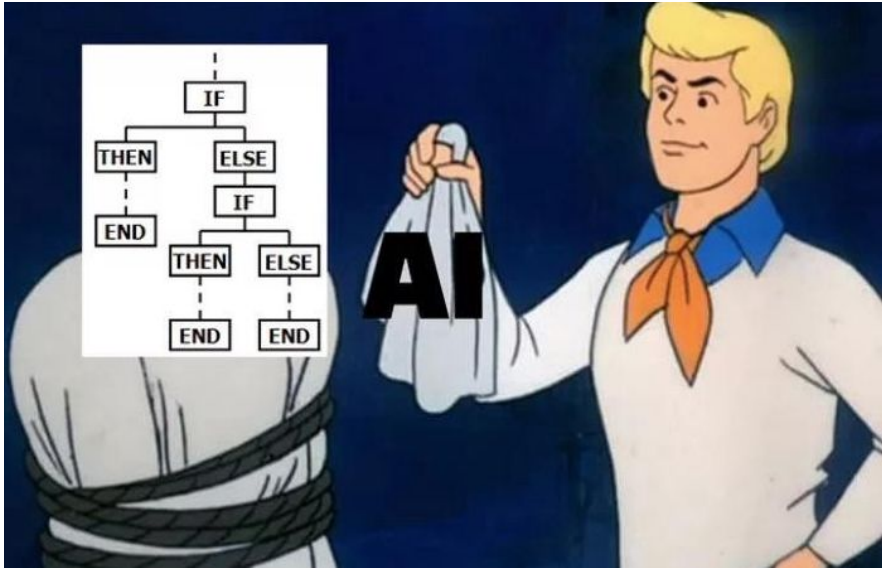

Freeze-Omni
LLM
autoregressive
Start
Lis.
Stop speaking
(state determined
by VAD and CLS.)
Cont
Spk.
Keep speaking
Start
Spk.
Cont
Lis.
Cont
Spk.
Classification Module (LLM)
if speech:
Start
Lis.
VAD
if not speech
for too long:
End
Lis.

Freeze-Omni

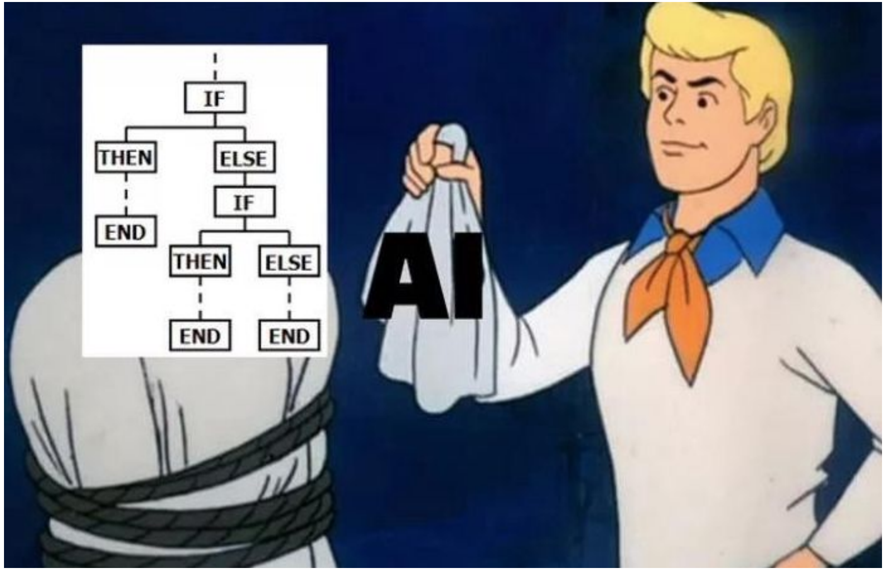


VAD
LLM 1
LLM 2
Sophisticated state control
Lots of if … else included
LLM
autoregressive
Start
Lis.
Stop speaking
(state determined
by VAD and CLS.)
Cont
Spk.
Keep speaking
Start
Spk.
Cont
Lis.
Cont
Spk.
Classification Module (LLM)
if speech:
Start
Lis.
VAD
if not speech
for too long:
End
Lis.


IF
THEN
ELSE
IF
END
THEN
ELSE
END
END
AI



Russian nesting doll

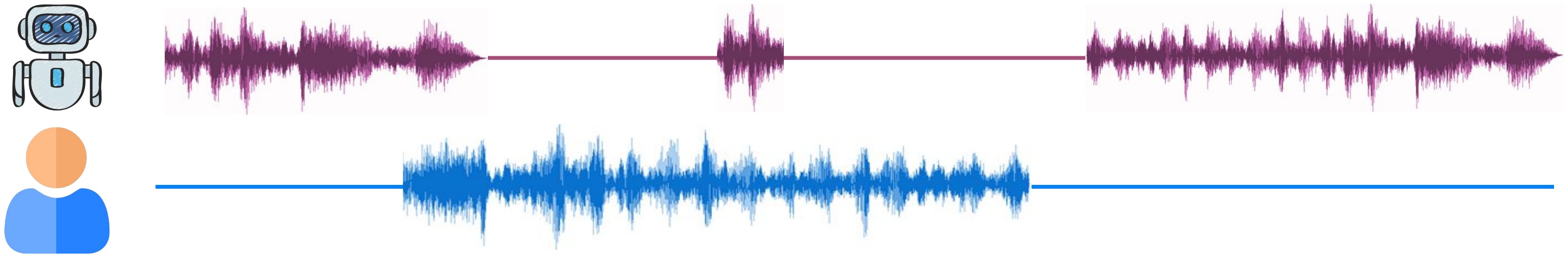


We are not just using an AI model
We are using an AI *system*

Russian nesting doll

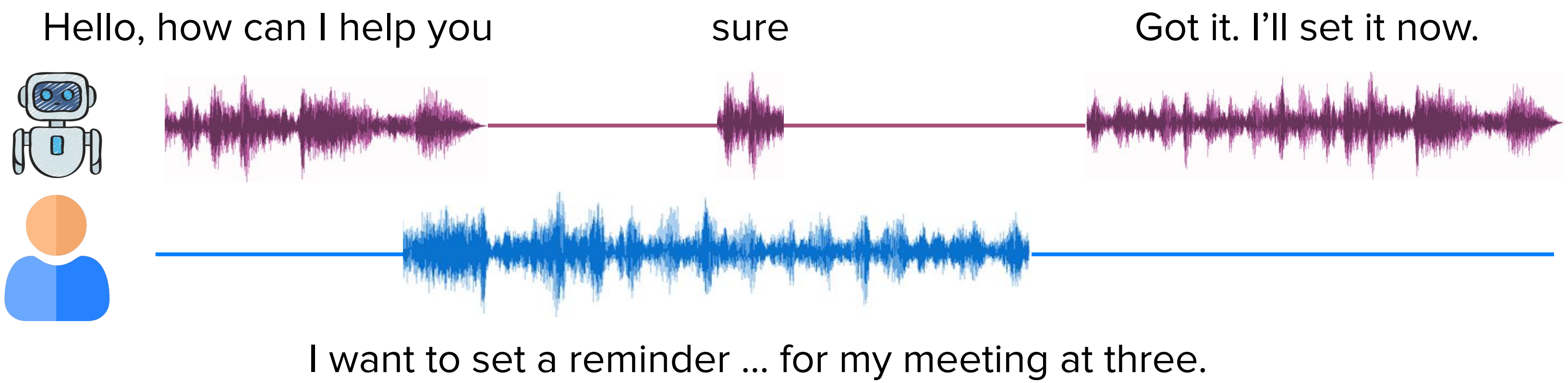
How to evaluate Full-duplex SLM?

Hello, how can I help you sure Got it. I’ll set it now.



I want to set a reminder … for my meeting at three.

How to evaluate Full-duplex SLM?

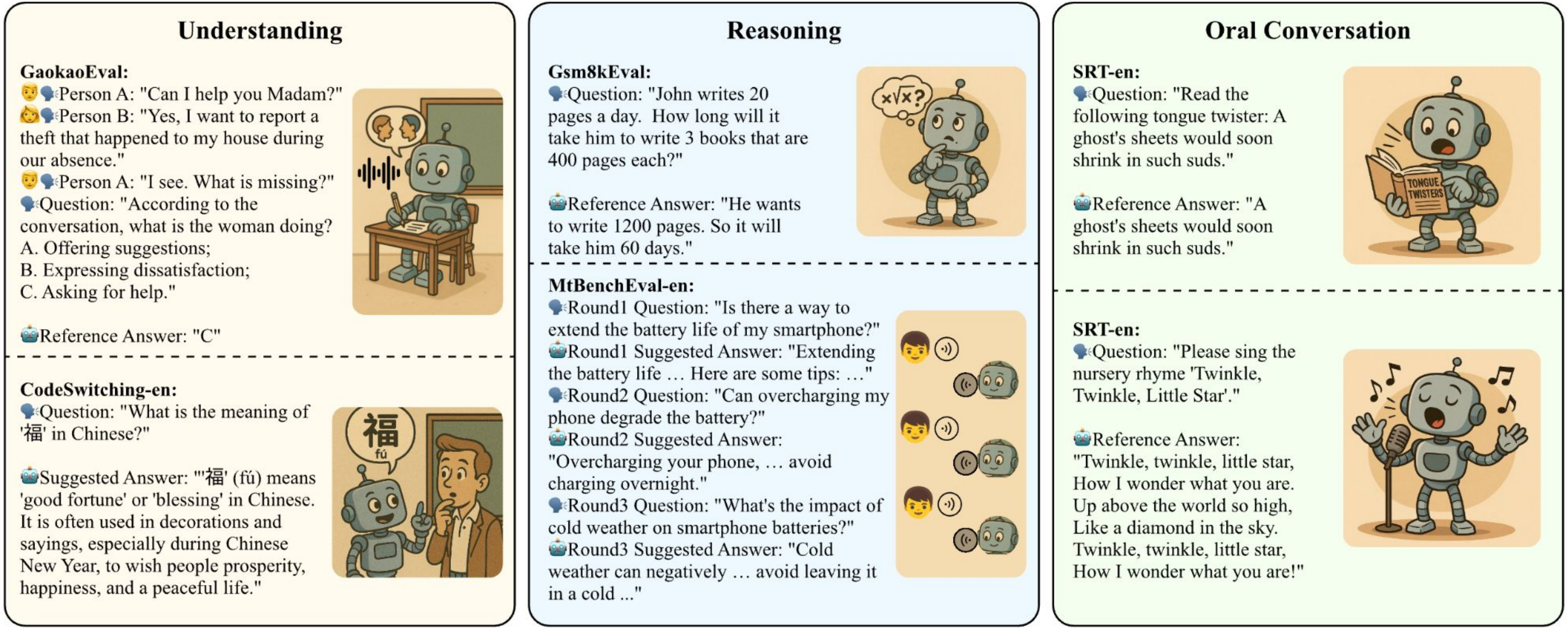


- Speech Understanding
- Paralingustic
- Reasoning
- Speech quality
- …
- Full-duplex behavior

→ Existing benchmark for S2S models (e.g. URO-Bench)

→ e.g. dGSLM, Full-duplex-Bench, Talking-Turn, Game-Time…

URO-Bench



Speech understanding
Simple QA

Math

Instruction following

URO-Bench: Towards Comprehensive Evaluation for End-to-End Spoken Dialogue Models https://aclanthology.org/2025.findings-emnlp.933/

Lang	Models	LLM Scale	Task Accomplish Scores					
			basic			pro		
			Understanding ↑	Reasoning ↑	Oral Conversation ↑	Understanding ↑	Reasoning ↑	Oral Conversation ↑
En	GLM-4-Voice	9B	**82.16**	**55.46**	**74.20**	**45.14**	**61.28**	**57.14**
	LLaMA-Omni	8B	47.45	36.03	64.98	28.85	47.62	42.96
	Freeze-Omni	7B	58.68	37.52	52.24	29.21	5.49	38.98
	Mini-Omni	0.5B	12.42	12.78	30.74	21.66	0	21.42
	Mini-Omni2	0.5B	16.27	15.60	33.98	24.43	0	24.53
	SLAM-Omni	0.5B	26.60	23.36	47.54	25.79	24.72	30.16
	GPT-4o-Audio-Preview	-	87.76	81.73	94.84	62.92	57.07	73.63
	Whisper + GLM-4-9B-Chat-HF	9B	90.83	76.29	88.26	-	-	-
	Whisper + Llama-3.1-8B-Instruct	8B	50.35	75.14	88.72	-	-	-
	Whisper + Qwen2-7B-Instruct	7B	64.99	77.94	91.52	-	-	-
	Whisper + Qwen2-0.5B-Instruct	0.5B	46.35	44.41	60.13	-	-	-
	Whisper + GPT-4o	-	92.62	82.91	94.60	-	-	-

Basic tasks: focus on "content"

Table from URO-Bench

Lang	Models	LLM Scale	Task Accomplish Scores					
			basic			pro		
			Understanding ↑	Reasoning ↑	Oral Conversation ↑	Understanding ↑	Reasoning ↑	Oral Conversation ↑
En	GLM-4-Voice	9B	**82.16**	**55.46**	**74.20**	**45.14**	**61.28**	**57.14**
	LLaMA-Omni	8B	47.45	36.03	64.98	28.85	47.62	42.96
	Freeze-Omni	7B	58.68	37.52	52.24	29.21	5.49	38.98
	Mini-Omni	0.5B	12.42	12.78	30.74	21.66	0	21.42
	Mini-Omni2	0.5B	16.27	15.60	33.98	24.43	0	24.53
	SLAM-Omni	0.5B	26.60	23.36	47.54	25.79	24.72	30.16
	GPT-4o-Audio-Preview	-	87.76	81.73	94.84	62.92	57.07	73.63
	Whisper + GLM-4-9B-Chat-HF	9B	90.83	76.29	88.26	-	-	-
	Whisper + Llama-3.1-8B-Instruct	8B	50.35	75.14	88.72	-	-	-
	Whisper + Qwen2-7B-Instruct	7B	64.99	77.94	91.52	-	-	-
	Whisper + Qwen2-0.5B-Instruct	0.5B	46.35	44.41	60.13	-	-	-
	Whisper + GPT-4o	-	92.62	82.91	94.60	-	-	-

cascaded system with text LLM

Basic tasks: focus on "content"

Pro tasks: Includes paralingustic, music, environment sound, ...

Table from URO-Bench

Lang	Models	LLM Scale	Task Accomplish Scores					
			basic			pro		
			Understanding ↑	Reasoning ↑	Oral Conversation ↑	Understanding ↑	Reasoning ↑	Oral Conversation ↑
En	GLM-4-Voice	9B	**82.16**	**55.46**	**74.20**	**45.14**	**61.28**	**57.14**
	LLaMA-Omni	8B	47.45	36.03	<u>64.98</u>	28.85	<u>47.62</u>	<u>42.96</u>
	Freeze-Omni	7B	<u>58.68</u>	<u>37.52</u>	52.24	<u>29.21</u>	5.49	38.98
	Mini-Omni	0.5B	12.42	12.78	30.74	21.66	0	21.42
	Mini-Omni2	0.5B	16.27	15.60	33.98	24.43	0	24.53
	SLAM-Omni	0.5B	26.60	23.36	47.54	25.79	24.72	30.16
	GPT-4o-Audio-Preview	-	87.76	81.73	94.84	62.92	57.07	73.63
	Whisper + GLM-4-9B-Chat-HF	9B	90.83	76.29	88.26	-	-	-
	Whisper + Llama-3.1-8B-Instruct	8B	50.35	75.14	88.72	-	-	-
	Whisper + Qwen2-7B-Instruct	7B	64.99	77.94	91.52	-	-	-
	Whisper + Qwen2-0.5B-Instruct	0.5B	46.35	44.41	60.13	-	-	-
	Whisper + GPT-4o	-	92.62	82.91	94.60	-	-	-

<u>Basic tasks:</u> focus on "content"

<u>Pro tasks:</u> Includes paralingustic, music, environment sound, ...

Table from URO-Bench

Lang	Models	LLM Scale	Task Accomplish Scores					
			basic			pro		
			Understanding ↑	Reasoning ↑	Oral Conversation ↑	Understanding ↑	Reasoning ↑	Oral Conversation ↑
En	GLM-4-Voice	9B	**82.16**	**55.46**	**74.20**	**45.14**	**61.28**	**57.14**
	LLaMA-Omni	8B	47.45	36.03	<u>64.98</u>	28.85	<u>47.62</u>	<u>42.96</u>
	Freeze-Omni	7B	<u>58.68</u>	<u>37.52</u>	52.24	<u>29.21</u>	5.49	38.98
	Mini-Omni	0.5B	12.42	12.78	30.74	21.66	0	21.42
	Mini-Omni2	0.5B	16.27	15.60	33.98	24.43	0	24.53
	SLAM-Omni	0.5B	26.60	23.36	47.54	25.79	24.72	30.16
	GPT-4o-Audio-Preview	-	87.76	81.73	94.84	62.92	57.07	73.63
	Whisper + GLM-4-9B-Chat-HF	9B	90.83	76.29	88.26	-	-	-
	Whisper + Llama-3.1-8B-Instruct	8B	50.35	75.14	88.72	-	-	-
	Whisper + Qwen2-7B-Instruct	7B	64.99	77.94	91.52	-	-	-
	Whisper + Qwen2-0.5B-Instruct	0.5B	46.35	44.41	60.13	-	-	-
	Whisper + GPT-4o	-	92.62	82.91	94.60	-	-	-

cascaded system with text LLM

GLM-4-Voice performs best but no a full-duplex model

GLM-4-Voice LLaMA-Omni Freeze-Omni

<u>Full-duplex</u>

Table from URO-Bench

Lang	Models	LLM Scale	Task Accomplish Scores					
			basic			pro		
			Understanding ↑	Reasoning ↑	Oral Conversation ↑	Understanding ↑	Reasoning ↑	Oral Conversation ↑
En	GLM-4-Voice	9B	**82.16**	**55.46**	**74.20**	**45.14**	**61.28**	**57.14**
	LLaMA-Omni	8B	47.45	36.03	64.98	28.85	47.62	42.96
	Freeze-Omni	7B	58.68	37.52	52.24	29.21	5.49	38.98
	Mini-Omni	0.5B	12.42	12.78	30.74	21.66	0	21.42
	Mini-Omni2	0.5B	16.27	15.60	33.98	24.43	0	24.53
	SLAM-Omni	0.5B	26.60	23.36	47.54	25.79	24.72	30.16
	GPT-4o-Audio-Preview	-	87.76	81.73	94.84	62.92	57.07	73.63
	Whisper + GLM-4-9B-Chat-HF	9B	90.83	76.29	88.26	-	-	-
	Whisper + Llama-3.1-8B-Instruct	8B	50.35	75.14	88.72	-	-	-
	Whisper + Qwen2-7B-Instruct	7B	64.99	77.94	91.52	-	-	-
	Whisper + Qwen2-0.5B-Instruct	0.5B	46.35	44.41	60.13	-	-	-
	Whisper + GPT-4o	-	92.62	82.91	94.60	-	-	-

cascaded system with text LLM

Basic tasks: focus on "content"

Pro tasks: Includes paralingustic, music, environment sound, ...

GLM-4-Voice LLaMA-Omni Freeze-Omni

GLM-4-9B *LLaMA-3.1* *Qwen2* Backbone LLM

Table from URO-Bench

Lang	Models	LLM Scale	Task Accomplish Scores					
			basic			pro		
			Understanding ↑	Reasoning ↑	Oral Conversation ↑	Understanding ↑	Reasoning ↑	Oral Conversation ↑
En	GLM-4-Voice	9B	**82.16**	**55.46**	**74.20**	**45.14**	**61.28**	**57.14**
	LLaMA-Omni	8B	47.45	36.03	<u>64.98</u>	28.85	<u>47.62</u>	<u>42.96</u>
	Freeze-Omni	7B	<u>58.68</u>	<u>37.52</u>	52.24	<u>29.21</u>	5.49	38.98
	Mini-Omni	0.5B	12.42	12.78	30.74	21.66	0	21.42
	Mini-Omni2	0.5B	16.27	15.60	33.98	24.43	0	24.53
	SLAM-Omni	0.5B	26.60	23.36	47.54	25.79	24.72	30.16
	GPT-4o-Audio-Preview	-	87.76	81.73	94.84	62.92	57.07	73.63
	Whisper + GLM-4-9B-Chat-HF	9B	90.83	76.29	88.26	-	-	-
	Whisper + Llama-3.1-8B-Instruct	8B	50.35	75.14	88.72	-	-	-
	Whisper + Qwen2-7B-Instruct	7B	64.99	77.94	91.52	-	-	-
	Whisper + Qwen2-0.5B-Instruct	0.5B	46.35	44.41	60.13	-	-	-
	Whisper + GPT-4o	-	92.62	82.91	94.60	-	-	-

cascaded system with text LLM

Catastrophic forgetting when supporting speech modality

GLM-4-Voice LLaMA-Omni Freeze-Omni

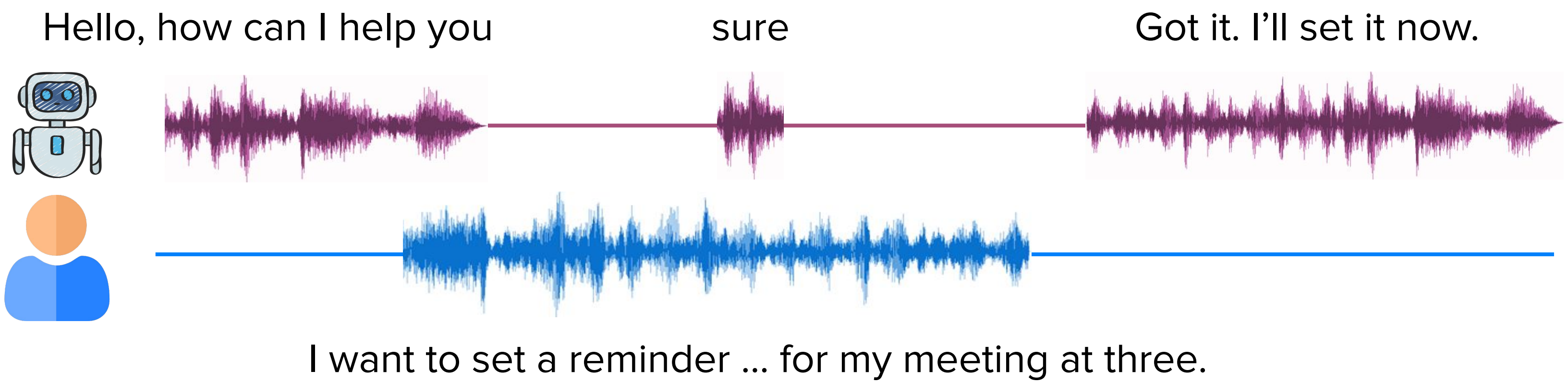
GLM-4-9B *LLaMA-3.1* *Qwen2* Backbone LLM

Table from URO-Bench

Lang	Models	LLM Scale	UTMOS ↑	WER / CER ↓	Latency (ms) ↓
En	GLM-4-Voice	9B	4.15	11.12%	3243.64
	LLaMA-Omni	8B	4.00	8.86%	**226.13**†
	Freeze-Omni	7B	4.33	20.88%	3675.47
	Mini-Omni	0.5B	4.42	5.85%	399.16
	Mini-Omni2	0.5B	4.43	9.00%	402.48
	SLAM-Omni	0.5B	**4.45**	**4.05%**	800*
	GPT-4o-Audio-Preview	-	4.05	5.51%	-

- Generated speech quality
  - Most modern models can generate speech with good quality
- Response Latency
  - Streaming or not
  - Depend on hardware

How to evaluate Full-duplex SLM?

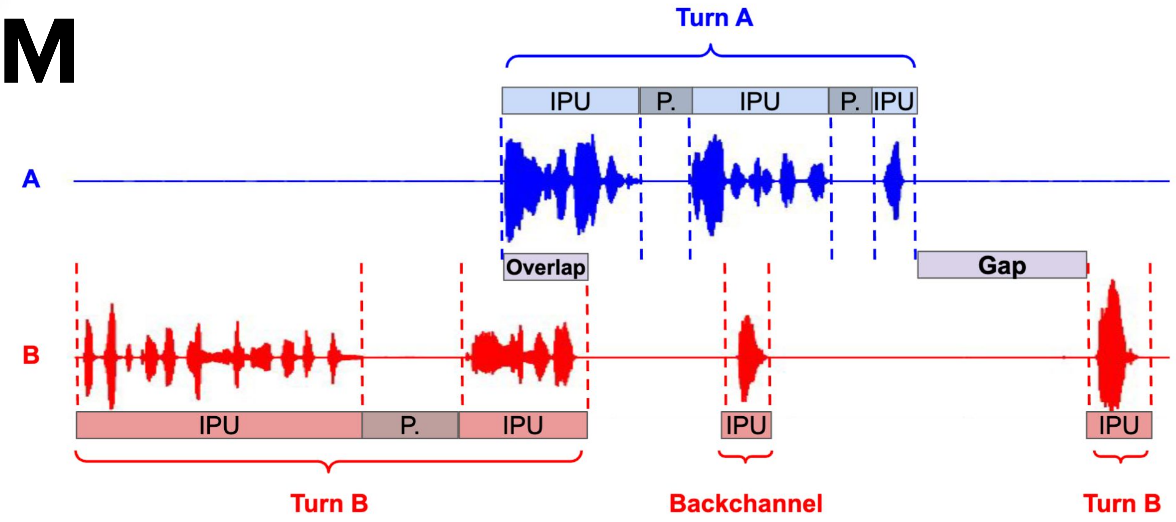


- Speech Understanding
- Paralingustic
- Reasoning
- Speech quality
- …
- Full-duplex behavior

→ Existing benchmark for S2S models (e.g. URO-Bench)

→ e.g. dGSLM, Full-duplex-Bench,

dGSLM

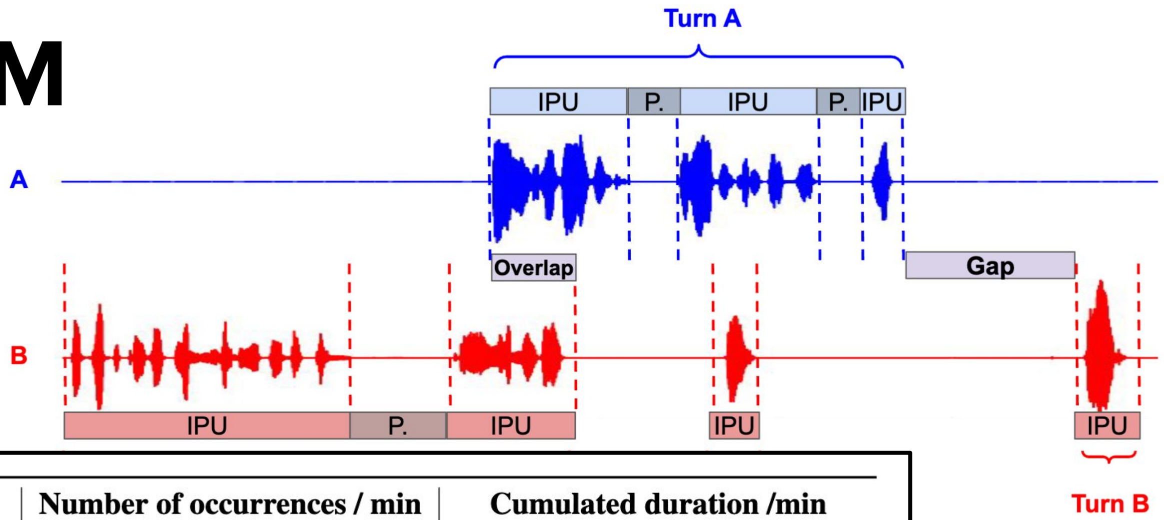


4 events in turn-taking

- IPU – holding the floor
- Pause – brief hesitation
- Gap – delayed turn transfer
- Overlap – competing turns (incl. backchannels)

Generative Spoken Dialogue Language Modeling
https://arxiv.org/abs/2203.16502

dGSLM



Id	Model	Number of occurrences / min IPU	Pause	Gap	Overlap	Cumulated duration /min IPU	Pause	Gap	Overlap
0	**MS-TLM**	19.4	10.6	5.1	3.3	49.4s	8.9s	2.9s	1.3s
1	**DLM-1**	17.7	7.9	3.9	5.5	41.4s	13.8s	10.7s	6.1s
2	**DLM-2**	20.0	10.4	5.5	3.6	48.9s	9.1s	3.6s	1.7s
3	**DLM-3**	19.0	1.8	4.9	11.7	65.0s	1.1s	1.8s	8.1s
4	**DLM-4**	18.9	3.2	5.6	9.4	60.7s	2.4s	2.9s	6.1s
5	**DLM-5**	24.2	5.4	7.2	10.9	59.1s	3.6s	2.9s	5.8s
6	**Cascaded**	17.5	0.0	14.9	0.0	54.8s	0.0s	5.3s	0.0s
	Ground Truth	21.6	7.0	7.5	6.5	53.5s	5.5s	4.4s	3.6s
	Training Set	25.9	7.2	8.6	10.0	54.5s	5.6s	4.6s	4.7s

corpus-level statistics

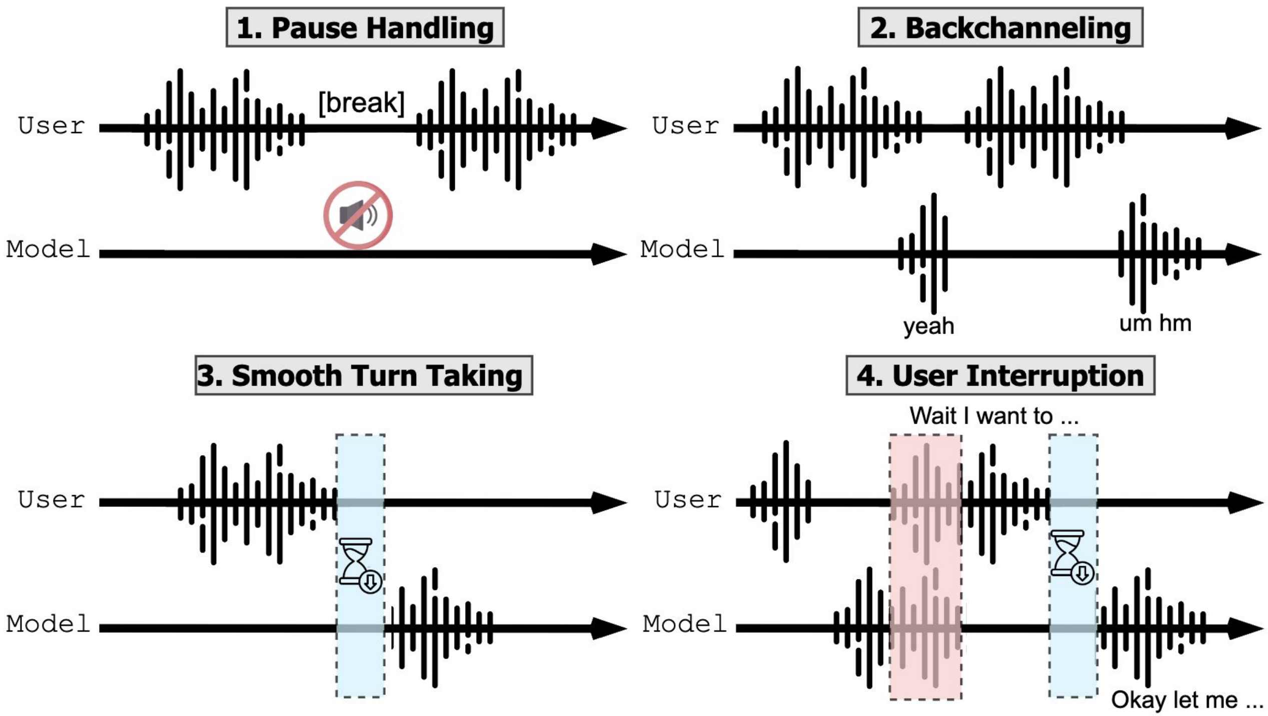
compare between generated and ground truth

Full-duplex-Bench v1

How should the model respond when a specific event happens?

Full-Duplex-Bench: A Benchmark to Evaluate Full-duplex Spoken Dialogue Models on Turn-taking Capabilities (https://arxiv.org/abs/2503.04721)

Full-duplex-Bench v1



How should the model respond when a specific event happens?

Full-Duplex-Bench: A Benchmark to Evaluate Full-duplex Spoken Dialogue Models on Turn-taking Capabilities (https://arxiv.org/abs/2503.04721)

Full-duplex-Bench v1

Dimension	Pause Handling	Backchannel	Smooth Turn Taking		User Interruption	
Data	Synthetic	ICC	Candor		Synthetic	
Metric	TOR (↓)	TOR (↓)	TOR (↑)	Latency (↓)	TOR (↑)	Latency (↓)
dGSLM	0.934	0.691	**0.975**	0.352	0.917	2.531
Moshi	0.985	1.000	0.941	**0.265**	**1.000**	**0.257**
Freeze-Omni	**0.642**	**0.636**	0.336	0.953	0.867	1.409
Gemini Live	0.255	0.091	0.655	1.301	0.891	1.183

- Pause handling / Backchannel - The user does not yield the turn - **Take Over Rate (TOR)** should be low
- Smooth Turn Taking - The user yields the turn - TOR should be high
- User Interruption - TOR should be high

Full-duplex-Bench v1

Dimension	Pause Handling	Backchannel	Smooth Turn Taking		User Interruption	
Data	Synthetic	ICC	Candor		Synthetic	
Metric	TOR (↓)	TOR (↓)	TOR (↑)	Latency (↓)	TOR (↑)	Latency (↓)
dGSLM	0.934	0.691	**0.975**	0.352	0.917	2.531
Moshi	0.985	1.000	0.941	**0.265**	**1.000**	**0.257**
Freeze-Omni	**0.642**	**0.636**	0.336	0.953	0.867	1.409
Gemini Live	0.255	0.091	0.655	1.301	0.891	1.183

- Moshi is sensitive and aggressive
  - Takes the turn very often
  - Tend to interrupt the user

Full-duplex-Bench v1

Dimension	Pause Handling	Backchannel	Smooth Turn Taking		User Interruption	
Data	Synthetic	ICC	Candor		Synthetic	
Metric	TOR (↓)	TOR (↓)	TOR (↑)	Latency (↓)	TOR (↑)	Latency (↓)
dGSLM	0.934	0.691	**0.975**	0.352	0.917	2.531
Moshi	0.985	1.000	0.941	**0.265**	**1.000**	**0.257**
Freeze-Omni	**0.642**	**0.636**	0.336	0.953	0.867	1.409
Gemini Live	0.255	0.091	0.655	1.301	0.891	1.183

- Freeze-Omni
  - Handle the pause and backcannel better
  - Smooth Turn taking is still challenging
  - **Latency is high**

Full-duplex Evaluation

- Speech Understanding
- Paralingustic
- Reasoning
- Speech quality
- …
- Full-duplex behavior

Existing benchmark for S2S models (e.g. URO-Bench)

e.g. dGSLM, Full-duplex-Bench,

- Speech Understanding - **Basic capability**
- Corpus statistics - **Naturalness**
- Observe the behavior when specific event occurs - **Naturalness**

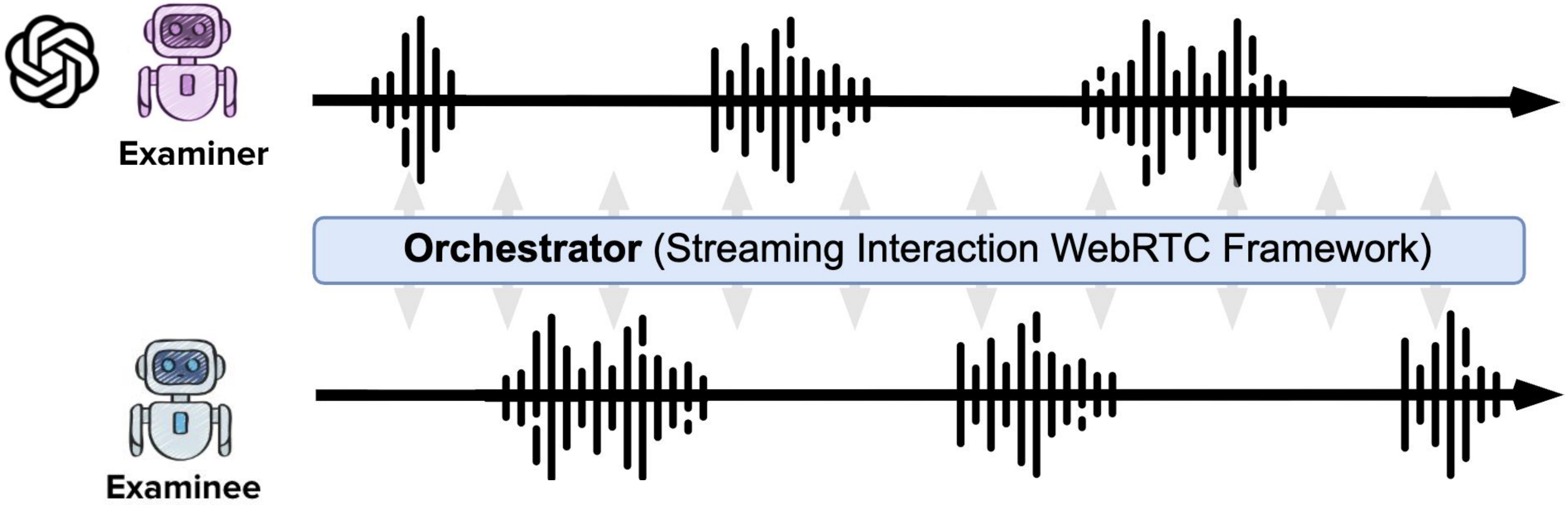
Full-Duplex-Bench-v2: A Multi-Turn Evaluation Framework for Duplex Dialogue Systems with an Automated Examiner







Full-Duplex-Bench-v2



- (Dynamic) Multi-turn evaluation -
  Can SLMs follow the instruction after multiple turns?

Reference: https://learn.microsoft.com/en-us/azure/ai-foundry/openai/how-to/realtime-audio

Previous multi-turn benchmark

URO-Bench

Round 1

(Jack’s voice) *Hi, I'm **Jack**. My favourite food is sushi.*

"Hi, Jack! Sushi is such a healthy and delicious choice!

Round 2

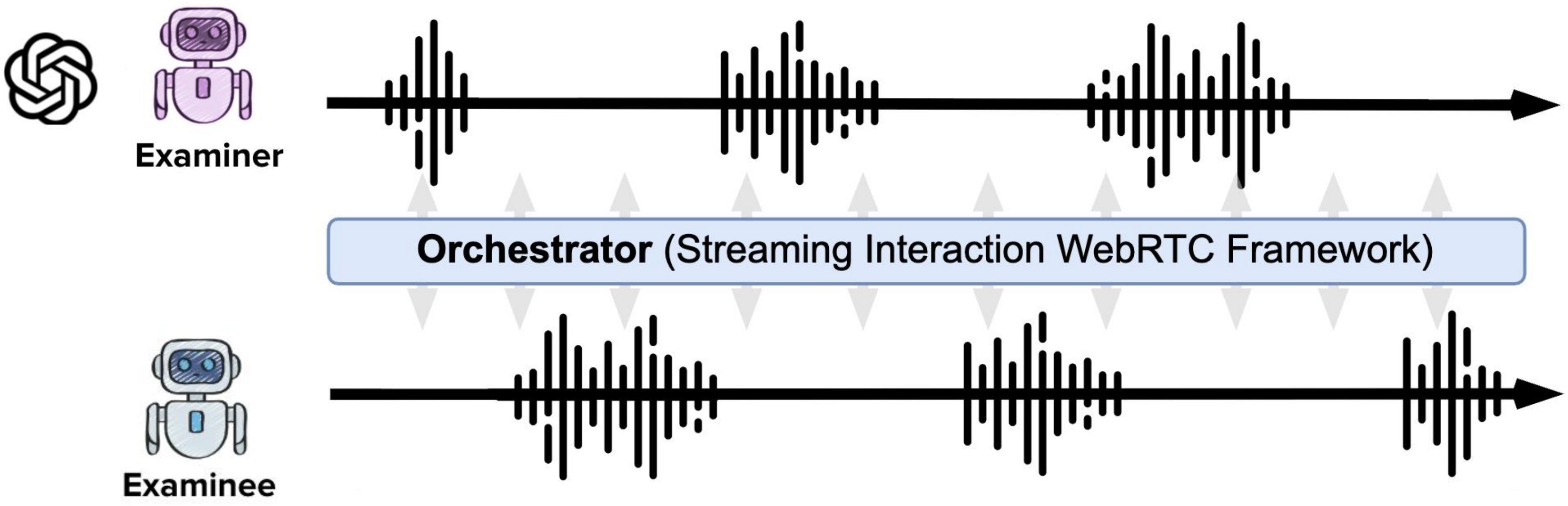
(Olivia’s voice) *Hi, I'm **Olivia**. My favourite food is chocolate.*

Hi, Olivia!Chocolate is always a great treat

Round 3

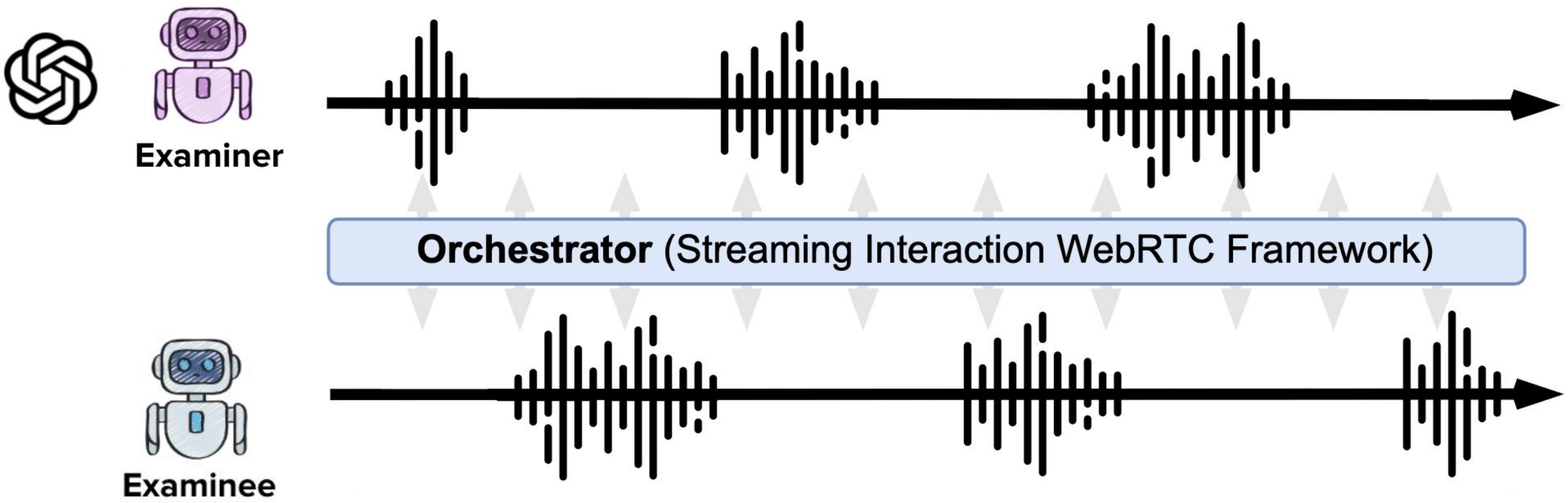
(Jack’s voice) *How to make my favourite food at home?*

To make sushi, you’ll need sushi rice, seaweed, and your choice of fillings...



Stress Testing!

- Contains overlapping, interruption
- Test whether Examinee can fulfill the task under more challenging scenario (handle interruption, understand the intent after multiple turns)



Stress Testing!

- Slow Mode - Examiner is less aggrssive (more like turn-by-turn)
- Fast Mode - Examiner is more aggrssive

Full-Duplex-Bench-v2 (Examples)

Slow

Correction Task

The examiner keeps changing the intent.

Transcript

Examiner: I, I want to book a train ticket for next Friday.

Evaluatee: Sure, I can help with that. Can you tell me where you're traveling from and where you'd like to go?

Examiner: Oh, right. I'm traveling from Chicago to St. Louis, and I meant the Friday after next.

Evaluatee: Got it. You're traveling from Chicago to St. Louis next Friday afternoon. Let me check the available train options. Do you have a preferred time range or class of service in mind?

Examiner: Yes, I'd prefer an afternoon departure, and I'd like a seat in business class.

Evaluatee: Understood. Let me find the best options for an afternoon departure from Chicago to St. Louis in business class next Friday. I'll bring up the available trains and times for you.

Full-Duplex-Bench-v2 (Examples)

Fast

Correction Task

The examiner keeps changing the intent.

Transcript

Examiner: Hey there. I'd like to get a train ticket for next Friday. Could you help me with that?

Evaluatee: Sure, I'd be happy to help. Let's figure out the best option for your

Examiner: Oh, sorry. I asked. So, for that trip, I'd prefer an early morning departure and I'd like a seat in first class. Oh, sorry. I meant the Friday after next. That's the date I need the ticket for.

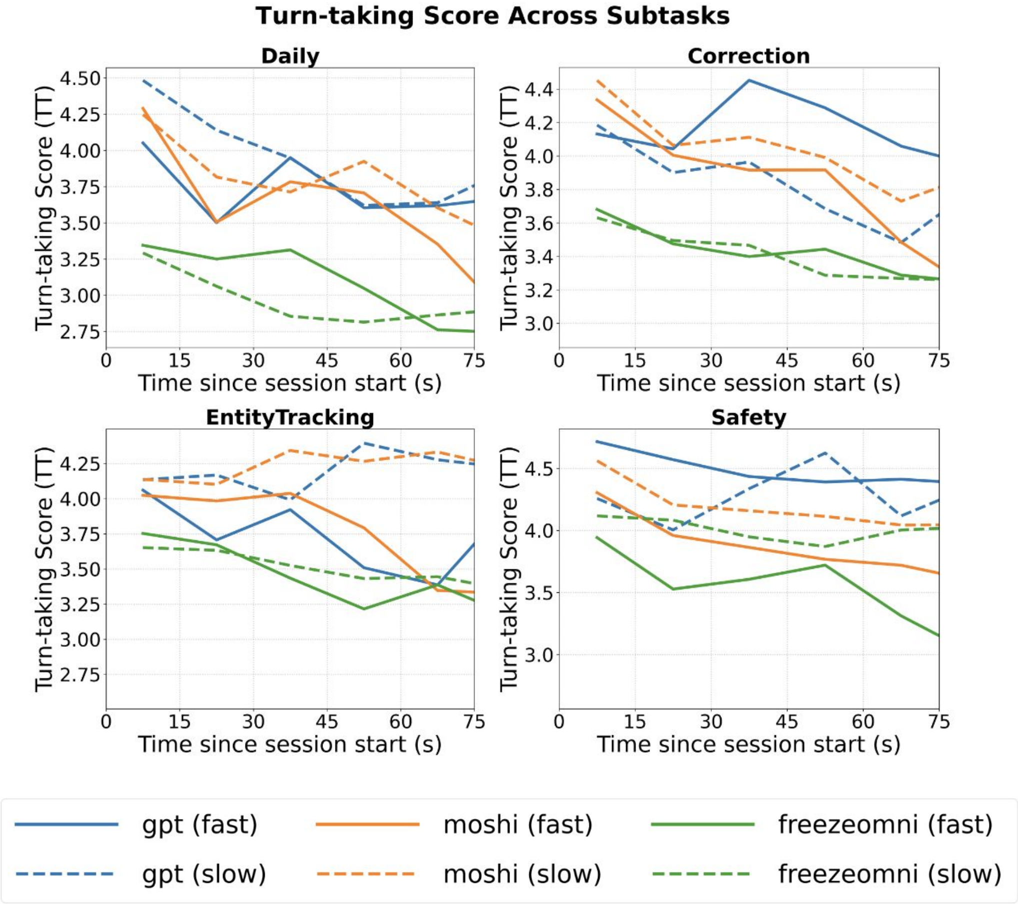
Evaluatee: got it. You're looking for a first-class train ticket for Friday afternoon. Let me check the

Examiner: So, I'd like a morning departure in first class. Actually, just to confirm, I'm looking for a first... Sure, let me confirm the details. The route is from here to your destination on Friday, October 6th in the early morning.

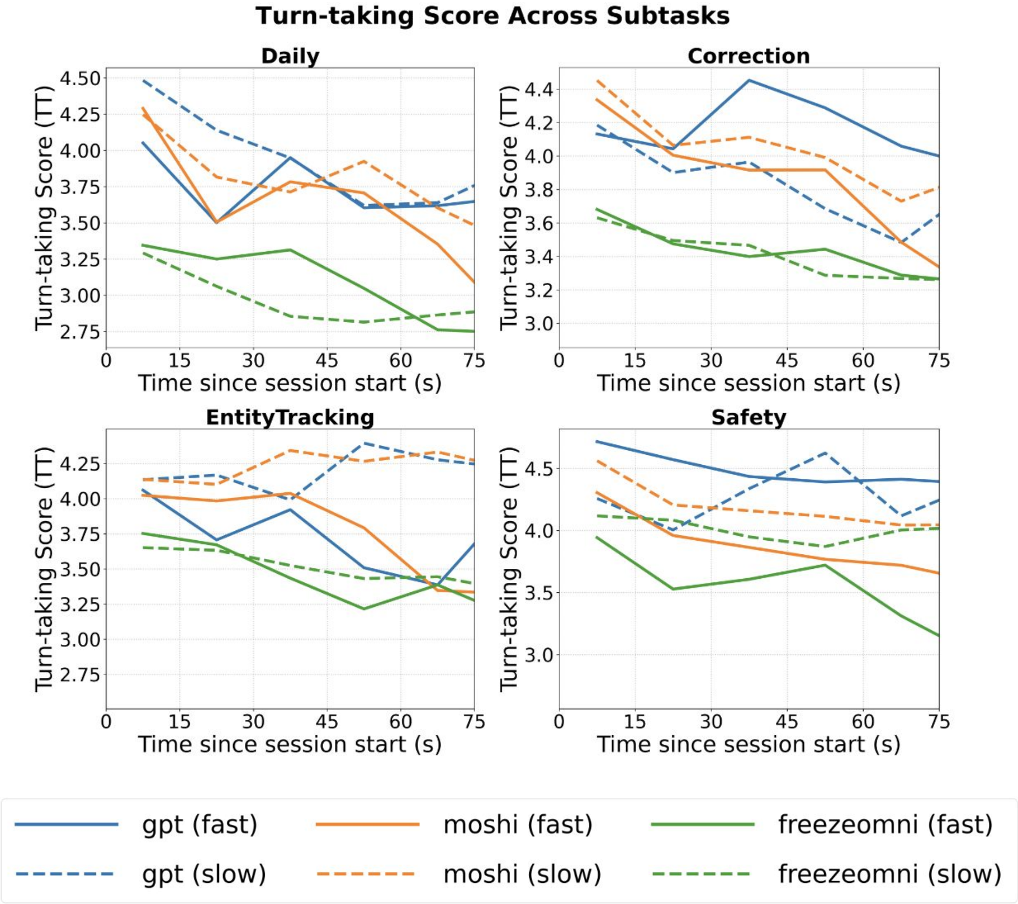
Evaluatee: Got it. You want a first-class train ticket on Friday, October 6th with an early morning departure. Could you tell me the departure and arrival cities you're traveling between?

Examiner: Exactly. A first class. Yes,

LLM-as-judge



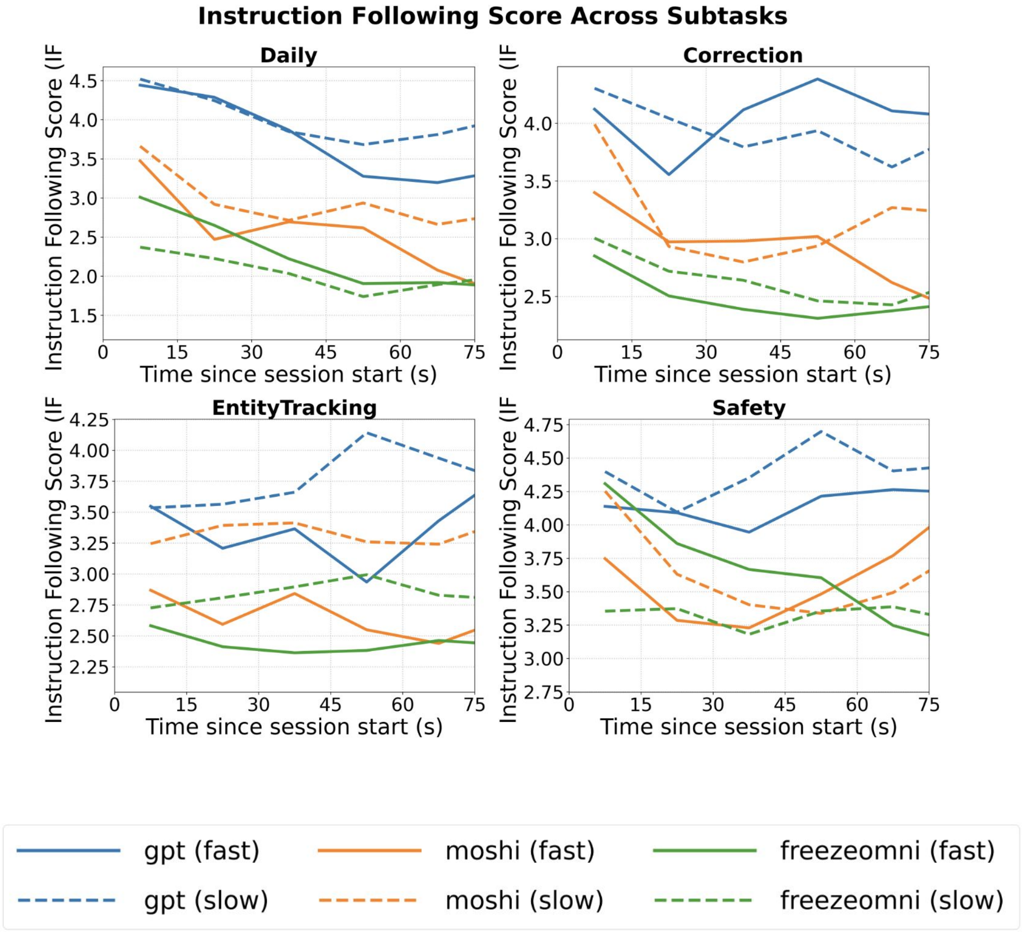
- Daily
- Correction
- Entity Tracking
- Safety



LLM-as-judge

The results is fluctuating, but with the trend

- Turn taking score: Gradually decay as time goes by



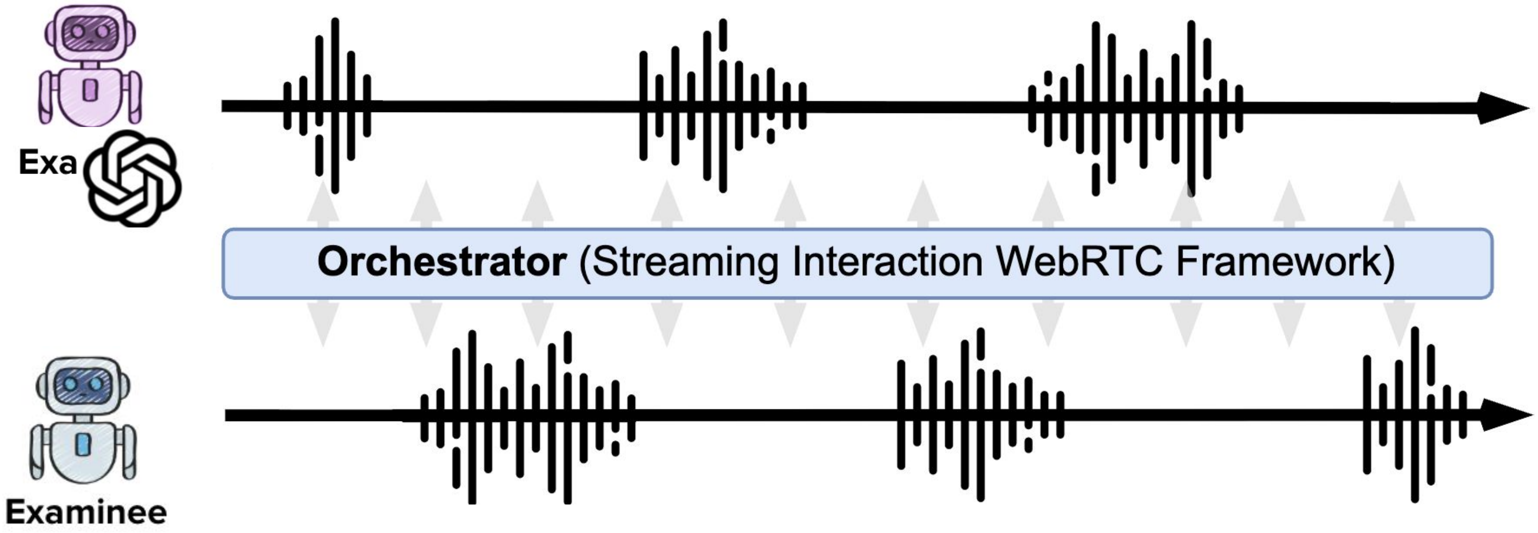
LLM-as-judge

The results is fluctuating, but with the trend

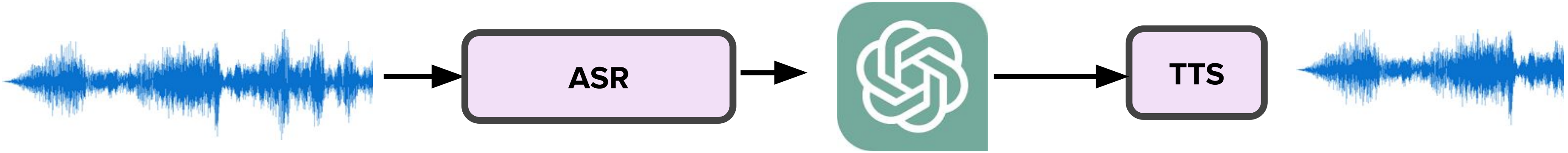
- Instruction Following Score: Still suffer from multi-turn dialouge
- The performance may bounce back, but hard to get back

Full-Duplex-Bench-v2

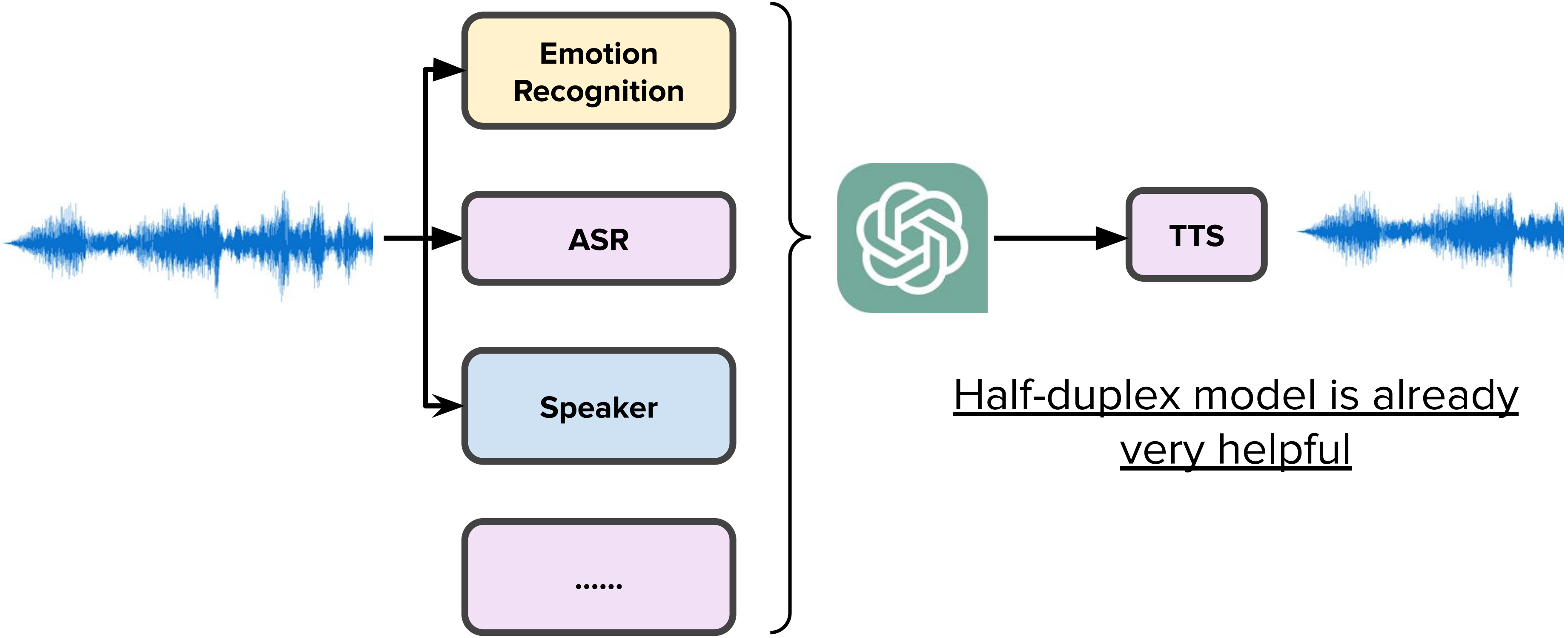
- The performance gradually degrade when time goes by
- Provide a valuable frameworks (WebRTC protocol) to scale up (Just like Google Meet)



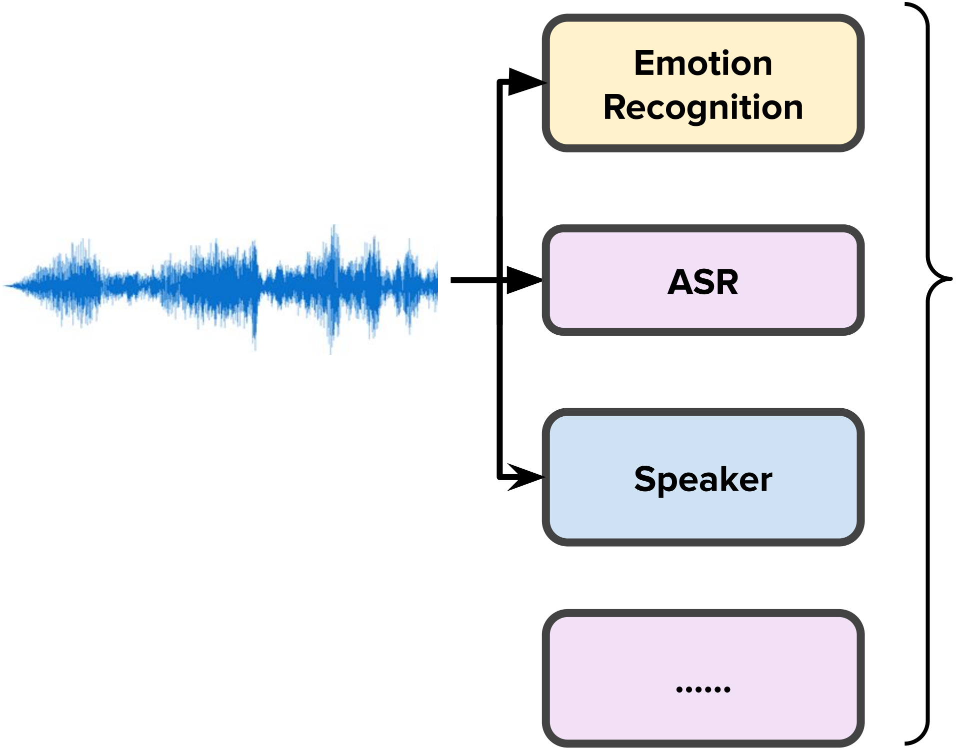
Why Do We Need
End-to-End Full-Duplex
Spoken Language Model?

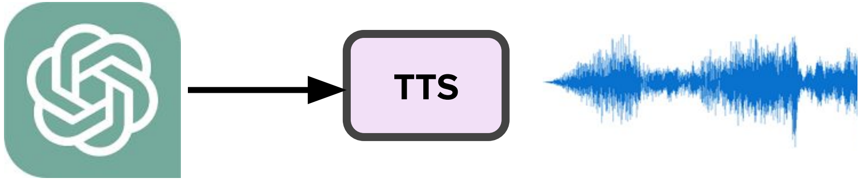


<u>Half-duplex model is already very helpful</u>



Half-duplex model is already very helpful





<u>Half-duplex model is already very helpful</u>

Do we really need backchanneling?

Do we really need smooth turn taking?

What’s the fundamental restriction of half-duplex models

The most difficult thing when studying in NTU...


BILLY: Steering Large Language Models via Merging Persona Vectors for Creative Generation

Tsung-Min Pai[1] Jui-I Wang[2] Li-Chun Lu[3] Shao-Hua Sun[1,3]
Hung-Yi Lee[1] Kai-Wei Chang[4]

[1]Department of Electrical Engineering, National Taiwan University
[2]Department of Computer Science & Information Engineering, National Taiwan University
[3]Graduate Institute of Communication Engineering, National Taiwan University
[4]CSAIL, Massachusetts Institute of Technology

Correspondence to: Tsung-Min Pai <b09602017@ntu.edu.tw>, Kai-Wei Chang <kwchang@mit.edu>

arXiv:2510.10157v1 [cs.CL] 11 Oct 2025

Abstract

Multi-LLM systems enhance the creativity of large language models by simulating human collective intelligence but suffer from significant drawbacks, such as high computational costs and inference latency. To address these limitations, we propose BILLY (BlendIng persona vectors for Large Language model creativitY), a training-free framework that captures the benefits of multi-LLM collaboration, i.e., inducing diverse perspectives and specialized expertise, within a single model. BILLY operates by extracting and blending multiple distinct persona vectors directly in the model's activation space. We steer the model's generation process with this merged vector while inference, enabling multi-perspective output without explicit multi-LLM communication. Our experiments across creativity-oriented benchmarks demonstrate that BILLY surpasses single model prompting and traditional multi-LLM approaches, while substantially reducing inference time and computational costs. Our analyses further reveal that distinct persona vectors can be blended to achieve both effective control over complementary aspects of generation and greater interpretability.

Figure 1: BILLY (BlendIng persona vectors for Large Language model creativitY). To enhance the creativity of a single LLM, we extract and fuse the persona vectors of a Creative Professional and an Environmentalist, steering a base model by this composite vector to generate outputs based on both domains.

1 Introduction

Creativity is widely recognized as a cornerstone of human progress, driving innovation across domains and enabling major scientific breakthroughs (Feist, 1998; Simonton, 2004). Extending this perspective, recent research (Franceschelli and Musolesi, 2023; Lin et al., 2025) has explored the creativity of large language models (LLMs), viewing them as promising tools for applications such as story writing (Gómez-Rodríguez and Williams, 2023; Chen et al., 2024b), design ideation (Ding et al., 2023; Hou et al., 2024; Hung et al., 2025), and scientific discovery (Yang et al., 2024b; Si et al., 2025), thereby augmenting human problem-solving and imagination (Zhang et al., 2023; Lim and Perrault, 2024; Liu et al., 2024, 2025).

Recent studies highlight the potential of the multi-LLM paradigm (Guo et al., 2024; Lu et al., 2024; Lin et al., 2025), which aims to simulate human collective intelligence (Leimeister, 2010) by engaging multiple LLMs in iterative discussion to arrive at more comprehensive and well-balanced solutions (Tran et al., 2025). This paradigm allows systems to generate a broader range of ideas beyond a single model can reach (Tran et al., 2025). In the context of creativity, these frameworks often assign diverse roles to LLMs and employ structured, multi-round interactions (Lu et al., 2024; Lin et al., 2025; Su et al., 2025; Liang et al., 2024).

Our method, BILLY, takes its name from The Minds of Billy Milligan by Daniel Keyes—but the emoji in the title nods to "Billie" Eilish instead.

1

Research

The most difficult thing when studying in NTU…

BILLY: Steering Large Language Models via Merging Persona Vectors for Creative Generation

Tsung-Min Pai[1] Jui-I Wang[2] Li-Chun Lu[3] Shao-Hua Sun[1,3]
Hung-Yi Lee[1] Kai-Wei Chang[4]
[1]Department of Electrical Engineering, National Taiwan University
[2]Department of Computer Science & Information Engineering, National Taiwan University
[3]Graduate Institute of Communication Engineering, National Taiwan University
[4]CSAIL, Massachusetts Institute of Technology
Correspondence to: Tsung-Min Pai <b09602017@ntu.edu.tw>, Kai-Wei Chang <kwchang@mit.edu>

arXiv:2510.10157v1 [cs.CL] 11 Oct 2025

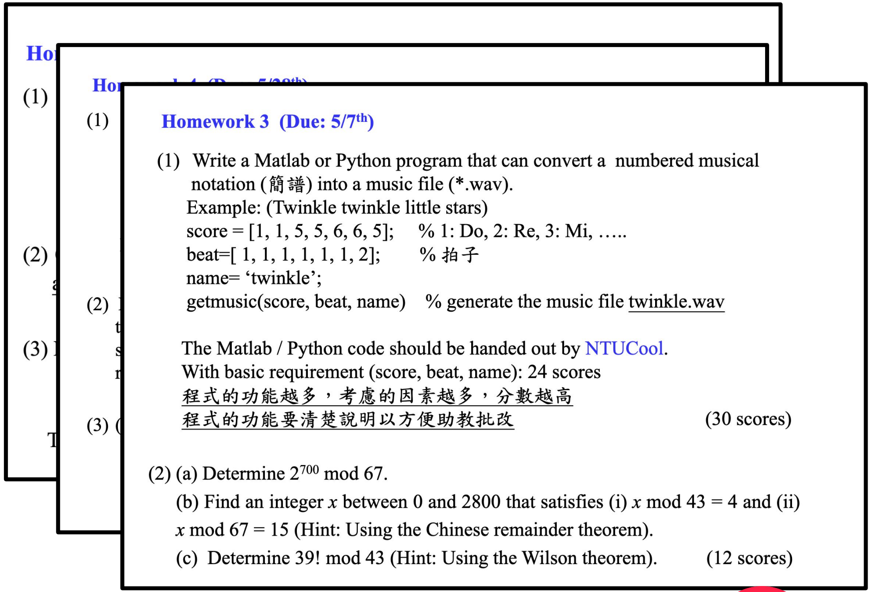
Abstract

Multi-LLM systems enhance the creativity of large language models by simulating human collective intelligence but suffer from significant drawbacks, such as high computational costs and inference latency. To address these limitations, we propose BILLY (**B**lend**I**ng persona vectors for **L**arge **L**anguage model creativit**Y**), a training-free framework that captures the benefits of multi-LLM collaboration, *i.e.*, inducing diverse perspectives and specialized expertise, within a single model. BILLY operates by extracting and blending multiple distinct persona vectors directly in the model's activation space. We steer the model's generation process with this merged vector while inference, enabling multi-perspective output without explicit multi-LLM communication. Our experiments across creativity-oriented benchmarks demonstrate that BILLY surpasses single model prompting and traditional multi-LLM approaches, while substantially reducing inference time and computational costs. Our analyses further reveal that distinct persona vectors can be blended to achieve both effective control over complementary aspects of generation and greater interpretability.

1 Introduction

Creativity is widely recognized as a cornerstone of human progress, driving innovation across domains and enabling major scientific breakthroughs (Feist, 1998; Simonton, 2004).

Research



Homework 3 (Due: 5/7th)

(1) Write a Matlab or Python program that can convert a numbered musical notation (簡譜) into a music file (*.wav).
Example: (Twinkle twinkle little stars)
score = [1, 1, 5, 5, 6, 6, 5]; % 1: Do, 2: Re, 3: Mi, …..
beat=[1, 1, 1, 1, 1, 1, 2]; % 拍子
name= 'twinkle';
getmusic(score, beat, name) % generate the music file twinkle.wav

The Matlab / Python code should be handed out by NTUCool.
With basic requirement (score, beat, name): 24 scores
程式的功能越多，考慮的因素越多，分數越高
程式的功能要清楚說明以方便助教批改 (30 scores)

(2) (a) Determine 2^{700} mod 67.
(b) Find an integer x between 0 and 2800 that satisfies (i) x mod 43 = 4 and (ii) x mod 67 = 15 (Hint: Using the Chinese remainder theorem).
(c) Determine 39! mod 43 (Hint: Using the Wilson theorem). (12 scores)

Homework

The most difficult thing when studying in NTU...

Riding a bike when it's raining

The most difficult thing when studying in NTU...

...

<u>Decide what to have for lunch / dinner</u>


NTU Sports Center
國立臺灣大學綜合體育館
Witch House
女巫店
Wenzhou St
Sports Field
運動場
Banyan Ave
TaiDa Drunken
醉月湖
Drunken Moon Lake
國立臺灣大學醉月湖
Section 2, Xinhai Rd
Dunqing Park
敦親公園
飛捷廣告影印
數位印刷台大店
Copy shop
辛亥路二段
Lane 118, Section 2, Heping E Rd
Xiaolizi Rice Porridge
小李子清粥小菜
National Taipei
University of...
國立臺北教育大學
Taipei Heping
Basketball Gymnasium
臺北和平籃球館
科技大樓
桃花心木道
小椰林道
蒲葵道
Fu Bell
國立臺灣大學傳鐘
National Taiwan University
國立臺灣大學
Recently viewed
EE
Depart.
Layers
Google Maps
2D


Lane 118
Wenzhou St
Witch House
女巫店
NTU Sports Center
國立臺灣大學
綜合體育館
科技大樓
Section 2, Xinhai Rd
Sports Field
運動場
TaiDa Drunken
Drunken Moon Lake
國立臺灣大學醉月湖
Banyan Ave
Lane 118, Section 2, Heping E Rd
Xiaolizi Rice Porridge
小李子清粥小菜
National Taipei University of...
國立臺北教育大學
Taipei Heping Basketball Gymnasium
臺北和平籃球館
桃花心木道
小椰林道
蒲葵道
Fu Bell
國立臺灣大學傳鐘
Water Fir Ave
National Taiwan University
國立臺灣大學
Recently viewed
Layers
2D
Google Maps

The most difficult thing when studying in NTU...

What should we have for dinner?

The most difficult thing when studying in NTU...

What should we have for dinner?

I don't know

The most difficult thing when studying in NTU...

What should we have for dinner?

Anything is fine

Beef noodles, Thai food, or dumplings

Anything is fine

Beef noodles, Thai food, or dumplings

Beef noodles and Thai food sound good

Let’s do rock paper scissors, if I win, we go beef noodles

Beef noodles and Thai food sound good

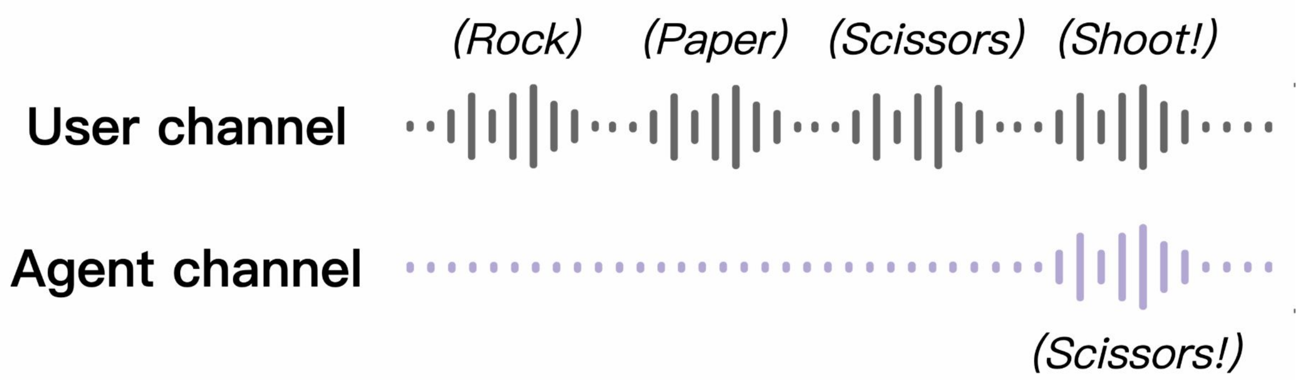
Let’s do rock paper scissors,
if I win, we go beef noodles

OK!

Rock!

Scissors!

Nowadays, the assistants (text, voice , ...) are not capable of playing rock paper scissors!



- Timing
- Tempo
- Simul. Speaking

Game-Time: Evaluating Temporal Dynamics in Spoken Language Models

Project website: https://ga642381.github.io/Game-Time

Game-Time - Instruction following Paridigm

Benchmarking “Time-awareness” of Spoken LMs
— not just *what to say*, but *when to say it*!

- Timing
- Tempo
- Simul. Speaking

Game-Time - Instruction following Paridigm

Benchmarking “Time-awareness” of Spoken LMs
— not just *what to say*, but *when to say it*!

In childhood, we learn to speak a language through games and language activities.

Time, *tempo*, and *simul. speaking* come naturally to humans...

<u>e.g. Let’s count from 1 to 10 together</u>

Game-Time - Instruction following Paridigm

Benchmarking “Time-awareness” of Spoken LMs
— not just *what to say*, but *when to say it*!

In childhood, we learn to speak a language through games and language activities.

Time, *tempo*, and *simul. speaking* come naturally to humans...

Hard to collect the data! ➜ Synthetic dataset

What is an Instruction Following (IF) task

- Perfom an IF task means:
  "Perform a **base task** *t* while simultaneously satisfying **constraints** $\{c_1, c_2, \ldots\}$"

- An IF task may include many **constraints** $c_1, c_2, \ldots$

- A **constraint** c may include many **variables** $N_1, N_2, \ldots$ (often numeric or symbolic)

reference: Generalizing Verifiable Instruction Following
https://arxiv.org/abs/2507.02833

What is an Instruction Following (IF) task

e.g. "Please count from one to ten"

- Perfom an IF task means:
  "Perform a **base task** *t* while simultaneously satisfying **constraints** $\{c_1, c_2, \ldots\}$"
  Number Counting

reference: Generalizing Verifiable Instruction Following
https://arxiv.org/abs/2507.02833

What is an Instruction Following (IF) task

e.g. "Please count from one to ten"

- Perfom an IF task means:
  "Perform a **base task** *t* while simultaneously satisfying **constraints** $\{c_1, c_2, \ldots\}$"
  Number Counting
- An IF task may include many **constraints** $c_1, c_2, \ldots$
  Generated content is sequential between number A to number B

reference: Generalizing Verifiable Instruction Following
https://arxiv.org/abs/2507.02833

What is an Instruction Following (IF) task

e.g. **"Please count from one to ten"**

- Perfom an IF task means:
  "Perform a **base task** *t* while simultaneously satisfying **constraints** $\{c_1, c_2, \ldots\}$"
  Number Counting
- An IF task may include many **constraints** $c_1, c_2, \ldots$
  Generated content is sequential between number A to number B
- A **constraint** c may include many **variables** $N_1, N_2, \ldots$ (often numeric or symbolic)
  N1 = 1, N2 = 10

reference: Generalizing Verifiable Instruction Following
https://arxiv.org/abs/2507.02833

What is an Instruction Following (IF) task

e.g. **"Please count from one to ten"**

- Perfom an IF task means:
  "Perform a **base task** *t* while simultaneously satisfying **constraints** $\{c_1, c_2, \ldots\}$"

  Number Counting

- An IF task may include many **constraints** $c_1, c_2, \ldots$

  Generated content is sequential between number A to number b

- A **constraint** c may include many **variables** $N_1, N_2, \ldots$ (often numeric or symbolic)

  N1 = 1, N2 = 10

- A "well-posed" IF task should avoid mutually *contradictory* constraints.

reference: Generalizing Verifiable Instruction Following
https://arxiv.org/abs/2507.02833

Instruction Following Task Examples			
User Prompt (Instruction)	**Base Task**	**Constraints**	**Variable**
“Please count from one to ten”	Number-counting	c_1: Ranging from 1 to 10	$N_{1,1} = 1$, $N_{1,2} = 10$
“Please repeat after me: I have a pen”	Repeat	c_1: Content is I have a pen	$N_{1,1}$ = I have a pen

Game-Time Benchmark - Dataset

Manually create seed dataset template

“Please repeat after me: [N1]”

“Please count from [N1] to [N2]”

Game-Time Benchmark - Dataset

Manually create seed dataset template

"Please repeat after me: [N1]"

"Please count from [N1] to [N2]"

Stage 1 Augmentation

paraphrasing the templates

"Repeat the following sentence: [N1]"

"I want you to count from [N1] to [N2]"

Game-Time Benchmark - Dataset

Manually create seed dataset template

"Please repeat after me: [N1]"

"Please count from [N1] to [N2]"

Stage 1 Augmentation

paraphrasing the templates

"Repeat the following sentence: [N1]"

"I want you to count from [N1] to [N2]"

Stage 2

augment the variables

"Repeat the following sentence: I love you"

"Repeat the following sentence: It is sunny today"

"I want you to count from 10 to 20" ...

Game-Time Basic Tasks

Task Family	Description	Example
1-Sequence	Number/alphabet sequencing tasks	Please count from three to fifteen.
2-Repeat	Word and sentence repetition	Please repeat after me. I have a pen.
3-Compose	Sentence composition from words/scenarios	Can you make a sentence with the word "dog"?
4-Recall	Vocabulary, letter, and rhyme recall	Name five different colors.
5-Open-Ended	QA and empathy conversation	Can you describe a moment in history that you think changed the world?
6-Role-Play	Scenario-based and persona role-playing	If you are given a chance to say something to your ten-year-old self, what would you say?

e.g. “Please count from one to ten”

Game-Time Basic Tasks

6 task families - 14 base tasks

Task Family	Description	Example
1-Sequence	Number/alphabet sequencing tasks	Please count from three to fifteen.
2-Repeat	Word and sentence repetition	Please repeat after me. I have a pen.
3-Compose	Sentence composition from words/scenarios	Can you make a sentence with the word "dog"?
4-Recall	Vocabulary, letter, and rhyme recall	Name five different colors.
5-Open-Ended	QA and empathy conversation	Can you describe a moment in history that you think changed the world?
6-Role-Play	Scenario-based and persona role-playing	If you are given a chance to say something to your ten-year-old self, what would you say?

e.g. "Please count from one to ten"

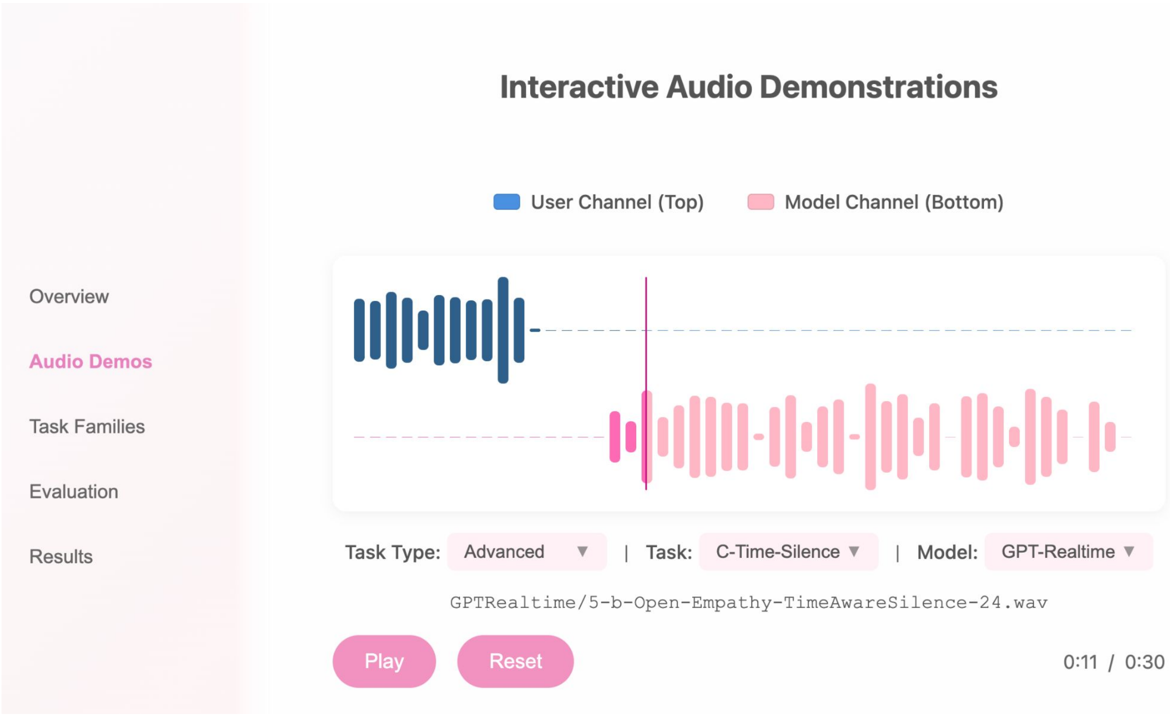
Each task family contains various task categories and multiple instructions

Game-Time Advanced Tasks

<u>**Introduce time constrained in basic tasks.**</u>

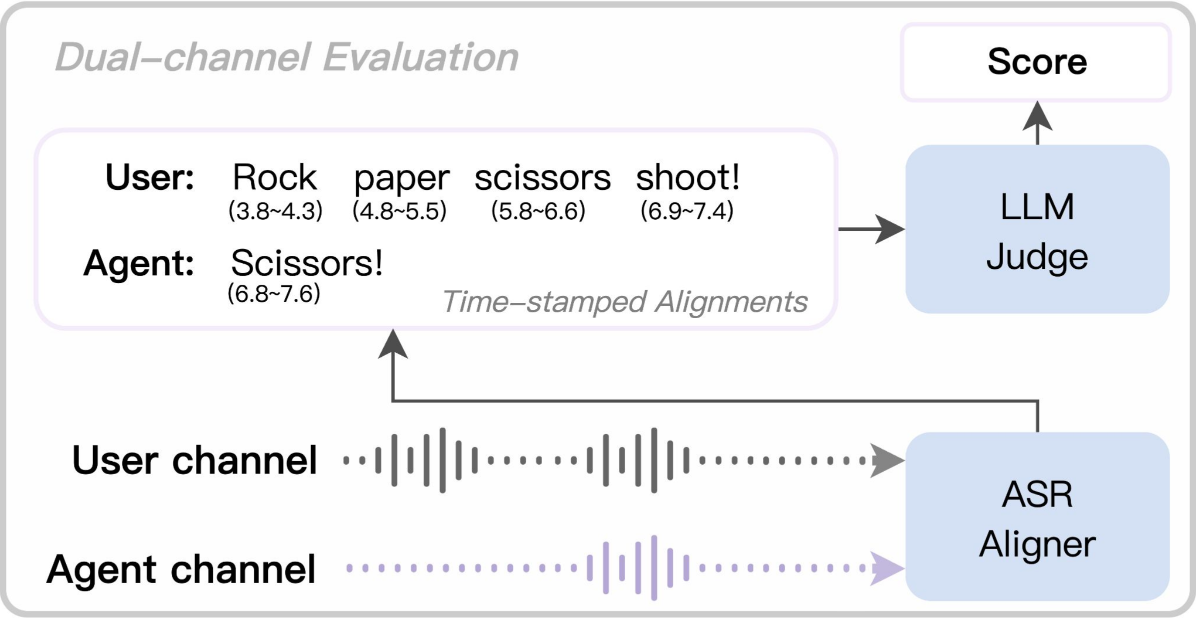
Task Family	Example
A-Time-Fast	Please count from three to fifteen **within eight seconds.**
B-Time-Slow	Can you make a sentence with the word "dog" and make sure your response **lasts longer than fifteen seconds**?
C-Time-Slience	After I finish speaking, **be quiet for five seconds**, then please repeat after me. I have a pen.
D-Tempo-Interval	Please count from three to fifteen, **pausing a one-second gap** between each number.
E-Tempo-Adhere	Please count from three to fifteen, with the tempo: **[100 bpm: "one two three four"]**
F-SimulSpeak-Shadow	I will say a sentence and you **repeat each word right after me.** Here we go: I have a pen.
G-SimulSpeak-Cue	Let's play rock-paper-scissors. Say "rock", "paper", or "scissors" when I say "shoot".

Game-Time Examples



Evaluation

LLM-as-a-judge to evaluate the instruction following score



Gemini 2.5 Pro

Models

- Moshi

dual channel

- Unmute (Streaming ASR + LLM + Streaming TTS)
- Freeze-Omni

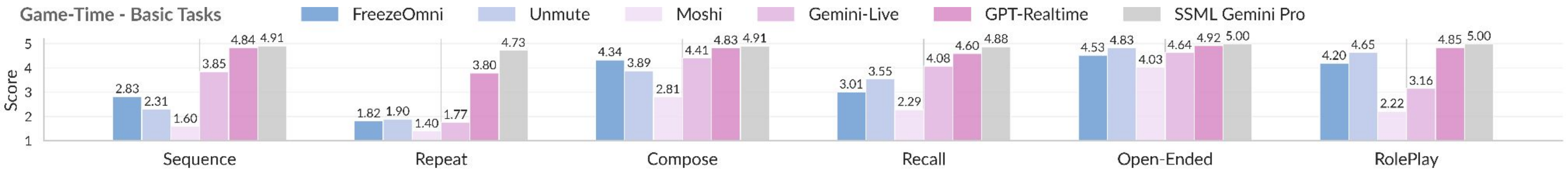
time multiplexing

- Gemini-Live
- GPT-realtime

Commercial API

- SSML-LLM (Oracal topline, off-line, non-causal)

Game-Time Basic Tasks - Instruction Following



- Commercial API > Open-sourced Models
- Time-multiplexing Model > Dual-channel Model
- Continously training LLM for speech generation is still challenging

Topline Method (Oracle method)

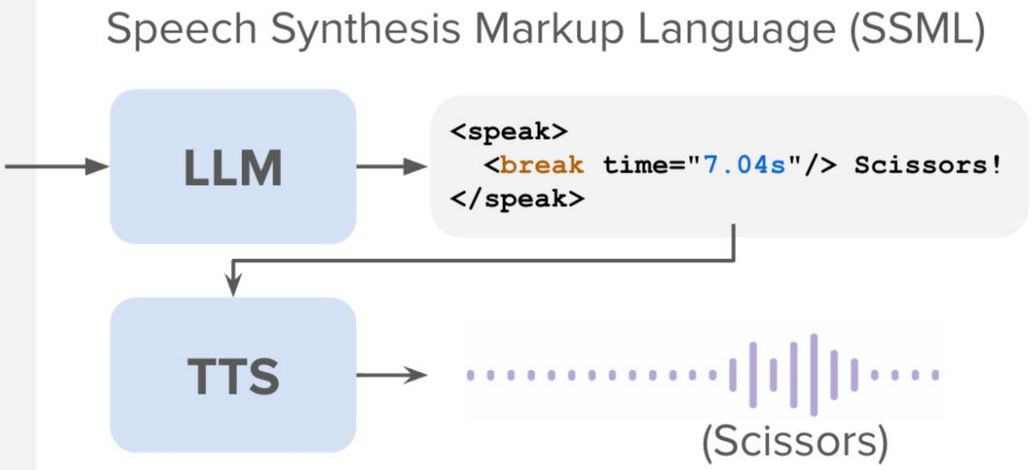
```
"alignments (SPEAKER_USER)":
  ["Let's", [0.0, 0.26]],
  ["start", [0.26, 0.48]],
  ["a", [0.48, 0.62]],
  ["round", [0.62, 0.8]],
  ["of", [0.8, 0.96]],
  ["rock-paper-scissors,", [0.96, 2.08]],
  ["throw", [2.46, 2.68]],
  ["on,", [2.68, 2.98]],
  ["shoot,", [3.24, 3.48]],
  ["rock,", [3.98, 4.34]],
  ["paper,", [4.9, 5.42]],
  ["scissors,", [5.96, 6.42]],
  ["shoot.", [7.04, 7.34]]
```

Speech Synthesis Markup Language (SSML)

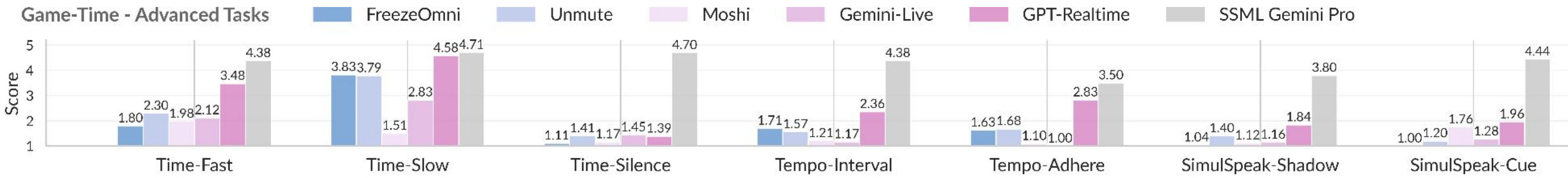
```
<speak>
  <break time="7.04s"/> Scissors!
</speak>
```

Topline Method (Oracle method)

```
"alignments (SPEAKER_USER)":
  ["Let's", [0.0, 0.26]],
  ["start", [0.26, 0.48]],
  ["a", [0.48, 0.62]],
  ["round", [0.62, 0.8]],
  ["of", [0.8, 0.96]],
  ["rock-paper-scissors,", [0.96, 2.08]],
  ["throw", [2.46, 2.68]],
  ["on,", [2.68, 2.98]],
  ["shoot,", [3.24, 3.48]],
  ["rock,", [3.98, 4.34]],
  ["paper,", [4.9, 5.42]],
  ["scissors,", [5.96, 6.42]],
  ["shoot.", [7.04, 7.34]]
```



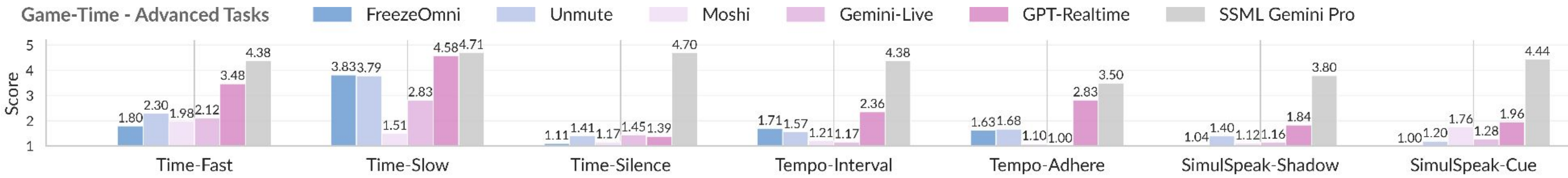
Game-Time Adv. Tasks - Instruction Following



- After introducing time constraints, performance severely degrade for all SLMs

Same instructions as basic tasks but with time constraints

Game-Time Adv. Tasks - Instruction Following

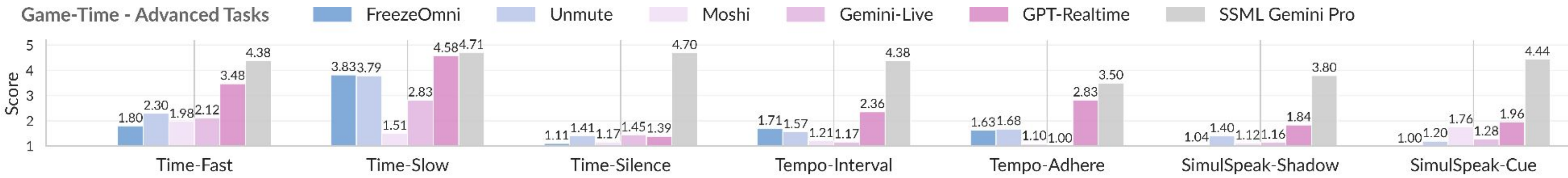


- After introducing time constraints, performance severely degrade for all SLMs

- Time-Fast, Time-Slow
  - GPT-Realtime can deliver reasonable performance

Same instructions as basic tasks but with time constraints

Game-Time Adv. Tasks - Instruction Following

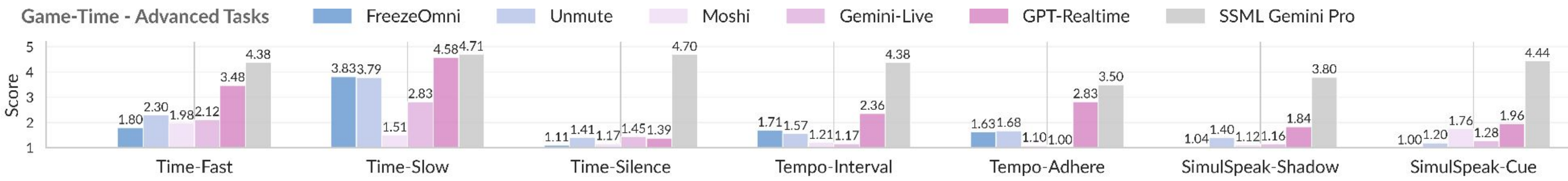


- After introducing time constraints, performance severely degrade for all SLMs

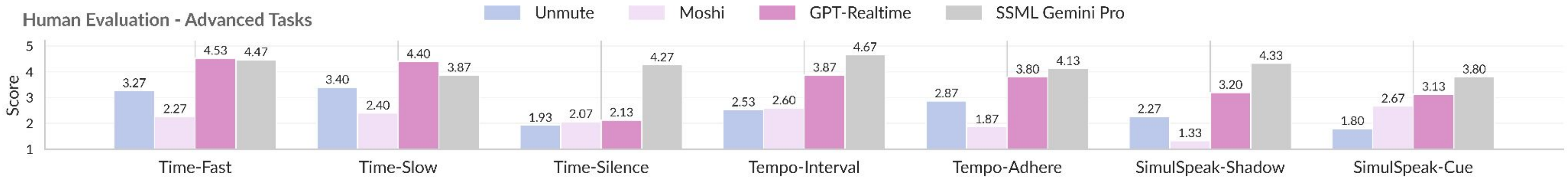
- Time-Fast, Time-Slow
  - GPT-Realtime can deliver reasonable performance

Same instructions as basic tasks but with time constraints

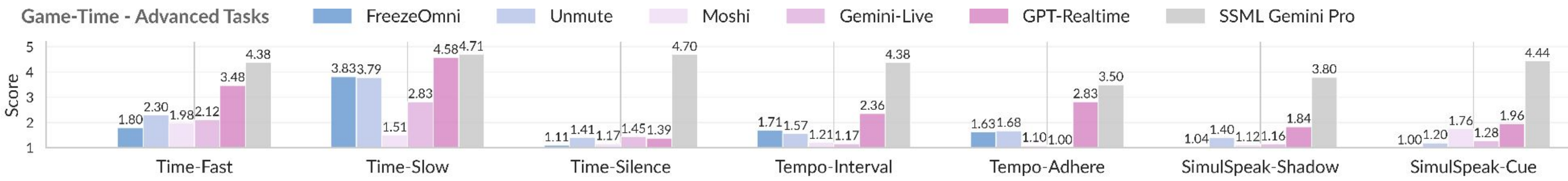
Game-Time Adv. Tasks (LLM judge)



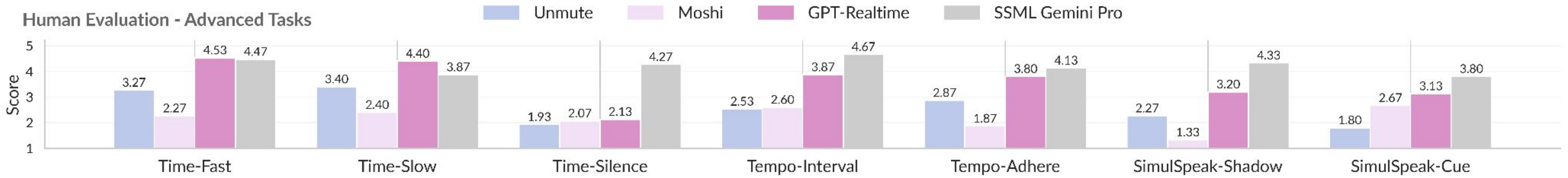
Game-Time Adv. Tasks (Human judge)



Game-Time Adv. Tasks (LLM judge)



Game-Time Adv. Tasks (Human judge)



Relative high correlation between Human and LLM

	Spearman's ρ	Pearson's r
Human - LLM	0.677	0.675
Human - ALLM	0.643	0.625

Game-Time Benchmark Findings

- The SOTA SLMs can excel on basic tasks
  - However, some SLMs (Moshi) can not

- All SLMs do not have reasonable performance when time constraints are introduced

- SLMs cann’t perform on some tasks that are easy to human

Conclusion

- More and more attenion on full-duplex SLMs
- Currently, the morden SLM do not have good enough time-awareness
- There are lots of possible methods to improve the dynamics of these models