Towards Time-Aware Full-Duplex Spoken Dialogue Models

Speaker: Kai-Wei Chang (MIT CSAIL)

MIT CSAIL
Computer Science &
Artificial Intelligence
Laboratory

About the Speaker

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

國立臺灣大學
National Taiwan University

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

About the Speaker

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

國立臺灣大學
National Taiwan University

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

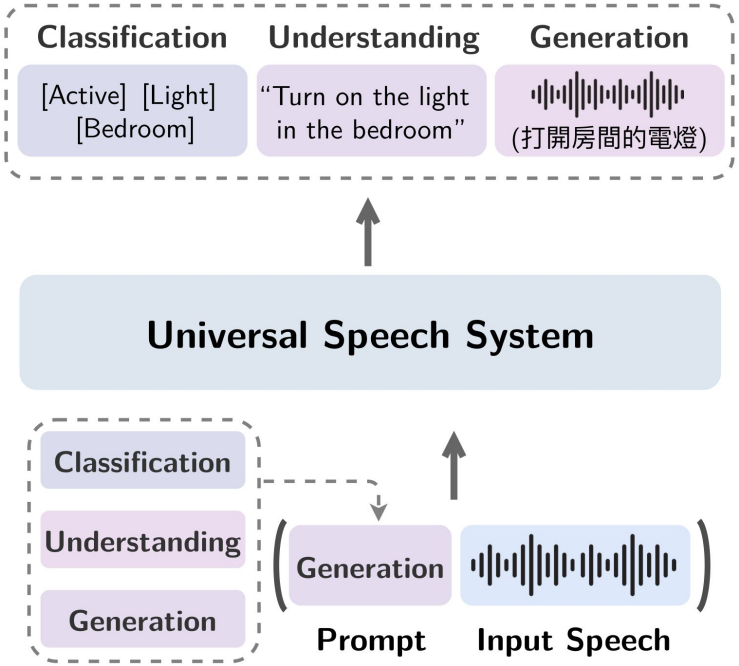
Thesis - *Towards a Universal Speech Model: Prompting Speech Language Models for Diverse Speech Processing Tasks*

About the Speaker

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

Chang+ [Interspeech 2022]
Chang+ [arXiv 2023]
Wu*, **Chang***, Wu*+ [arXiv 2023]
Chang+ [TASLP 2024]

About the Speaker

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

國立臺灣大學
National Taiwan University

2023 - Meta Reality Labs, Pittsburgh

Meta

REALITY LABS

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

About the Speaker

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

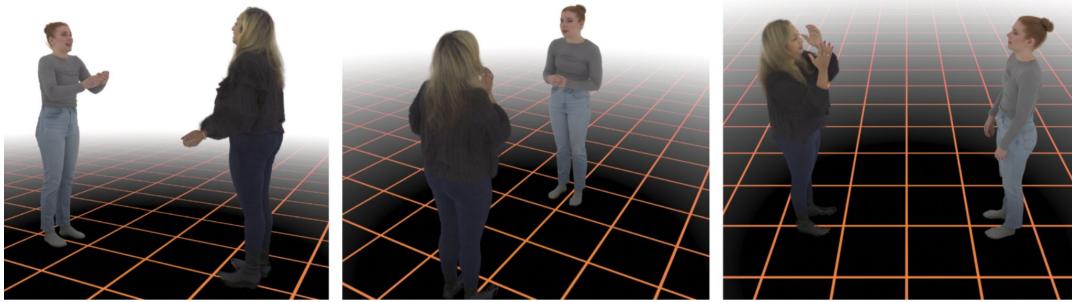
2023 - Meta Reality Labs, Pittsburgh

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

國立臺灣大學 National Taiwan University

Meta Reality Labs

Codec Avatar



Seamless Interaction dataset Agrawa+ [arXiv 2025]

About the Speaker

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

2023 - Meta Reality Labs, Pittsburgh

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

Now - Postdoc. at MIT CSAIL

國立臺灣大學
National Taiwan University

Meta

REALITY LABS

MIT CSAIL

About the Speaker

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

2023 - Meta Reality Labs, Pittsburgh

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

Now - Postdoc. at MIT CSAIL

Research Interest

(Full-duplex) Spoken Dialogue Model

E.g. Benchmark, Interpretability, Post-training

About the Speaker

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

2023 - Meta Reality Labs, Pittsburgh

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

Now - Postdoc. at MIT CSAIL

國立臺灣大學
National Taiwan University

Meta

REALITY LABS

MIT CSAIL

Research Interest

(Full-duplex) Spoken Dialogue Model

E.g. Benchmark, Interpretability, Post-training

Low-resource Speech Technologies

E.g. Taiwanese speech corpus –TaigiSpeech **Chang+** [Interspeech 2026]

Text assistant vs. Voice assistant

What’s the fundamental difference?

How's the weather today? → **Text Assistant** → It's sunny today


(How's the weather today?) → **Voice Assistant** → 
(It's sunny today.)

Emotion	**Non-verbal**	**Tone**
😡 😱 😢 😄	(sigh), (laughs), haha, lol	WHAT!! okay...?

Text can describe paralinguistics (to a certain degree).

Text assistant vs. Voice assistant

What’s the fundamental difference?

How's the weather today? → Text Assistant → It's sunny today

Time!

(How's the weather today?) → Voice Assistant → (It's sunny today.)

Emotion	Non-verbal	Tone
😠 😱 😢 😄	(sigh), (laughs), haha, lol	WHAT!! okay...?

Text can still describe paralinguistics to a certain degree.

How is “time” presented in speech conversation?

Text Conversation

Speech Conversation

Text Conversation

Turn-by-turn

Speech Conversation

Text Conversation

Speech Conversation

Turn-by-turn

Model’s turn

User’s turn

User’s turn

Model’s turn

Text Conversation

Speech Conversation

Turn-by-turn

Model's turn

User's turn

Model's turn

User's turn

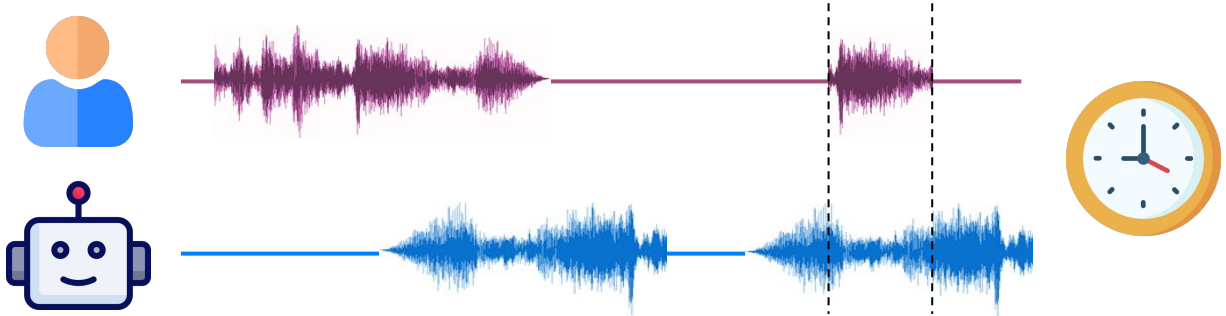
Usually, faster responses are better.

Text Conversation

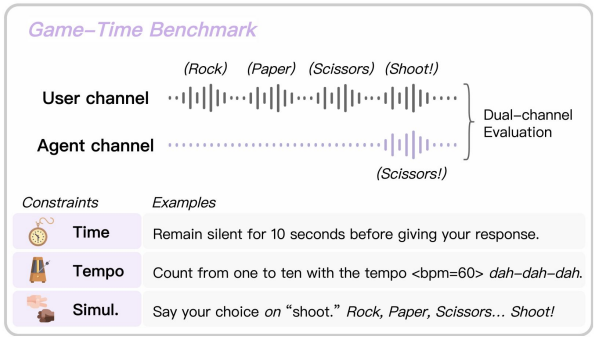
Speech Conversation

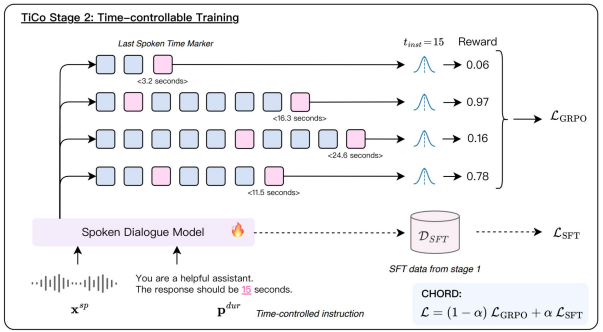
Full-duplex

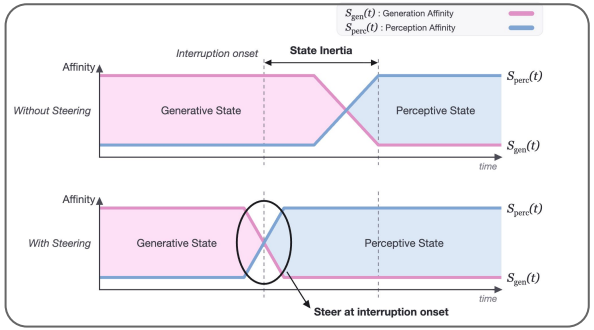
- Time keeps going forward
- At any time, both speakers
  - are listening and can speak
  - are able to overlap with each other
  - are making decisions



Outline







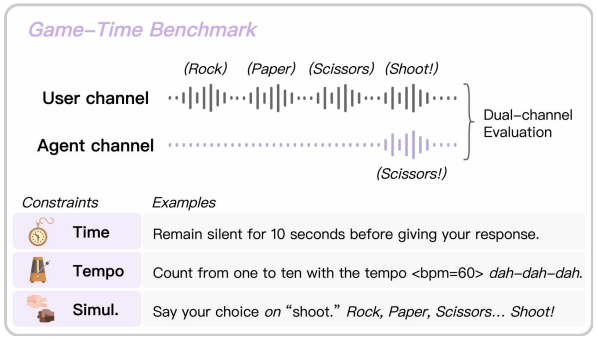
Game-Time → TiCo → State Inertia

Chang*, Hu*+ [ICASSP 2025]

Chang*, Chen*+ [arXiv 2026]

Chang, **Chang***+ [arXiv 2026]

Are current spoken dialogue models “time aware”?



Game-Time

Chang*, Hu*+ [ICASSP 2025]

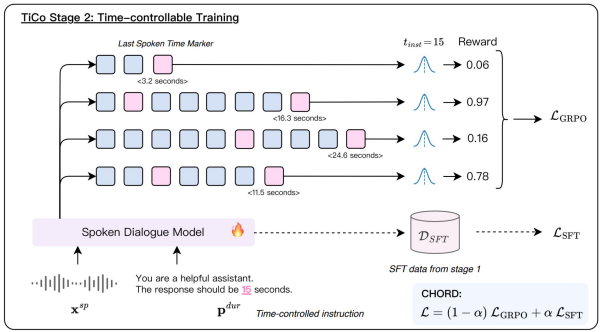
TiCo

Chang*, Chen*+ [arXiv 2026]

State Inertia

Chang, **Chang***+ [arXiv 2026]

Can we train spoken dialogue models to be “time aware”?



Game-Time

Chang*, Hu*+ [ICASSP 2025]

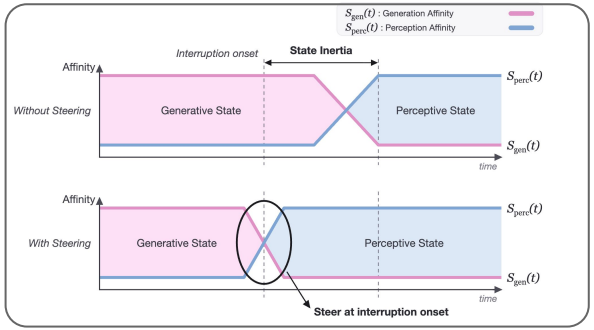
TiCo

Chang*, Chen*+ [arXiv 2026]

State Inertia

Chang, **Chang***+ [arXiv 2026]

How do full-duplex models coordinate listening and speaking?



Game-Time

Chang*, Hu*+ [ICASSP 2025]

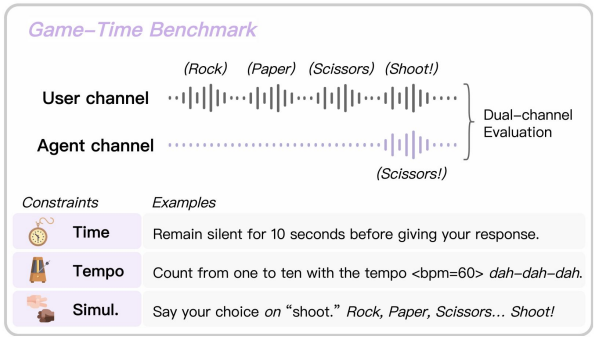
TiCo

Chang*, Chen*+ [arXiv 2026]

State Inertia

Chang, **Chang***+ [arXiv 2026]

Are current spoken dialogue models “time aware”?



Game-Time

Chang*, Hu*+ [ICASSP 2025]

TiCo

Chang*, Chen*+ [arXiv 2026]

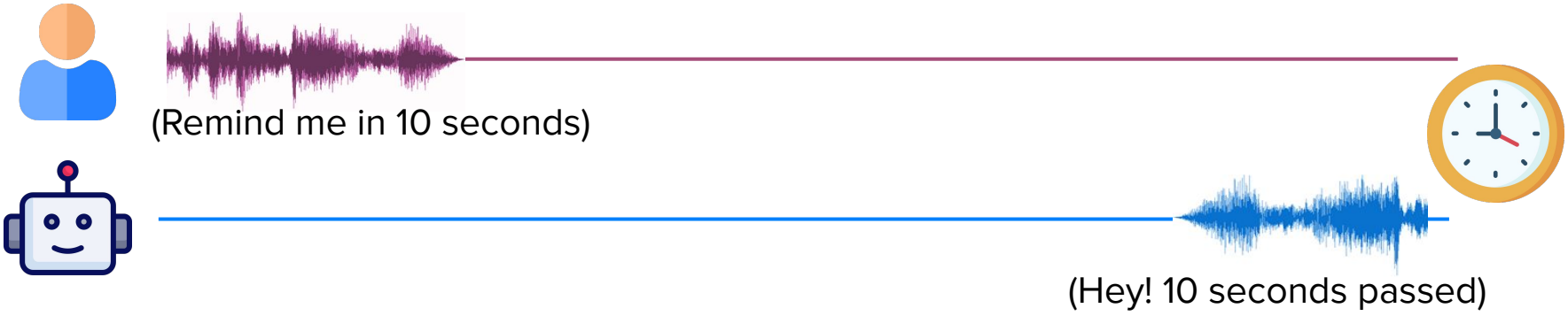
State Inertia

Chang, **Chang***+ [arXiv 2026]

Game-Time Benchmark

1. **Time awareness** - Knowing when and how long to respond
2. **Tempo** - Responding at a specified tempo
3. **Simul. Speaking** - Synchronizing and overlapping with the user

Time Awareness

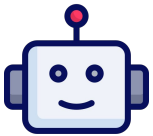


Game-Time Benchmark

1. **Time awareness** - Knowing when and how long to respond
2. **Tempo** - Responding at a specified tempo
3. **Simul. Speaking** - Synchronizing and overlapping with the user

Tempo

(Please count from one to five with the tempo one! two! three!)

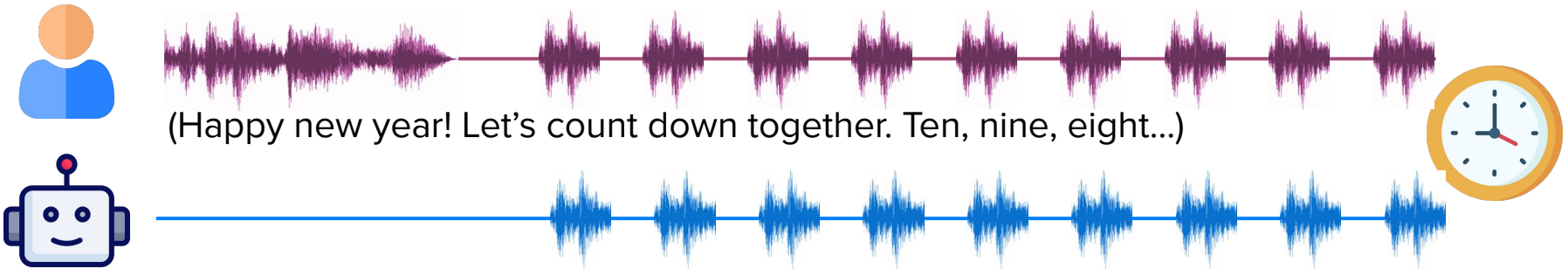


(one two three four five)

Game-Time Benchmark

1. **Time awareness** - Knowing when and how long to respond
2. **Tempo** - Responding at a specified tempo
3. **Simul. Speaking** - Synchronizing and overlapping with the user

Simul. Speaking (Overlapping)



Game-Time Benchmark

- **Time awareness** - Knowing when and how long to respond
- Tem
- Sim

Simul. Sp
(Overlap

(Happy new year! Let's count down together. Ten, nine, eight...)

Game-Time Inspiration

- How we learn a language in childhood
- Through playing games!

Game-Time Benchmark

Instruction following paradigm

Basic Tasks

Simple instructions

(Please count from 1 to 10)

Advanced Tasks

Basic tasks with time-related constraints

(Please count from 1 to 10 **in 5 seconds**)

Game-Time Basic Tasks

Task Family	Description	Example
1 - Sequence	Number / alphabet sequencing	“Please count from 3 to 15.”
2 - Repeat	Sentence repetition	“Please repeat after me. ‘I have a pen.’”
3 - Compose	Sentence composition from words	“Can you make a sentence with the word ‘dog’?”
4 - Recall	Vocabulary, letter, and rhyme recall	“Name 5 different colors.”
5 - Open	QA and empathy conversation	“What do you think makes a great movie?”
6 - Role-Play	Scenario and persona role-playing	“Imagine you're a psychologist, explaining anxiety.”

Game-Time Basic Tasks

Task Family	Description	Example
1 - Sequence	Number / alphabet sequencing	“Please count from 3 to 15.”
2 - Repeat	Sentence repetition	“Please repeat after me. ‘I have a pen.’”
3 - Compose	Sentence composition from words	“Can you make a sentence with the word ‘dog’?”
4 - Recall	Vocabulary, letter, and rhyme recall	“Name 5 different colors.”
5 - Open	QA and empathy conversation	“What do you think makes a great movie?”
6 - Role-Play	Scenario and persona role-playing	“Imagine you're a psychologist, explaining anxiety.”

Game-Time Basic Tasks

Task Family	Description	Example
1 - Sequence	Number / alphabet sequencing	“Please count from 3 to 15.”
2 - Repeat	Sentence repetition	“Please repeat after me. ‘I have a pen.’”
3 - Compose	Sentence composition from words	“Can you make a sentence with the word ‘dog’?”
4 - Recall	Vocabulary, letter, and rhyme recall	“Name 5 different colors.”
5 - Open	QA and empathy conversation	“What do you think makes a great movie?”
6 - Role-Play	Scenario and persona role-playing	“Imagine you're a psychologist, explaining anxiety.”

Game-Time Basic Tasks

Task Family	Description	Example
1 - Sequence	Number / alphabet sequencing	“Please count from 3 to 15.”
2 - Repeat	Sentence repetition	“Please repeat after me. ‘I have a pen.’”
3 - Compose	Sentence composition from words	“Can you make a sentence with the word ‘dog’?”
4 - Recall	Vocabulary, letter, and rhyme recall	“Name 5 different colors.”
5 - Open	QA and empathy conversation	“What do you think makes a great movie?”
6 - Role-Play	Scenario and persona role-playing	“Imagine you're a psychologist, explaining anxiety.”

Game-Time Basic Tasks

Task Family	Description	Example
1 - Sequence	Number / alphabet sequencing	“Please count from 3 to 15.”
2 - Repeat	Sentence repetition	“Please repeat after me. ‘I have a pen.’”
3 - Compose	Sentence composition from words	“Can you make a sentence with the word ‘dog’?”
4 - Recall	Vocabulary, letter, and rhyme recall	“Name 5 different colors.”
5 - Open	QA and empathy conversation	“What do you think makes a great movie?”
6 - Role-Play	Scenario and persona role-playing	“Imagine you're a psychologist, explaining anxiety.”

Game-Time Basic Tasks

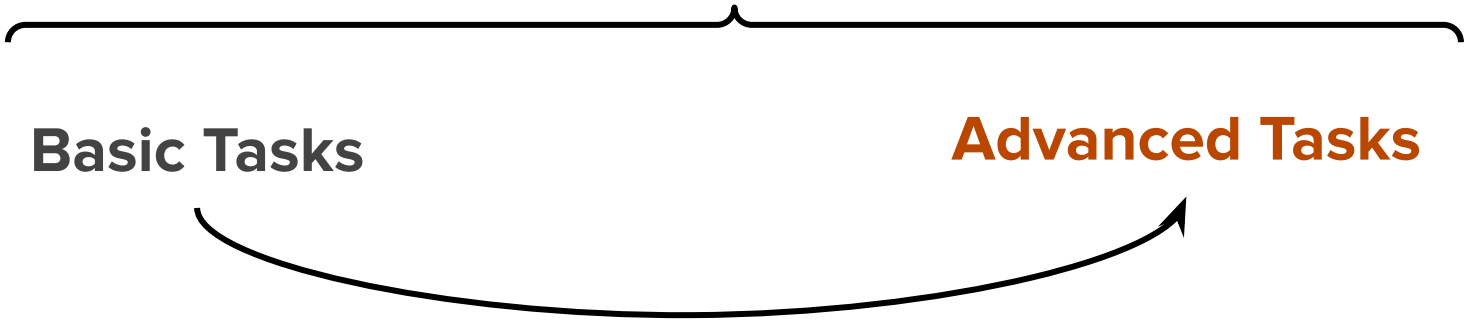
Task Family	Description	Example
1 - Sequence	Number / alphabet sequencing	“Please count from 3 to 15.”
2 - Repeat	Sentence repetition	“Please repeat after me. ‘I have a pen.’”
3 - Compose	Sentence composition from words	“Can you make a sentence with the word ‘dog’?”
4 - Recall	Vocabulary, letter, and rhyme recall	“Name 5 different colors.”
5 - Open	QA and empathy conversation	“What do you think makes a great movie?”
6 - Role-Play	Scenario and persona role-playing	“Imagine you're a psychologist, explaining anxiety.”

Game-Time Basic Tasks

Task Family	Description	Example
1 - Sequence	Number / alphabet sequencing	“Please count from 3 to 15.”
2 - Repeat	Sentence repetition	“Please repeat after me. ‘I have a pen.’”
3 - Compose	Sentence composition from words	“Can you make a sentence with the word ‘dog’?”
4 - Recall	Vocabulary, letter, and rhyme recall	“Name 5 different colors.”
5 - Open	QA and empathy conversation	“What do you think makes a great movie?”
6 - Role-Play	Scenario and persona role-playing	“Imagine you're a psychologist, explaining anxiety.”

Game-Time Benchmark

Instruction following paradigm



*Augment with **time-related constraints!***

Game-Time Benchmark

*Augment with **time-related constraints!***

- “Please count from 1 to 10”
  - longer than 20 seconds
  - shorter than 5 seconds
- “Repeat the following sentence: I have a pen”
  - Remain silent for 10 seconds, then respond
  - Repeat right after me (shadowing)

Game-Time Advanced Tasks

Constrained Type	Example
A - Time Fast	“Please count from 3 to 15 within 4 seconds”
B - Time Slow	“Please count from 3 to 15 lasting more than 10 seconds”
C - Time Silence	“Please remain silent for 10 seconds then count from 3 to 15”
D - Tempo Interval	“Please count from 3 to 15, pausing a 1 second gap between each number”
E - Tempo Adhere	“Please count from 3 to 15, with the tempo [bpm 100: one, two, three, four]”
F - Simul.Speak Shadow	“Please repeat each word right after me. Here we go: ‘I have a pen.’”
G - Simul.Speak Cue	“Let’s play rock-paper-scissors, Say ‘rock’, ‘paper’, or ‘scissors’ when I say ‘shoot’. .... ‘rock’, ‘paper’, ‘scissors’ shoot”

Game-Time Advanced Tasks

Constrained Type	Example
A - Time Fast	“Please count from 3 to 15 within 4 seconds”
B - Time Slow	“Please count from 3 to 15 lasting more than 10 seconds”
C - Time Silence	“Please remain silent for 10 seconds then count from 3 to 15”
D - Tempo Interval	“Please count from 3 to 15, pausing a 1 second gap between each number”
E - Tempo Adhere	“Please count from 3 to 15, with the tempo [bpm 100: one, two, three, four]”
F - Simul.Speak Shadow	“Please repeat each word right after me. Here we go: ‘I have a pen.’”
G - Simul.Speak Cue	“Let’s play rock-paper-scissors, Say ‘rock’, ‘paper’, or ‘scissors’ when I say ‘shoot’. .... ‘rock’, ‘paper’, ‘scissors’ shoot”

Game-Time Advanced Tasks

Constrained Type	Example
A - Time Fast	“Please count from 3 to 15 within 4 seconds”
B - Time Slow	“Please count from 3 to 15 lasting more than 10 seconds”
C - Time Silence	“Please remain silent for 10 seconds then count from 3 to 15”
D - Tempo Interval	“Please count from 3 to 15, pausing a 1 second gap between each number”
E - Tempo Adhere	“Please count from 3 to 15, with the tempo [bpm 100: one, two, three, four]”
F - Simul.Speak Shadow	“Please repeat each word right after me. Here we go: ‘I have a pen.’”
G - Simul.Speak Cue	“Let’s play rock-paper-scissors, Say ‘rock’, ‘paper’, or ‘scissors’ when I say ‘shoot’. .... ‘rock’, ‘paper’, ‘scissors’ shoot”

Game-Time Advanced Tasks

Constrained Type	Example
A - Time Fast	“Please count from 3 to 15 within 4 seconds”
B - Time Slow	“Please count from 3 to 15 lasting more than 10 seconds”
C - Time Silence	“Please remain silent for 10 seconds then count from 3 to 15”
D - Tempo Interval	“Please count from 3 to 15, pausing a 1 second gap between each number”
E - Tempo Adhere	“Please count from 3 to 15, with the tempo [bpm 100: one, two, three, four]”
F - Simul.Speak Shadow	“Please repeat each word right after me. Here we go: ‘I have a pen.’”
G - Simul.Speak Cue	“Let’s play rock-paper-scissors, Say ‘rock’, ‘paper’, or ‘scissors’ when I say ‘shoot’. .... ‘rock’, ‘paper’, ‘scissors’ shoot”

Basic Tasks (14 subtasks)

1-Sequence (3 subtasks)

- 1-a Sequence-Number
- 1-b Sequence-Alphabet
- 1-c Sequence-Spell

2-Repeat (2 subtasks)

- 2-a Repeat-Word
- 2-b Repeat-Sentence

3-Compose (2 subtasks)

- 3-a Compose-Word
- 3-b Compose-Scenario

4-Recall (3 subtasks)

- 4-a Recall-Vocabulary
- 4-b Recall-Letter
- 4-c Recall-Rhyme

5-Open-Ended (2 subtasks)

- 5-a OpenEnded-QA
- 5-b OpenEnded-Empathy

6-Role-Play (2 subtasks)

- 6-a RolePlay-Scenario
- 6-b RolePlay-Persona

Advanced Tasks (31 subtasks)

A-Time-Fast (10 subtasks)

- 1-a Sequence-Number-TimeFast
- 1-c Sequence-Spell-TimeFast
- 3-a Compose-Word-TimeFast
- 3-b Compose-Scenario-TimeFast
- 4-a Recall-Vocabulary-TimeFast
- 4-b Recall-Letter-TimeFast
- 5-a OpenEnded-QA-TimeFast
- 5-b OpenEnded-Empathy-TimeFast
- 6-a RolePlay-Scenario-TimeFast
- 6-b RolePlay-Persona-TimeFast

B-Time-Slow (7 subtasks)

- 1-a Sequence-Number-TimeSlow
- 3-a Compose-Word-TimeSlow
- 3-b Compose-Scenario-TimeSlow
- 5-a OpenEnded-QA-TimeSlow
- 5-b OpenEnded-Empathy-TimeSlow
- 6-a RolePlay-Scenario-TimeSlow
- 6-b RolePlay-Persona-TimeSlow

C-Time-Silence (4 subtasks)

- 2-b Repeat-Sentence-TimeSilence
- 4-a Recall-Vocabulary-TimeSilence
- 4-b Recall-Letter-TimeSilence
- 5-b OpenEnded-Empathy-TimeSilence

D-Tempo-Interval (4 subtasks)

- 1-a Sequence-Number-TempoInterval
- 1-c Sequence-Spell-TempoInterval
- 4-a Recall-Vocabulary-TempoInterval
- 4-c Recall-Rhyme-TempoInterval

E-Tempo-Adhere (4 subtasks)

- 1-a Sequence-Number-TempoAdhere
- 1-c Sequence-Spell-TempoAdhere
- 4-a Recall-Vocabulary-TempoAdhere
- 4-c Recall-Rhyme-TempoAdhere

F-SimulSpeak-Shadow (1 subtask)

- 2-b Repeat-Sentence-SimulSpeakShadow

G-SimulSpeak-Cue (1 subtask)

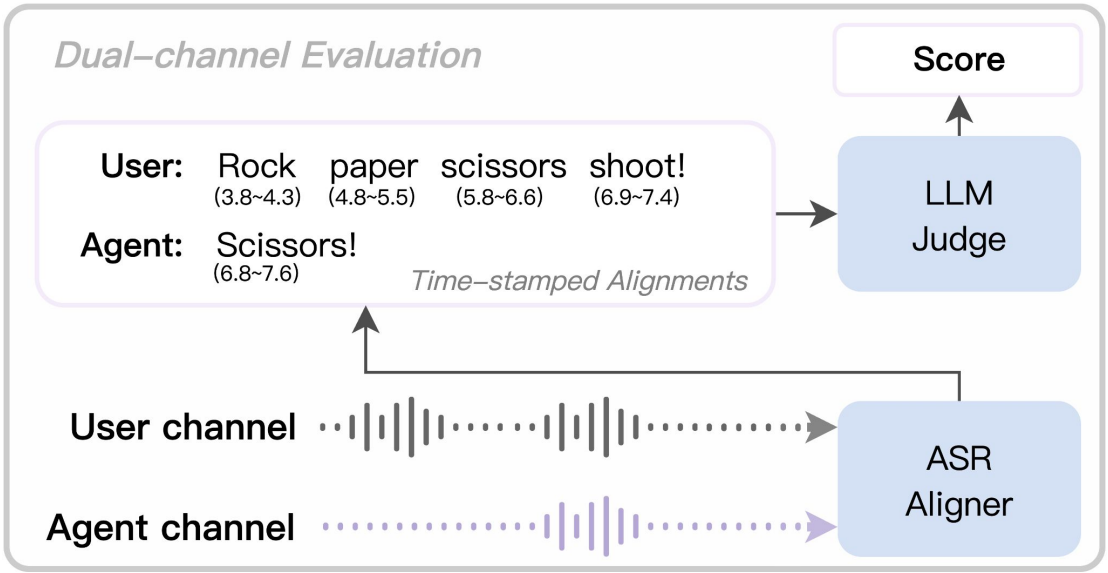
- 7-a Game-RockPaperScissors-SimulSpeakCue

Basic Tasks: 700 queries

Advanced Tasks: 775 queries

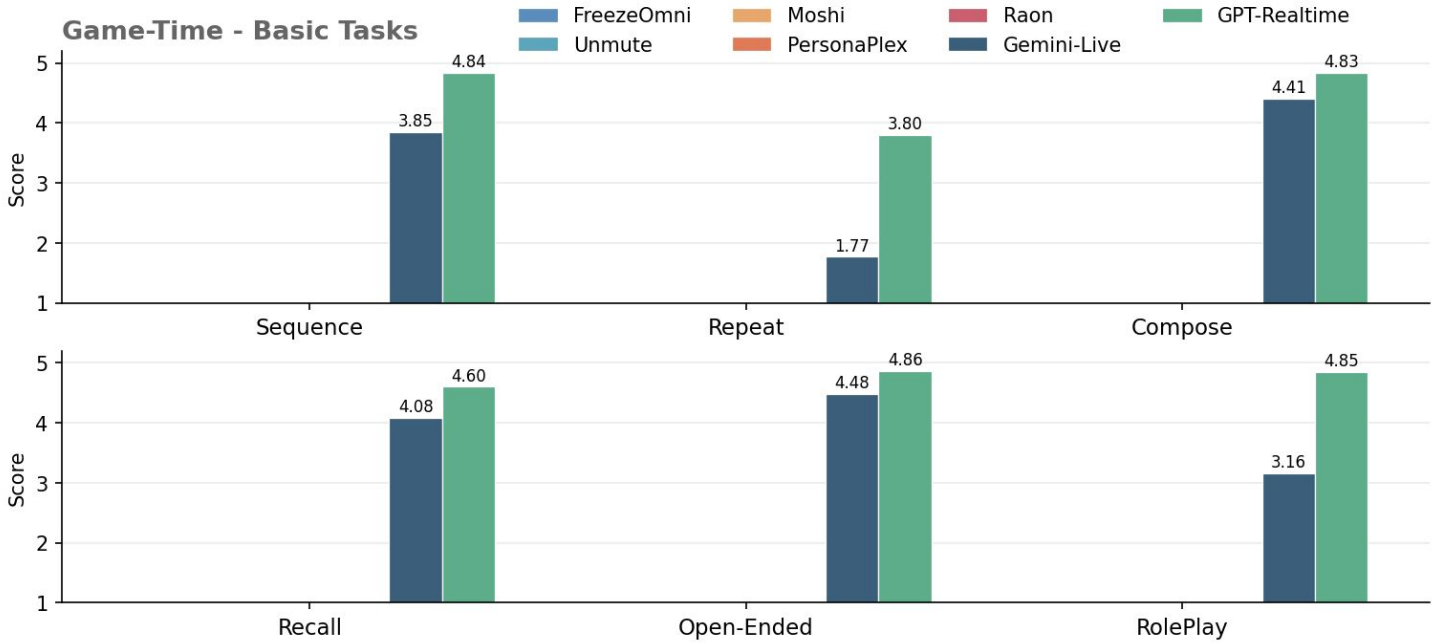
Evaluation

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



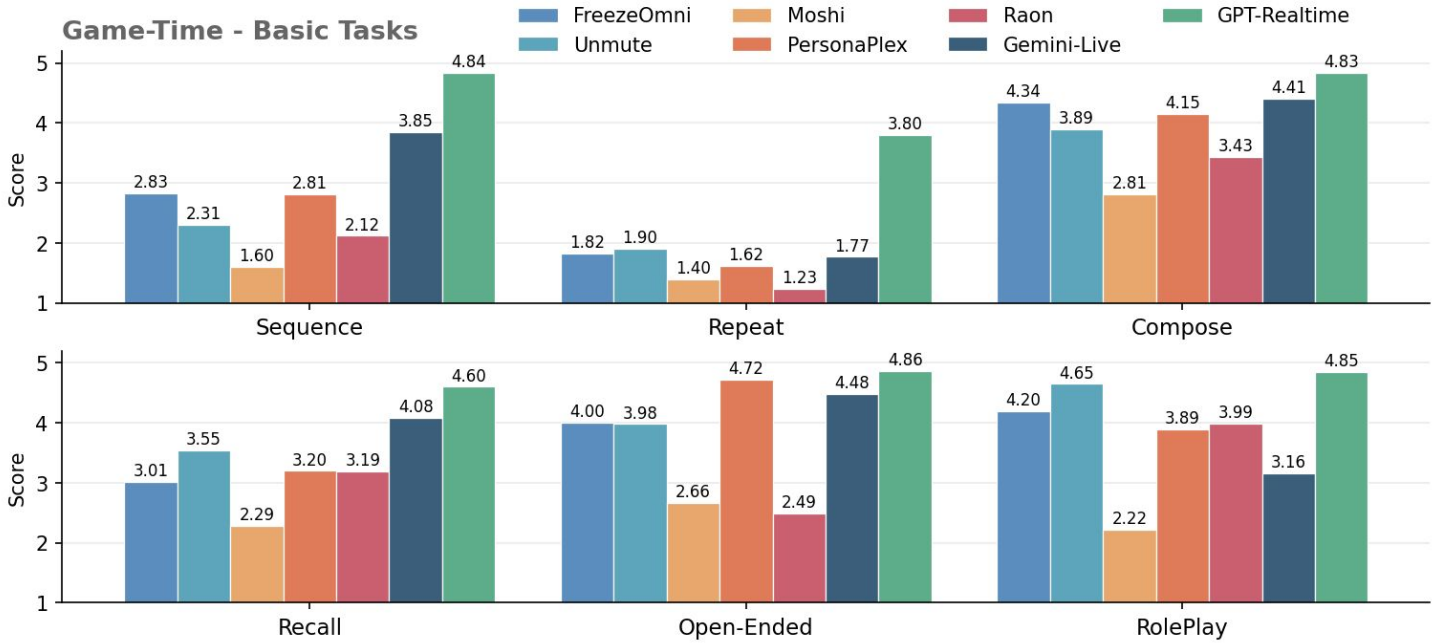
Gemini 2.5 Pro

Game-Time Basic Tasks Results



Commercial API performs well on basic tasks

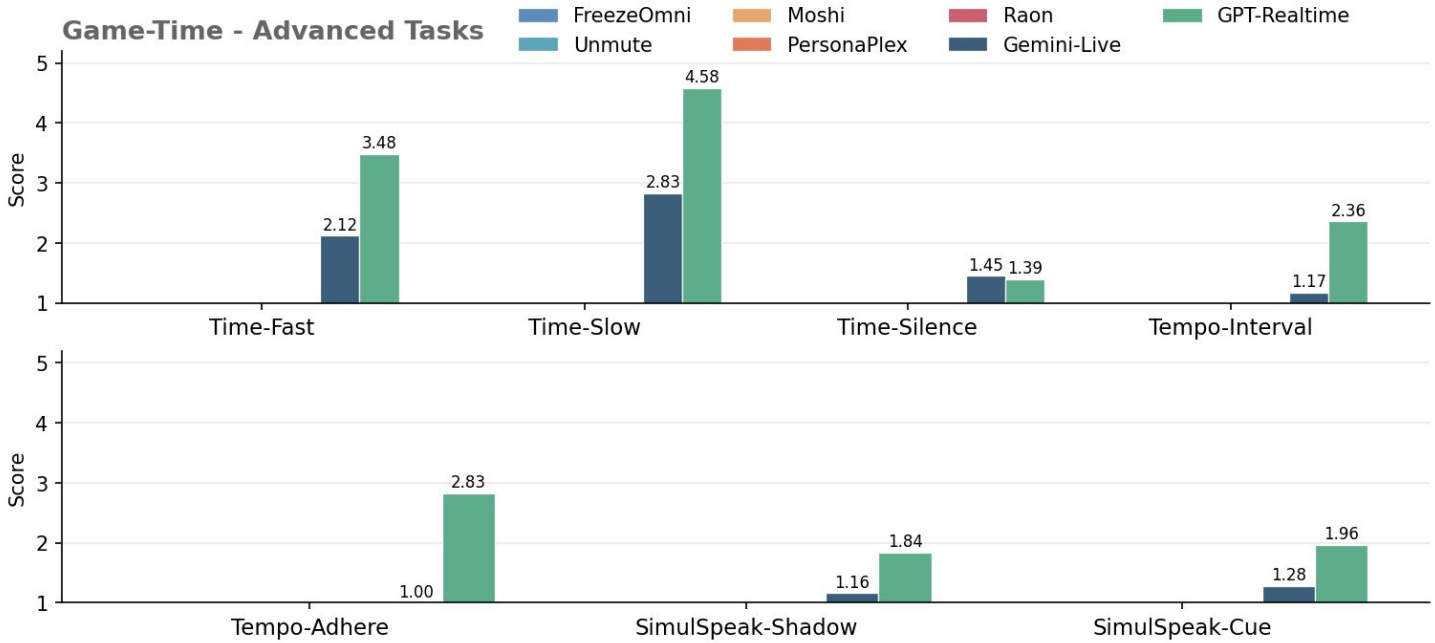
Game-Time Basic Tasks Results



Open-sourced models < Commercial API

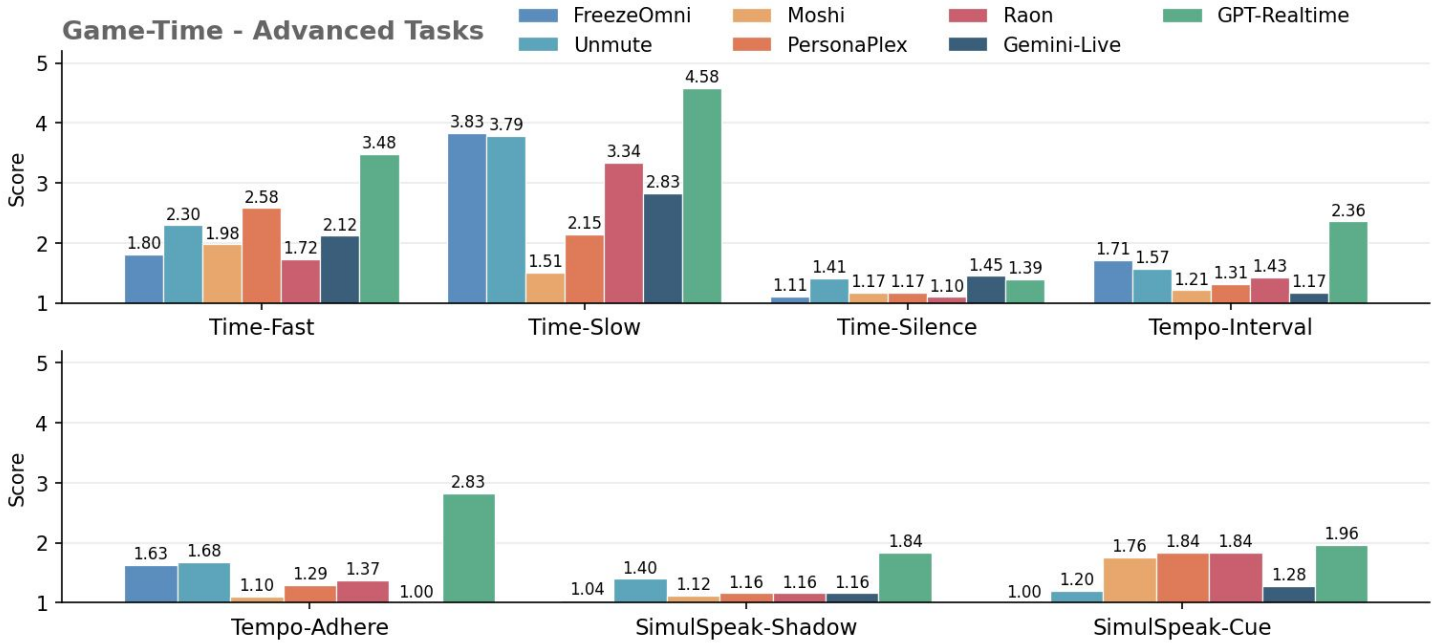
Wang+ [PMLR 2025]
D´efossez*, Mazar´e* [arXiv 2024]
KRAFTON [arXiv 2026]
Zeghidour+ [arXiv 2025]
Roy+ [ICASSP 2026]

Game-Time Advanced Tasks Results



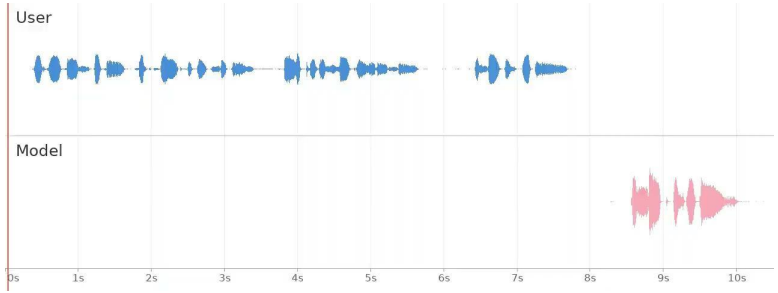
After introducing time related constraint, the model degrades a lot

Game-Time Advanced Tasks Results

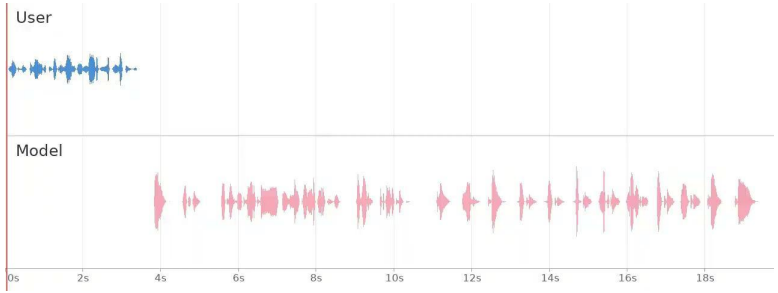


After introducing time related constraint, the model degrades a lot

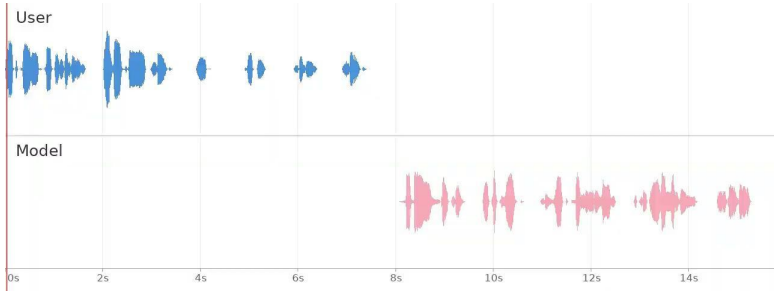
TimeSilence



TimeFast



SimulSpeak-Cue



Model:
GPT-realtime

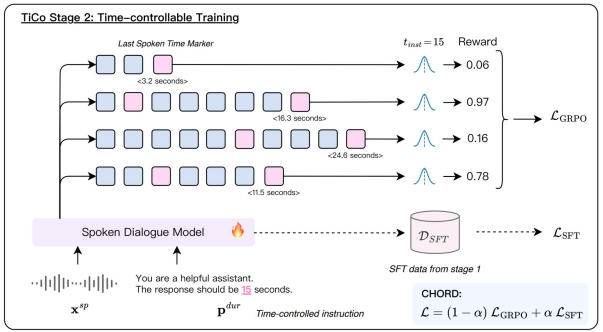
Game-Time Benchmark: Takeaway

- Most speech models can excel on basic tasks
  * However, some models (e.g. Moshi) still cannot

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

- Speech models cannot perform on some tasks that are easy to human!

Train spoken dialogue models to be “time aware”? – An attempt



Game-Time

Chang*, Hu*+ [ICASSP 2025]

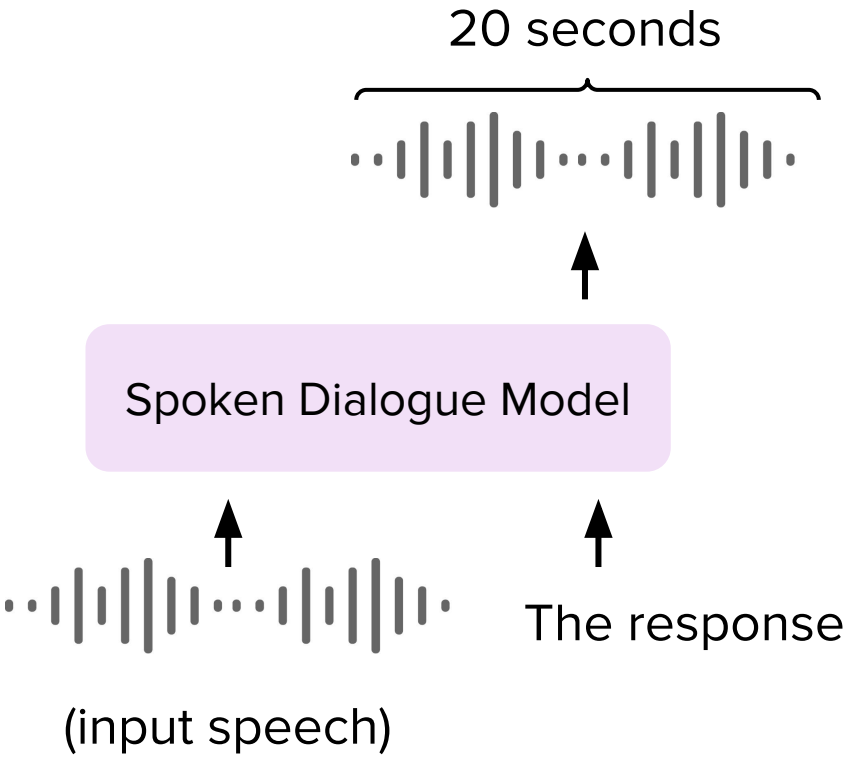
TiCo

Chang*, Chen*+ [arXiv 2026]

State Inertia

Chang, **Chang***+ [arXiv 2026]

Duration Controllability



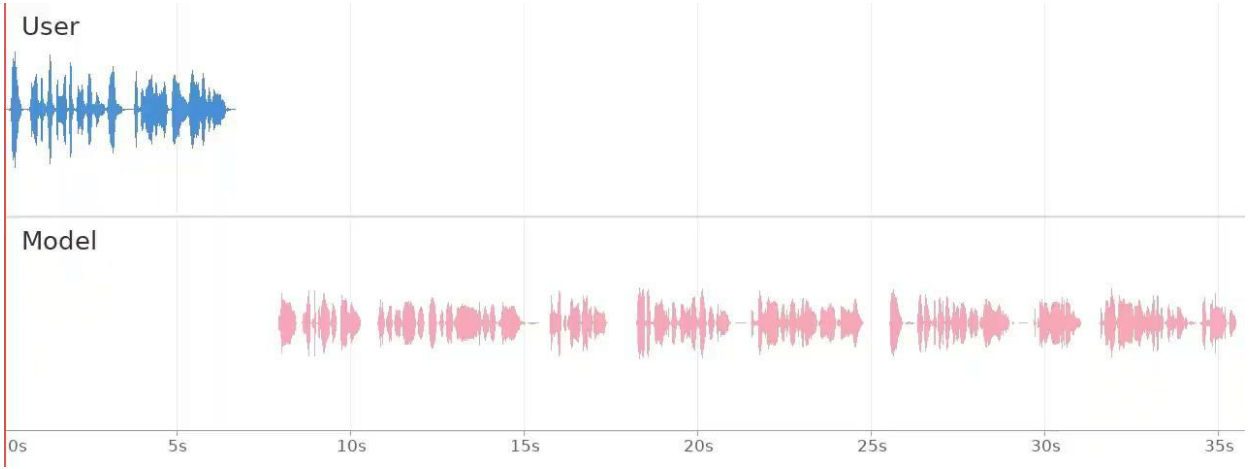
- Semantic planning
- Speaking rate control

TiCo
half-duplex models
Semanic planning

Target duration:
20 seconds
(text instruction)

Model:
Qwen-2.5 Omni

Qwen Team [arXiv 2025]

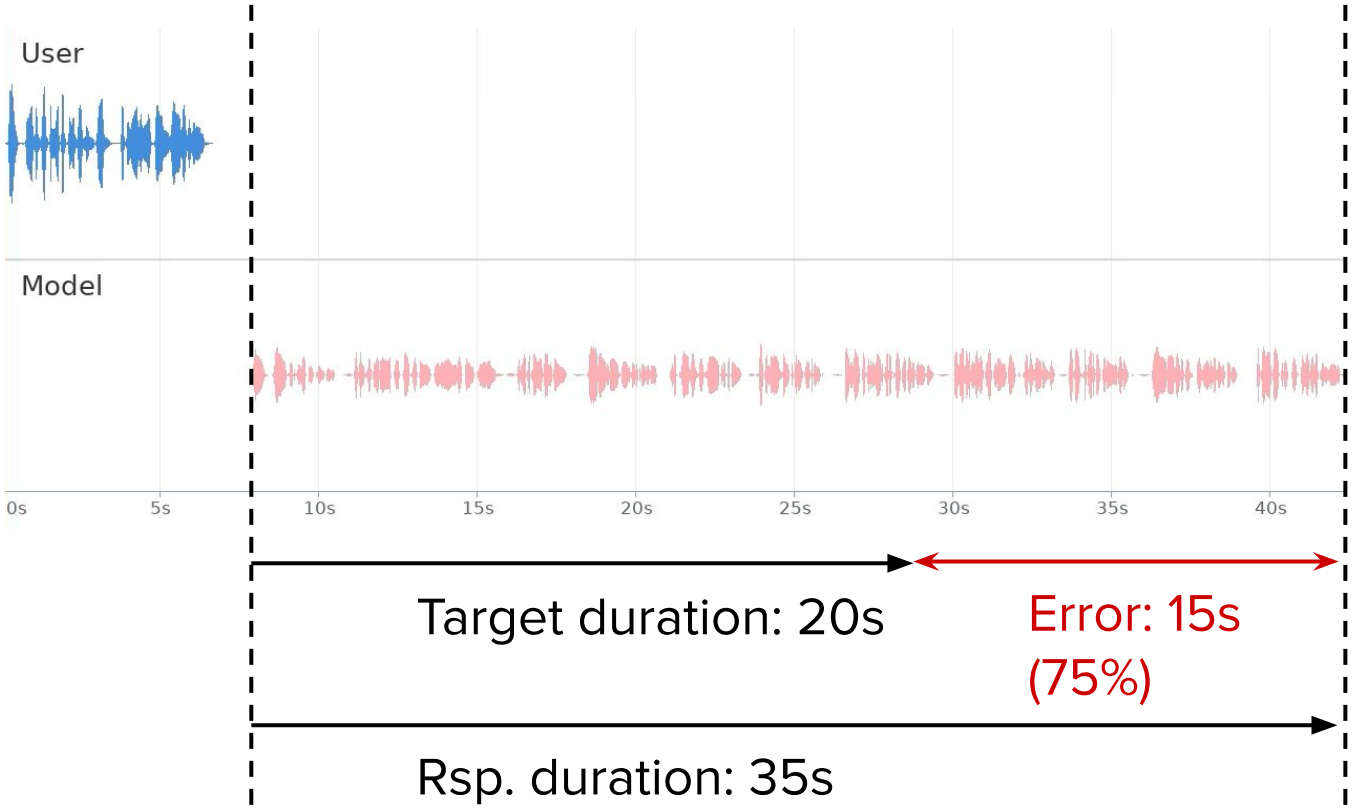


User: Can you suggest a method for reducing stress?
Model: Well, you could try deep breathing....

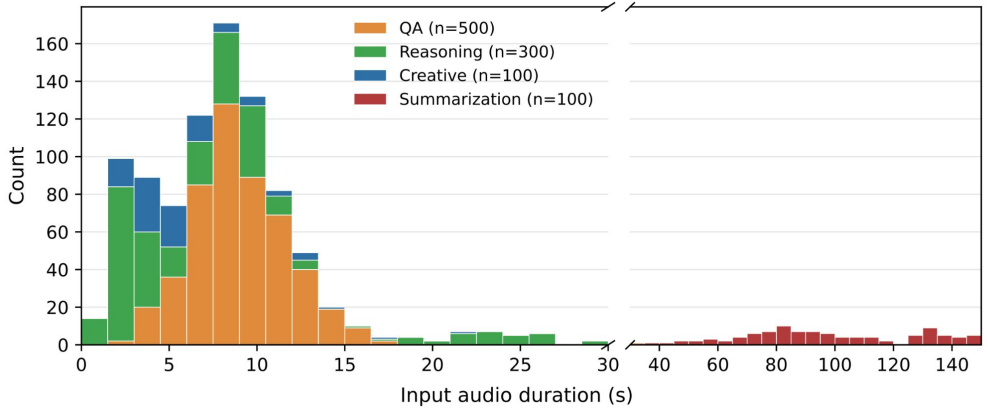
Current spoken dialogue models can not follow duration-constrained instructions

Target duration:
20 seconds
(text instruction)

Model:
Qwen-2.5 Omni

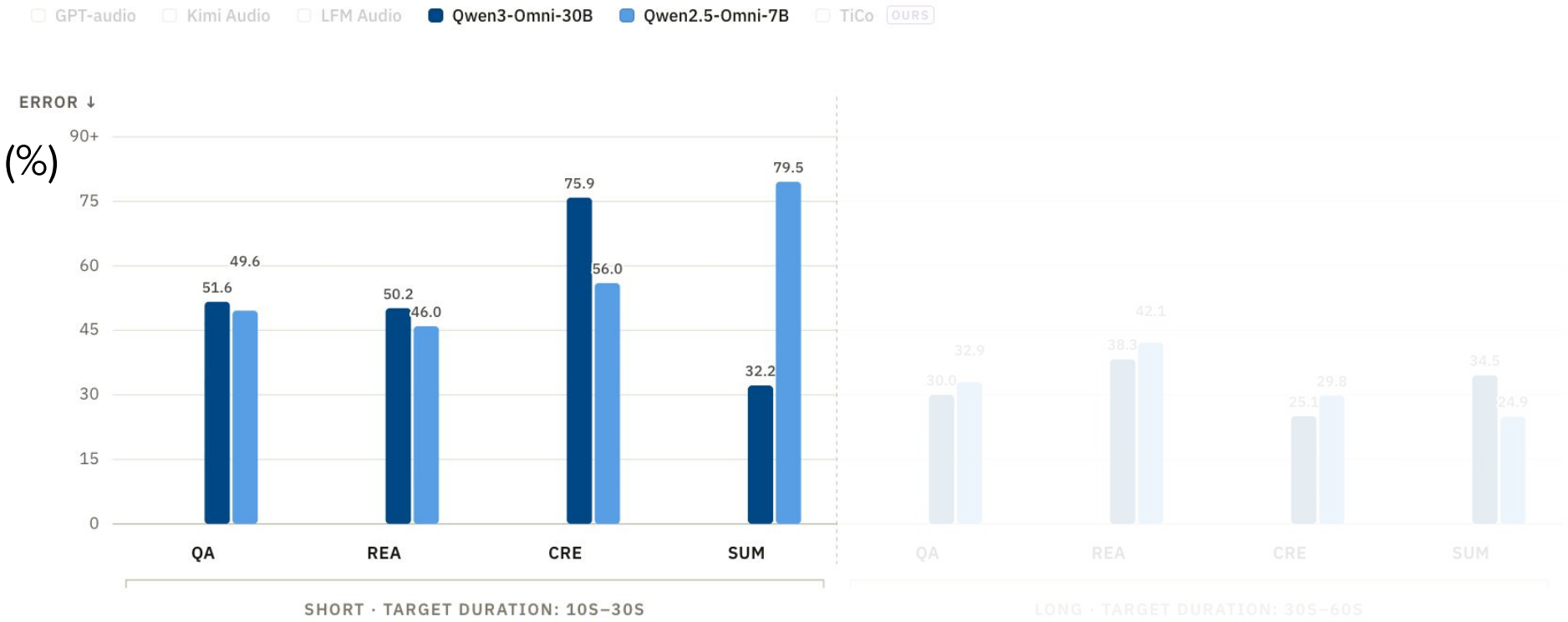


TiCo-Bench



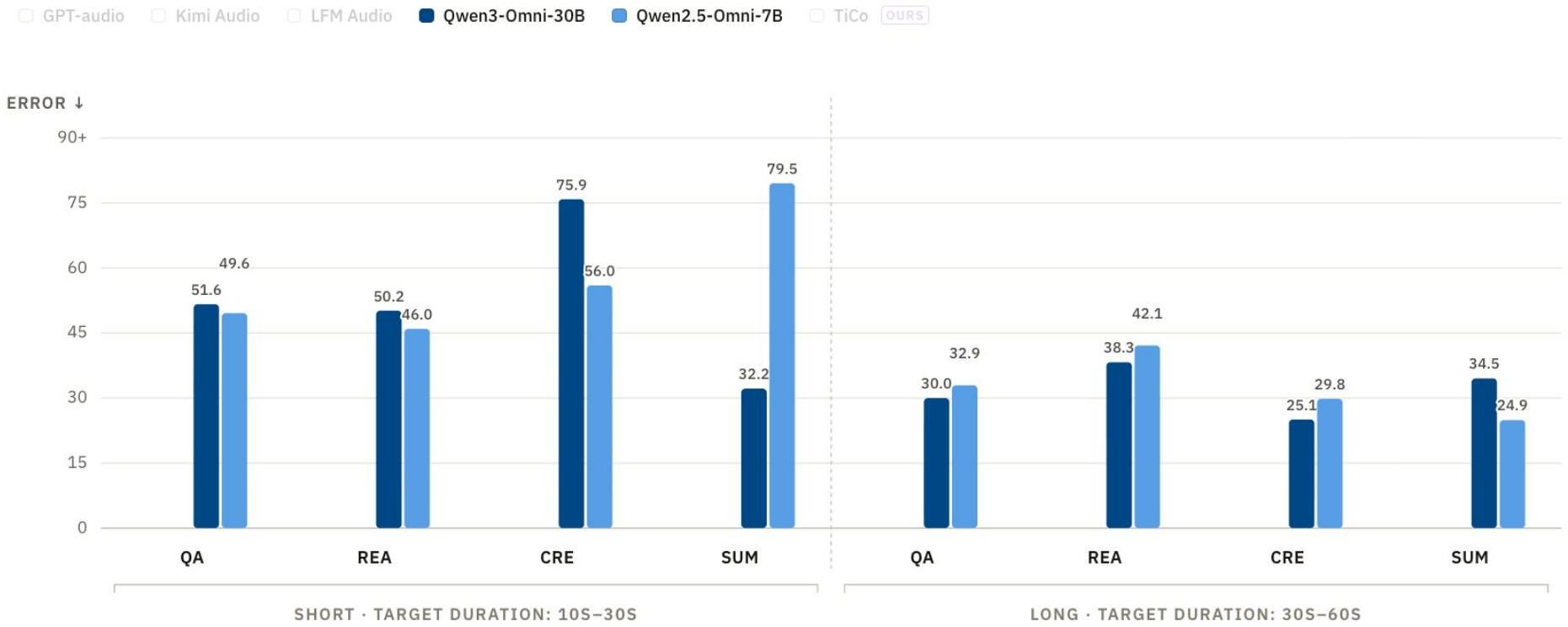
500 QA
300 Reasoning (REA)
100 Creative (CRE)
100 Summarization (SUM)

duration-controlled instructions
(e.g. The response should be 20 seconds)

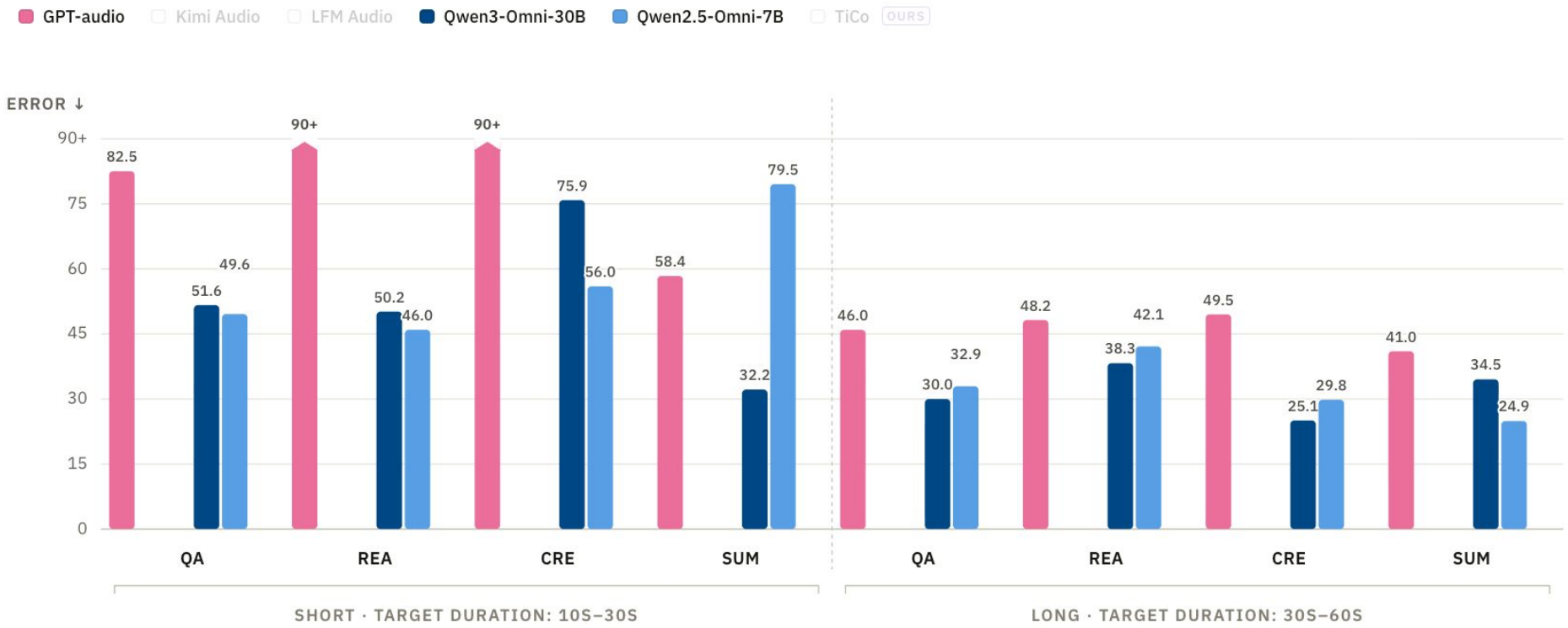


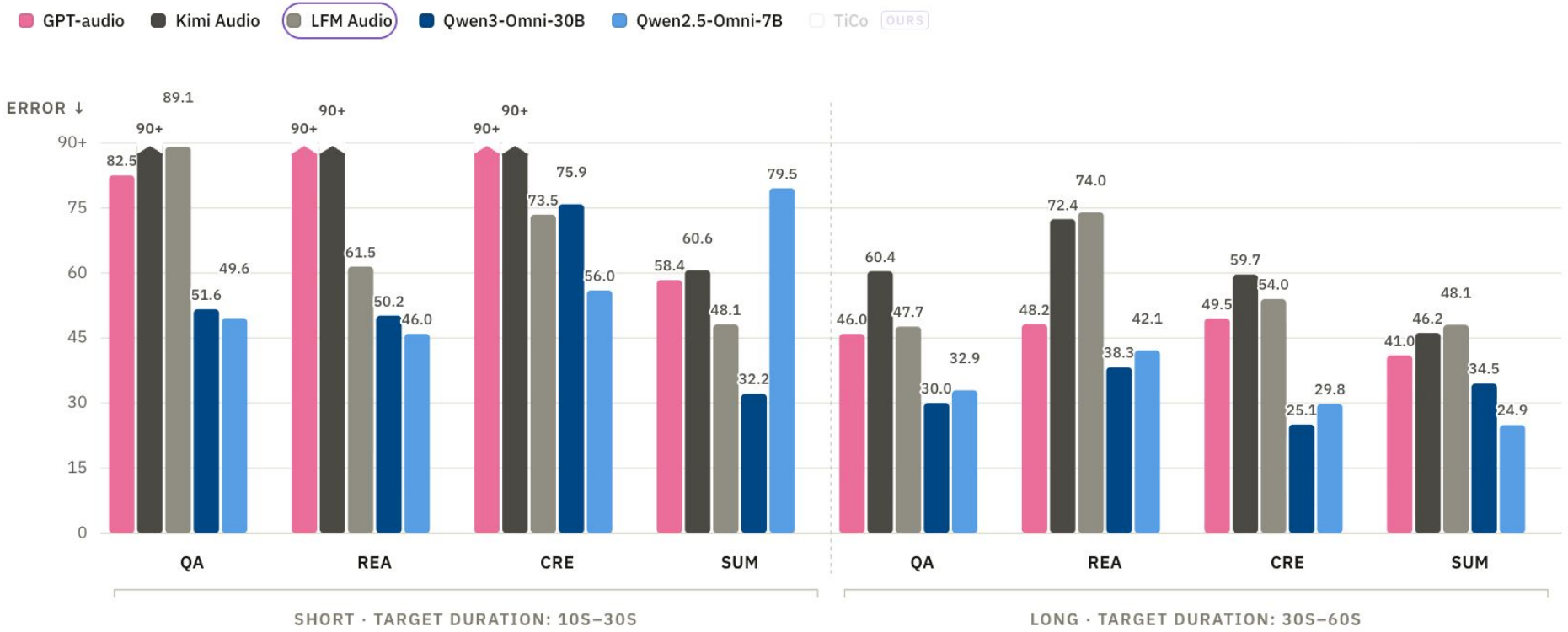
Qwen Team [arXiv 2025]

Qwen Team [arXiv 2025]



Tend to generate longer responses


GPT-audio
Kimi Audio
LFM Audio
Qwen3-Omni-30B
Qwen2.5-Omni-7B
TiCo
OURS
ERROR ↓
90+
75
60
45
30
15
0
82.5
51.6
49.6
90+
50.2
46.0
90+
75.9
56.0
58.4
32.2
79.5
46.0
30.0
32.9
48.2
38.3
42.1
49.5
25.1
29.8
41.0
34.5
24.9
QA
REA
CRE
SUM
QA
REA
CRE
SUM
SHORT · TARGET DURATION: 10S–30S
LONG · TARGET DURATION: 30S–60S



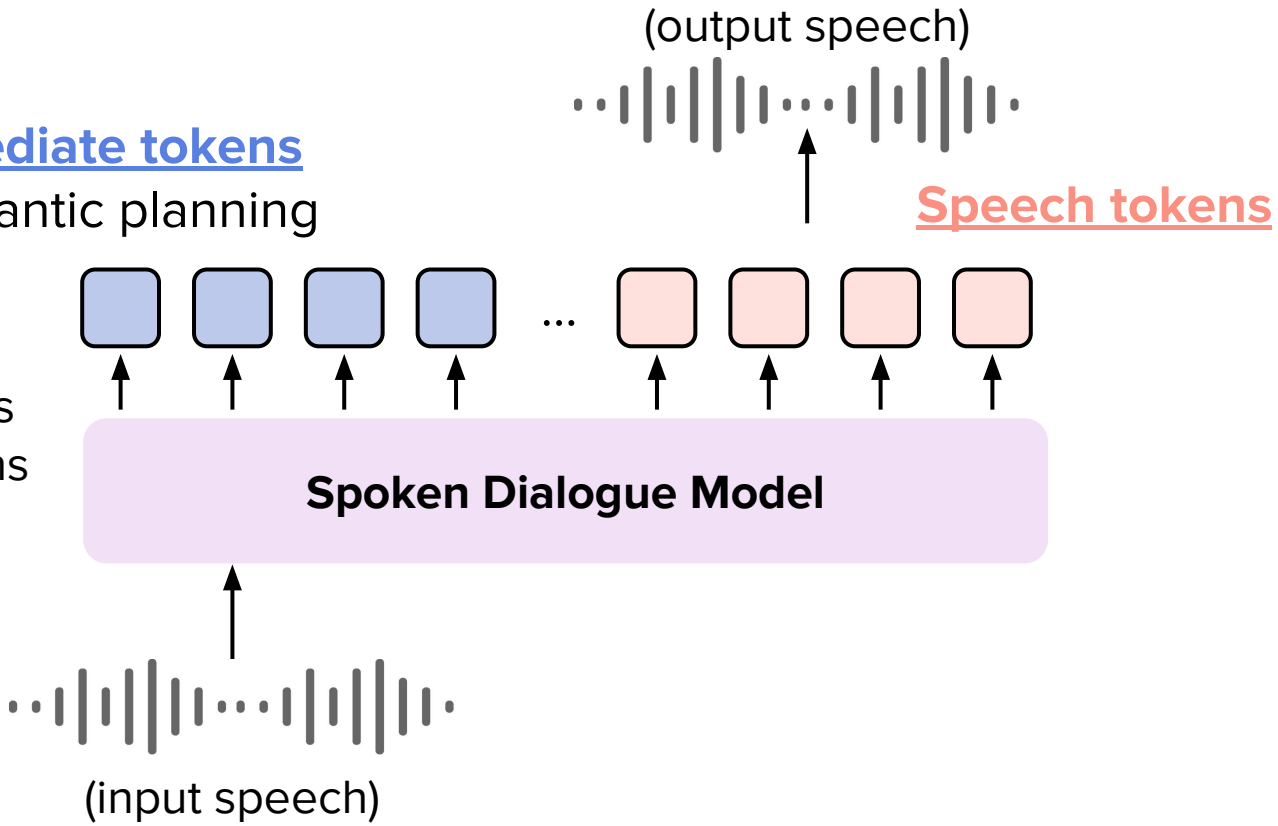
Off-the-shelf speech models cannot perform well
can not follow duration instructions

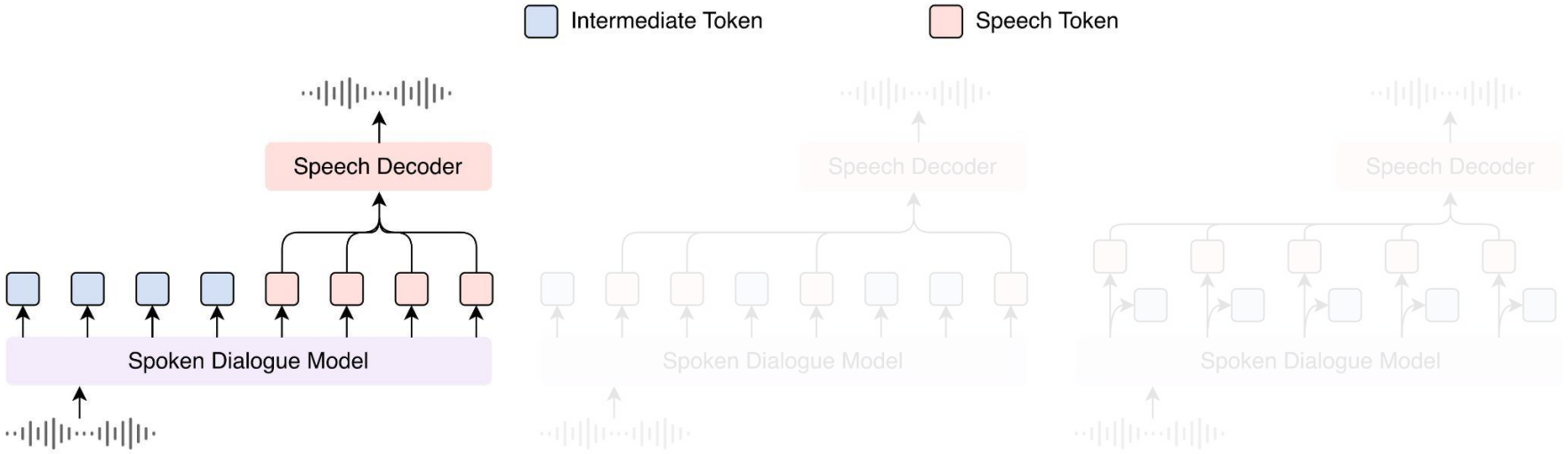
Kimi Team [arXiv 2025]
Liquid AI Team [arXiv 2025]

If we look inside today's Spoken Dialogue Models

- Text tokens (to be spoken)
- Reasoning tokens
- Tool calling tokens

...





(a) Sequential

- Qwen-Omni
- Freeze Omni

Qwen Team [arXiv 2025]
Wang+ [PMLR 2025]

(b) Interleaving

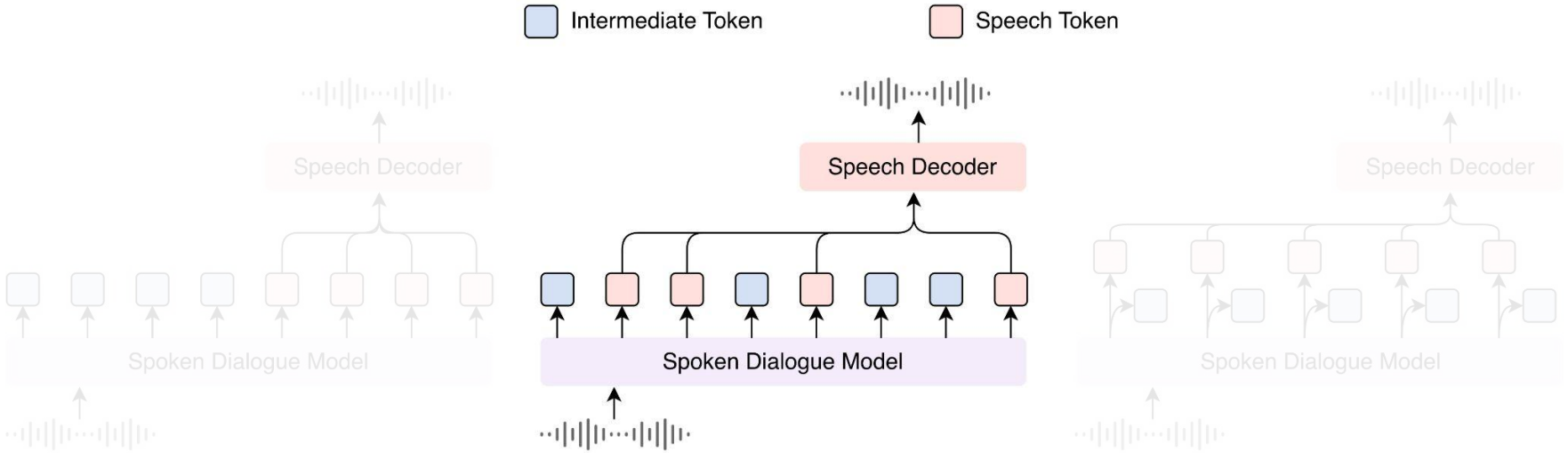
- Step-Audio
- LFM2-Audio

StepFun Audio Team [axXiv 2025]
Liquid AI Team [arXiv 2025]

(c) Parallel

- LLaMA-Omni
- Moshi

Fang+ [ICLR 2025]
D´efossez*, Mazar´e* [arXiv 2024]



(a) Sequential

- Qwen-Omni
- Freeze Omni

Qwen Team [arXiv 2025]
Wang+ [PMLR 2025]

(b) Interleaving

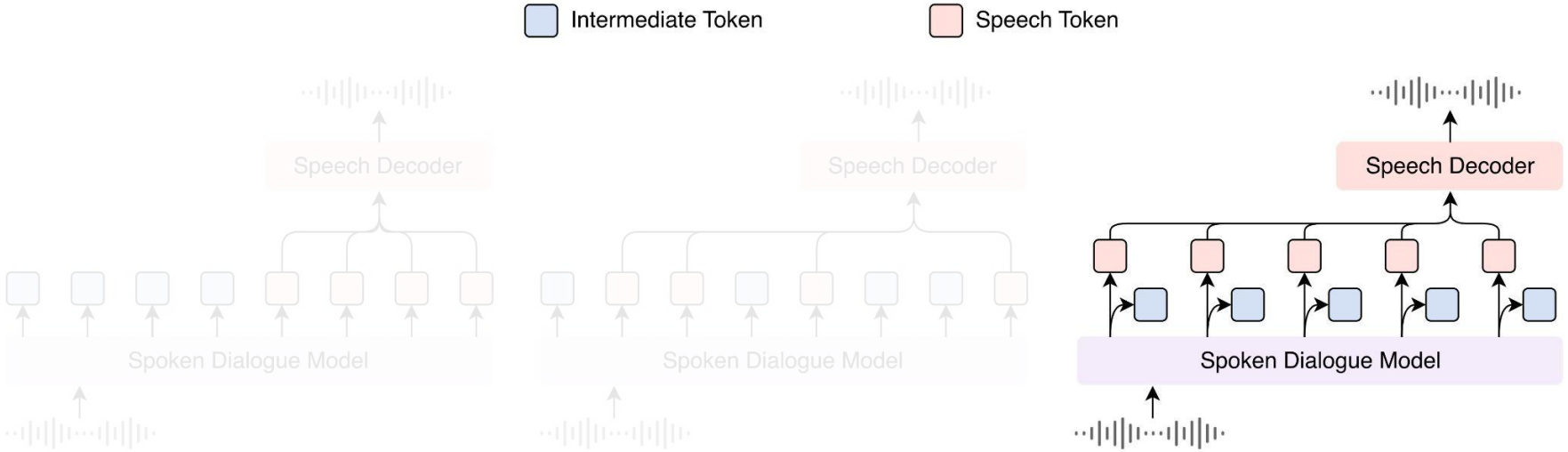
- Step-Audio 2
- LFM2-Audio

StepFun Audio Team [axXiv 2025]
Liquid AI Team [arXiv 2025]

(c) Parallel

- LLaMA-Omni
- Moshi

Fang+ [ICLR 2025]
D´efossez*, Mazar´e* [arXiv 2024]



(a) Sequential

- Qwen-Omni
- Freeze Omni

Qwen Team [arXiv 2025]
Wang+ [PMLR 2025]

(b) Interleaving

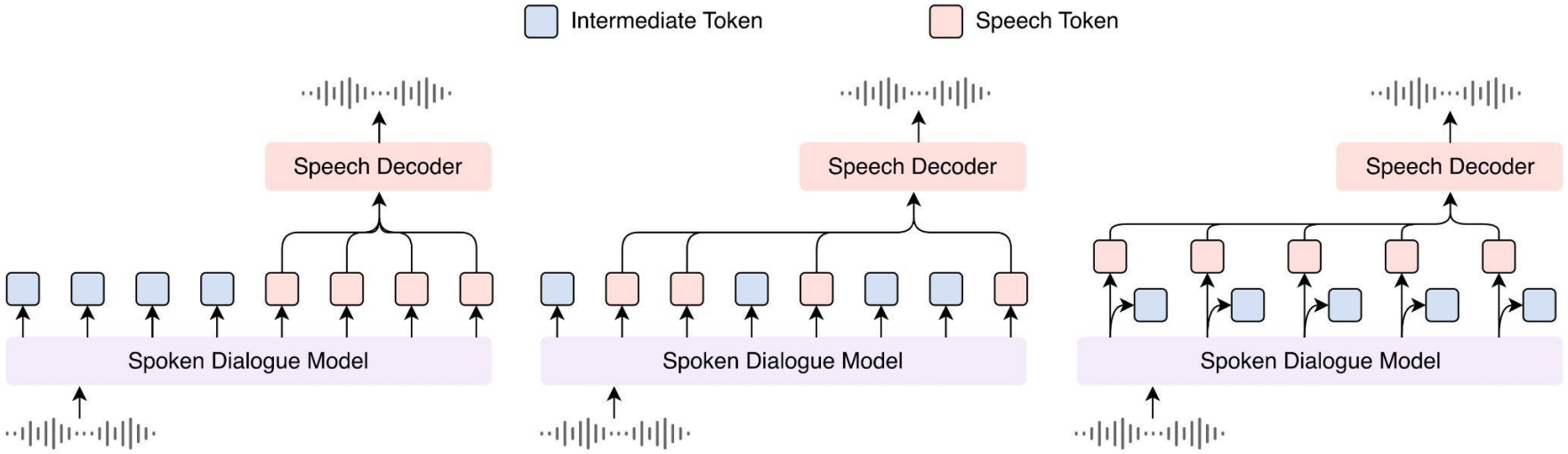
- Step-Audio
- LFM2-Audio

StepFun Audio Team [axXiv 2025]
Liquid AI Team [arXiv 2025]

(c) Parallel

- LLaMA-Omni
- Moshi

Fang+ [ICLR 2025]
D´efossez*, Mazar´e* [arXiv 2024]



(a) Sequential

- Qwen-Omni
- Freeze Omni

(b) Interleaving

- Step-Audio
- LFM2-Audio

(c) Parallel

- LLaMA-Omni
- Moshi

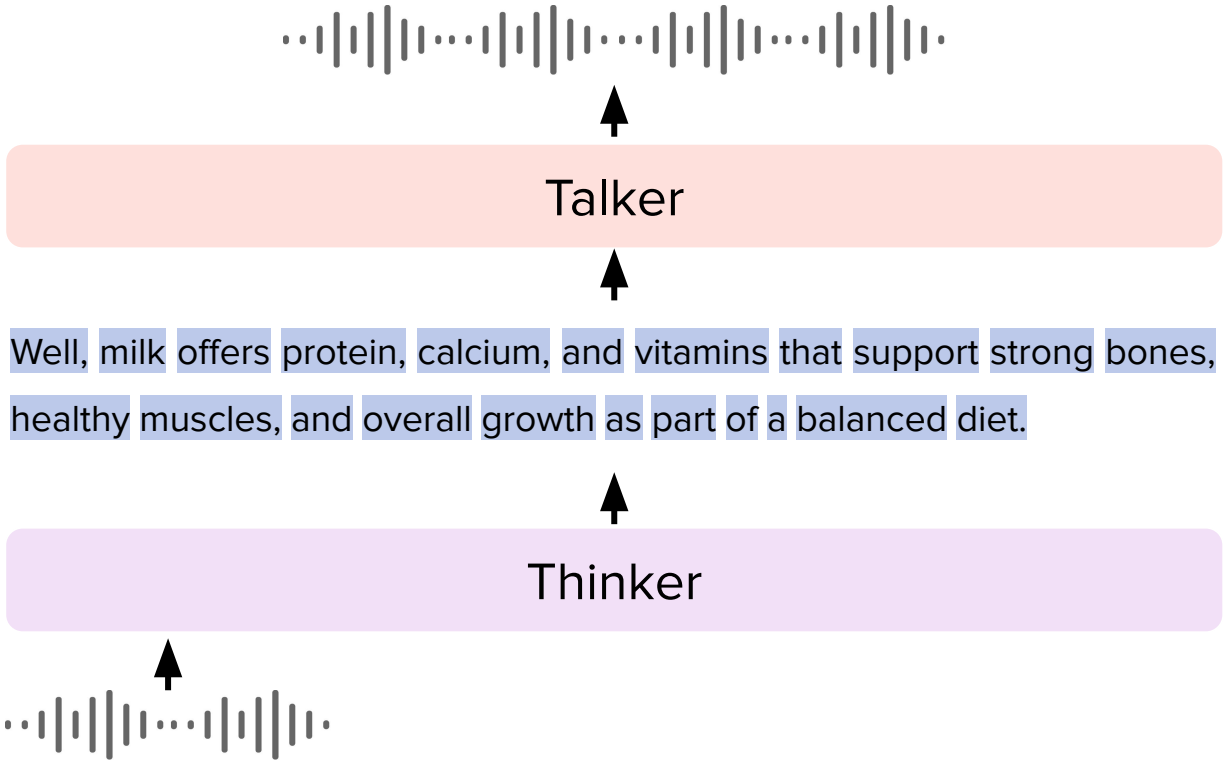
For a more comprehensive survey,
please refer to the TiCo paper.

Chang*, Chen*+ [arXiv 2026]

TiCo’s core idea

Let “Time tokens” participate in semantic planning

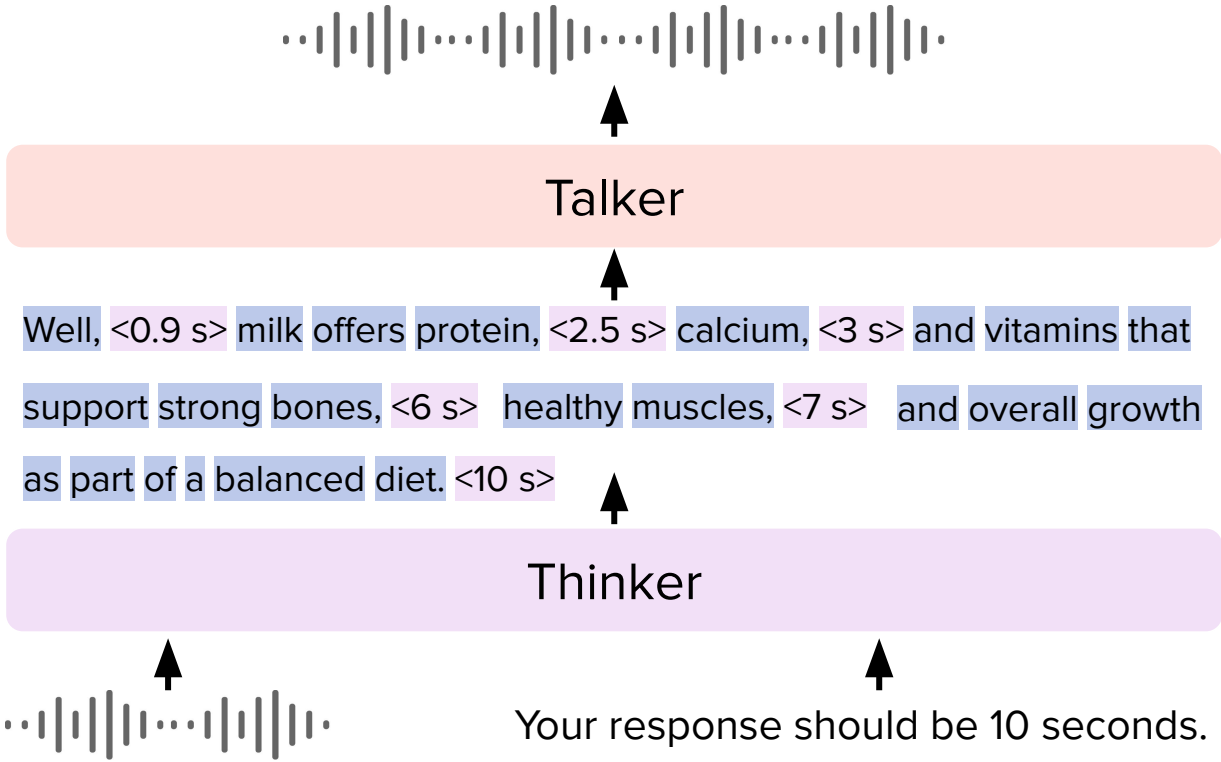
TiCo: Qwen-2.5 Omni as an example



(What benefits does milk offer?)

Qwen Team [arXiv 2025]

TiCo: Qwen-2.5 Omni as an example



(What benefits does milk offer?)

Qwen Team [arXiv 2025]

TiCo Training: Stage 1 - Time Awareness Training

Self-generation

Talker

Well, milk offers protein, calcium, and vitamins that support strong bones, healthy muscles, and overall growth as part of a balanced diet.

Thinker

(What benefits does milk offer?)

Alignment

well: 0.3s - 0.9s

milk: 1.2s - 1.8s

...

diet: 9.4s - 10.0s

Insert

Dataset: InstructS2S subset
6K user queries

Fang+ [ICLR 2025]

TiCo Training: Stage 1 - Time Awareness Training

Self-generation

Talker

Well, <0.9 s> milk offers protein, <2.5 s> calcium, <3 s> and vitamins that support strong bones, <6 s> healthy muscles, <7 s> and overall growth as part of a balanced diet. <10 s>

Thinker

(What benefits does milk offer?)

Alignment

well: 0.3s - 0.9s

milk: 1.2s - 1.8s

...

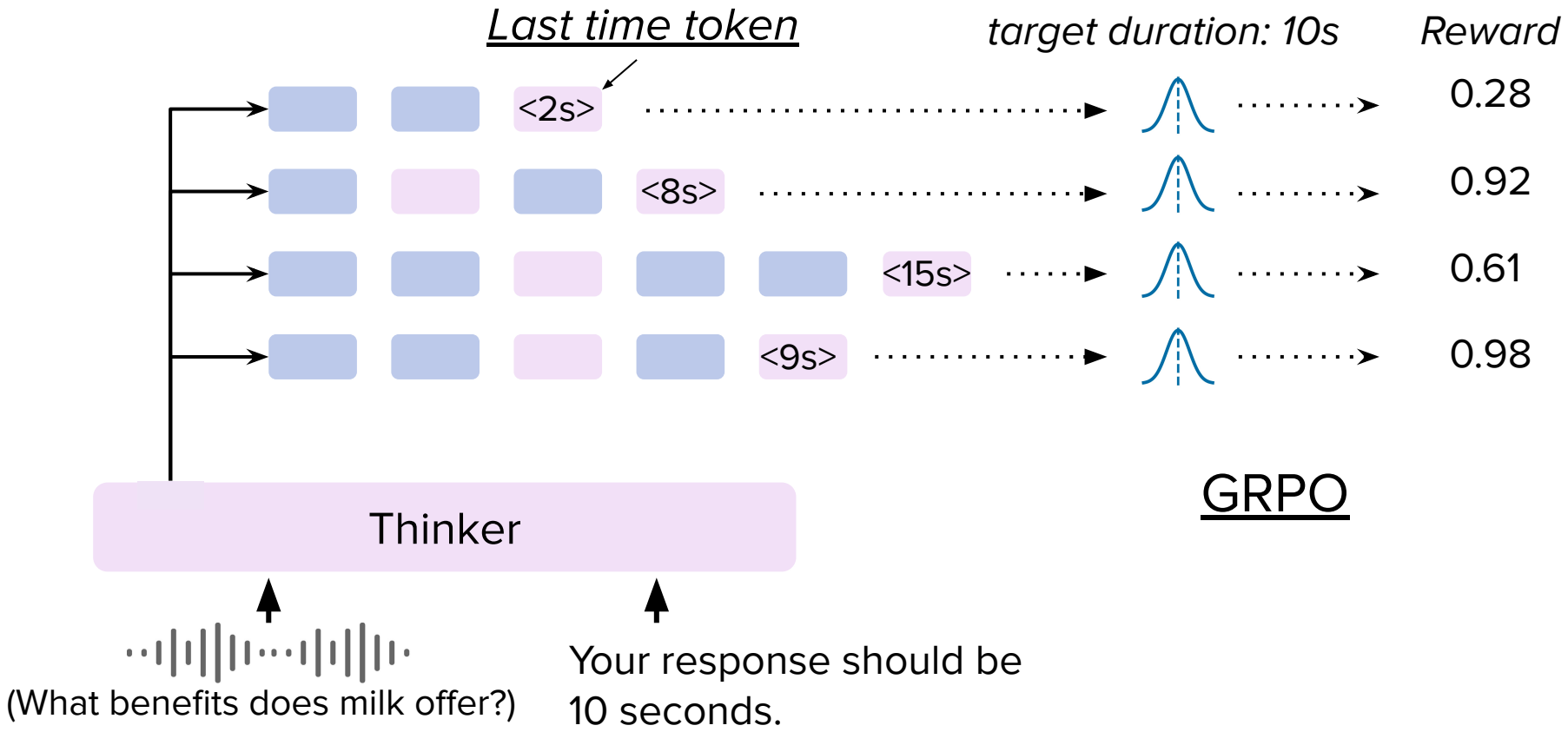
diet: 9.4s - 10.0s

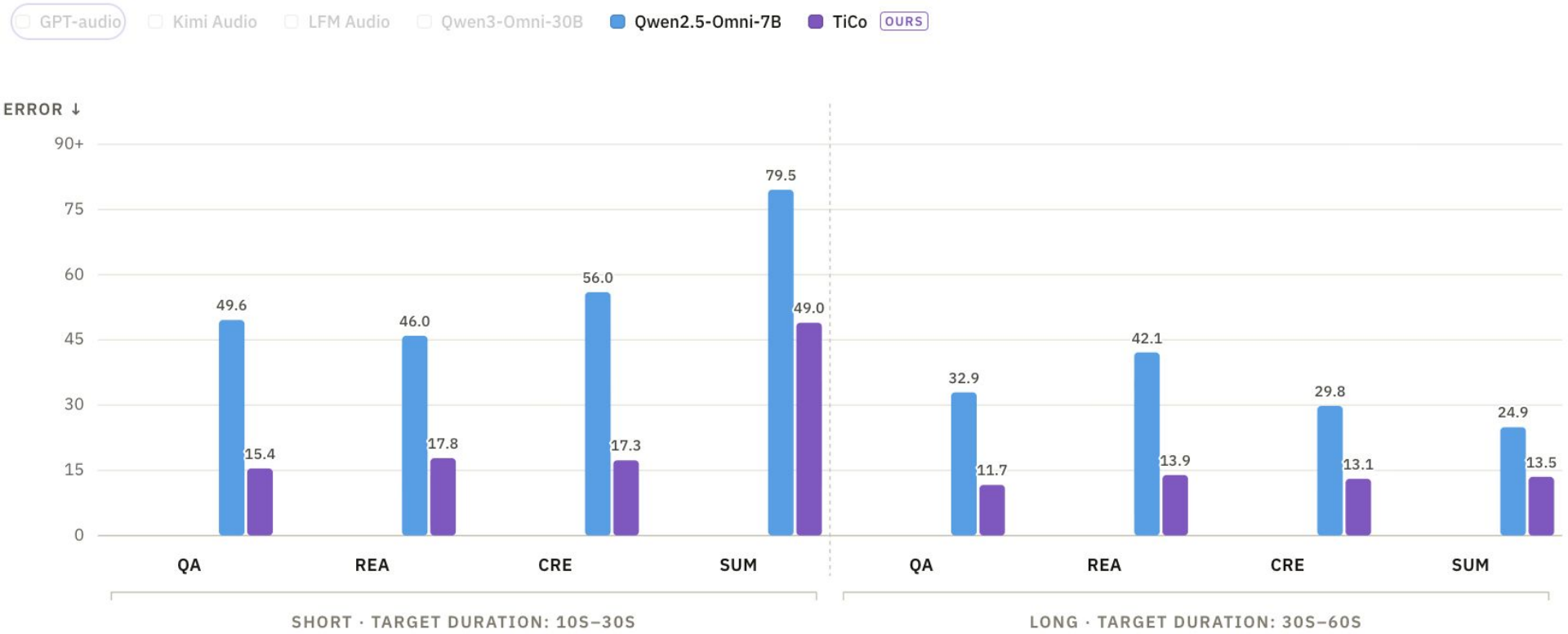
New prediction target for the thinker

Dataset: InstructS2S subset
6K user queries

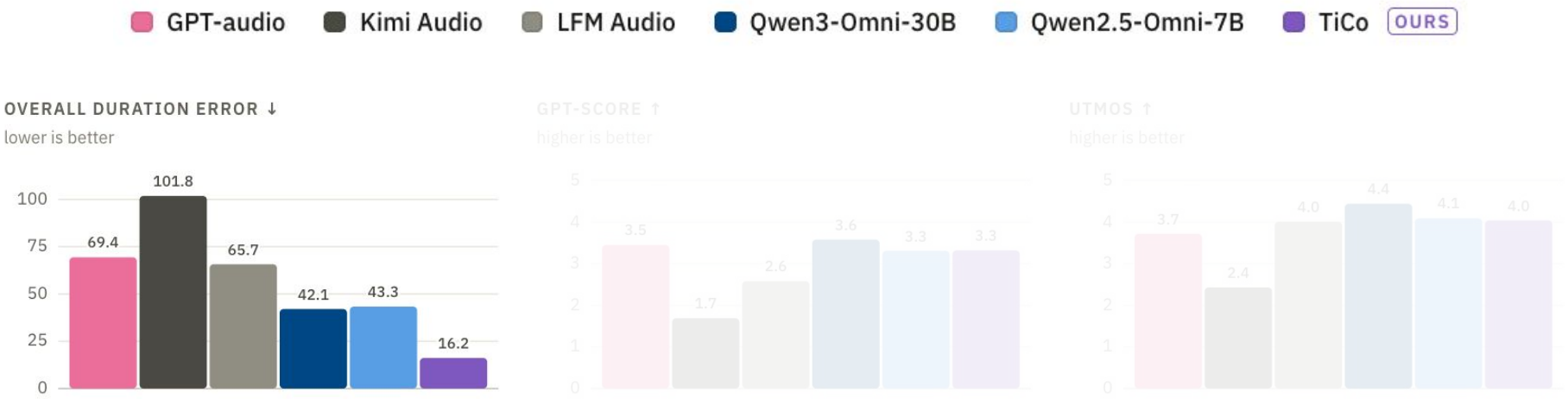
Fang+ [ICLR 2025]

TiCo Training: Stage 2 - Time Controllable Training

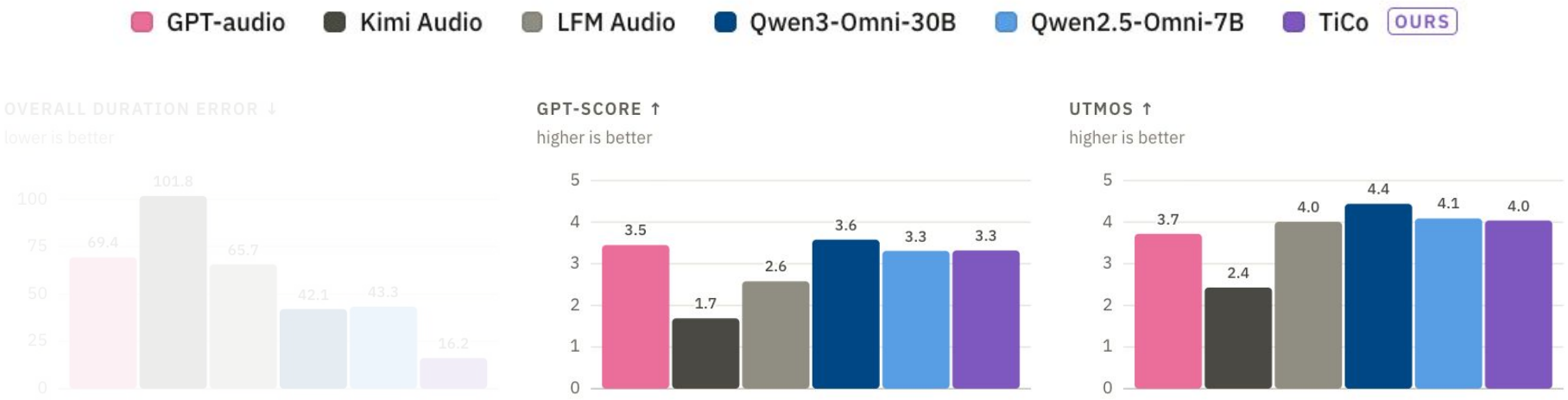




- Duration error drops a lot
- The only exception is the summarization task



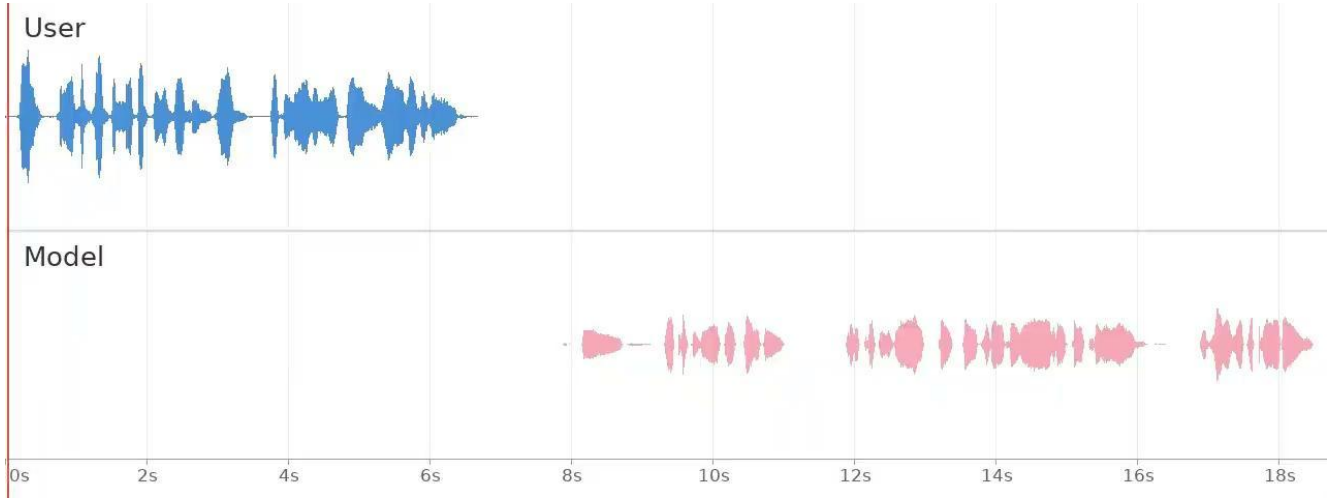
- TiCo achieves lowest duration error



- TiCo achieves lowest duration error
- TiCo does not degrade on response quality, both in semantic (GPT-SCORE) and acoustic (UTMOS)

Target duration:
10 seconds
(text instruction)

Model:
TiCo



- The ocean is very deep in some places
- The Mariana Trench is the deepest part
- Average ocean depth: ~3.7 km
  - Examples:
    - Shallow bays
    - The middle of the ocean
  - Clarification
    - 3.7 km is an average not a fixed depth

How deep is the ocean?, target duration = 15 s

Prompt: *How deep is the ocean? Please respond with exactly 15 seconds.*

Speech response:

Well, <0.9 seconds> you know, <1.6 seconds> the ocean is really deep in some places. <3.8 seconds> The Mariana Trench is the deepest part, <7.2 seconds> it's about 36 kilometers deep. <9.4 seconds> But overall, <11.0 seconds> the average depth of the world's oceans is around 3.7 kilometers. <15.0 seconds>

Target duration: *15.0 seconds*
Actual duration: *15.2 seconds*

How deep is the ocean?, target duration = 40 s

Prompt: *How deep is the ocean? Please respond with exactly 40 seconds.*

Speech response:

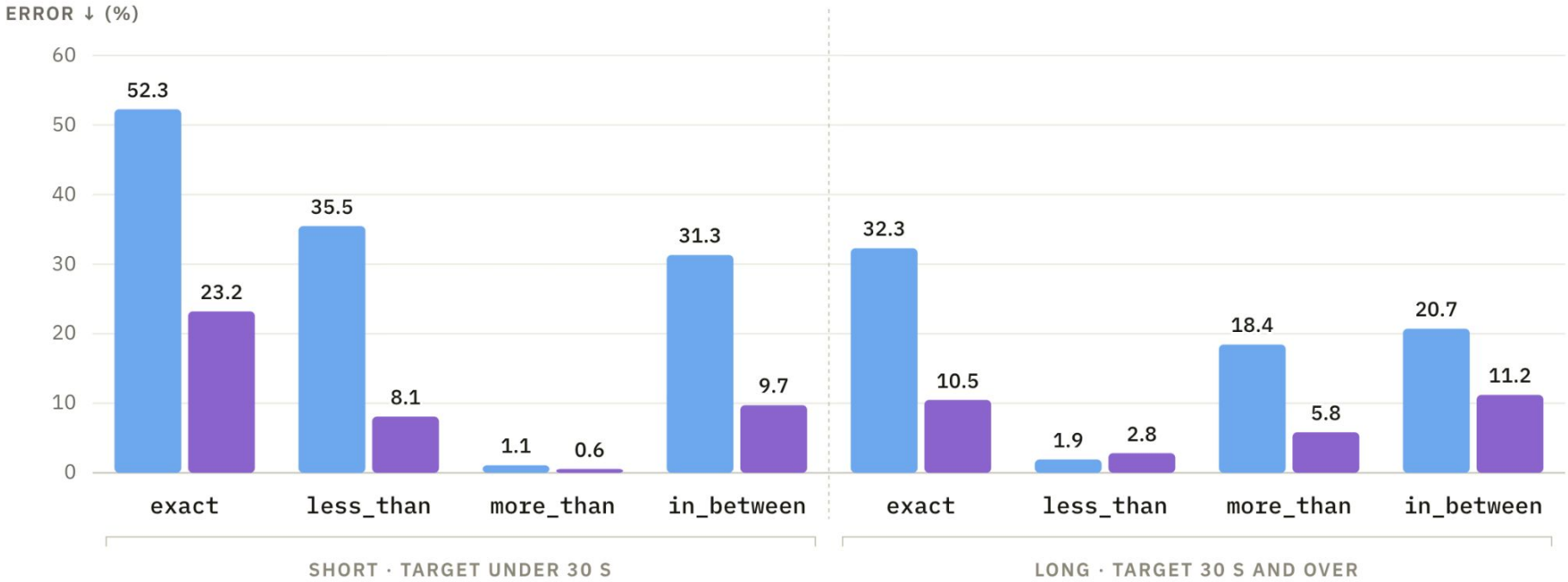
Well, <1.5 seconds> you know, <2.3 seconds> the ocean is really deep in some places. <4.6 seconds> The Mariana Trench is one of the deepest parts. <7.9 seconds> It's about 36 kilometers deep. <10.8 seconds> But overall, <12.5 seconds> if we're talking about an average depth for the whole ocean, <15.5 seconds> it's around 3.7 kilometers. <18.0 seconds> However, <19.4 seconds> different areas can have very different depths. <22.3 seconds> Some shallow bays might be just a few meters deep while others in the middle of the ocean could be much deeper than that. <29.0 seconds> So, <30.0 seconds> it's not really a fixed number like 3.7 kilometers everywhere.

<33.0 seconds> If you want to know more about why the ocean has such varying depths or anything else related, <39.0 seconds> feel free to ask. <40.0 seconds>

Target duration: *40.0 seconds*
Actual duration: *41.6 seconds*

A longer time constraint leads to **more detailed content**

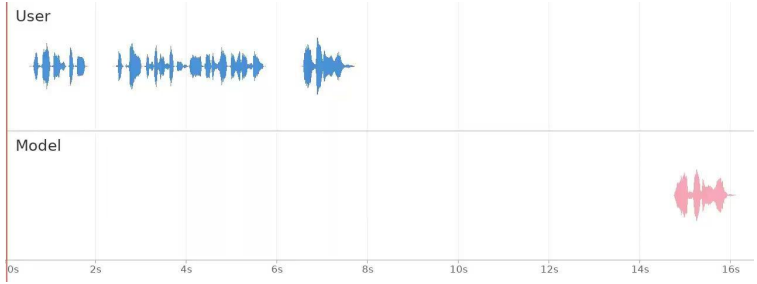
TiCo – Beyond exact duration match



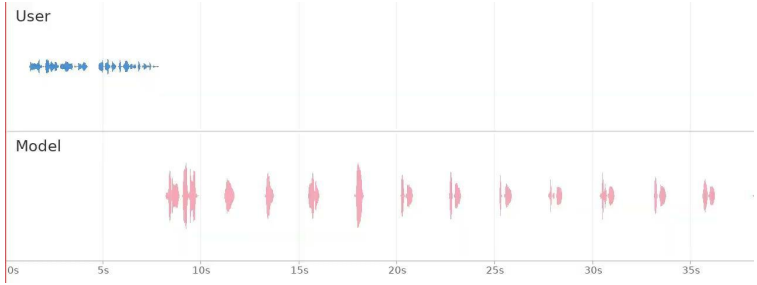
TiCo can generalize to other types of duration instruction

TiCo – Beyond Half-duplex Models

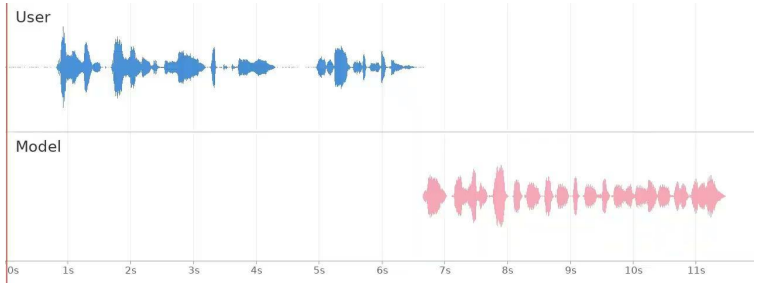
Time-Silence



Time-Slow



Time-Fast



TiCo with Full-duplex Model

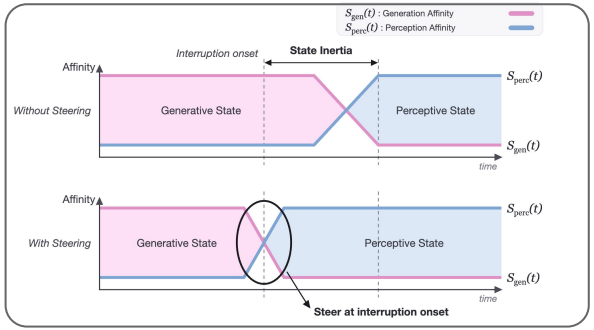
Model: Raon-SpeechChat
Data: Game-Time Benchmark

KRAFTON [arXiv 2026]
Chang*, Hu* [ICASSP 2026]

TiCo Takeaway

- Current spoken dialogue models still lack precise duration controllability.
- We can post-train speech models to explicitly condition on “time-tokens”.
- TiCo generalizes well across various instructions and can potentially be applied to diverse model architectures.

How do full duplex models coordinate speaking and listening?



Game-Time

Chang*, Hu*+ [ICASSP 2025]

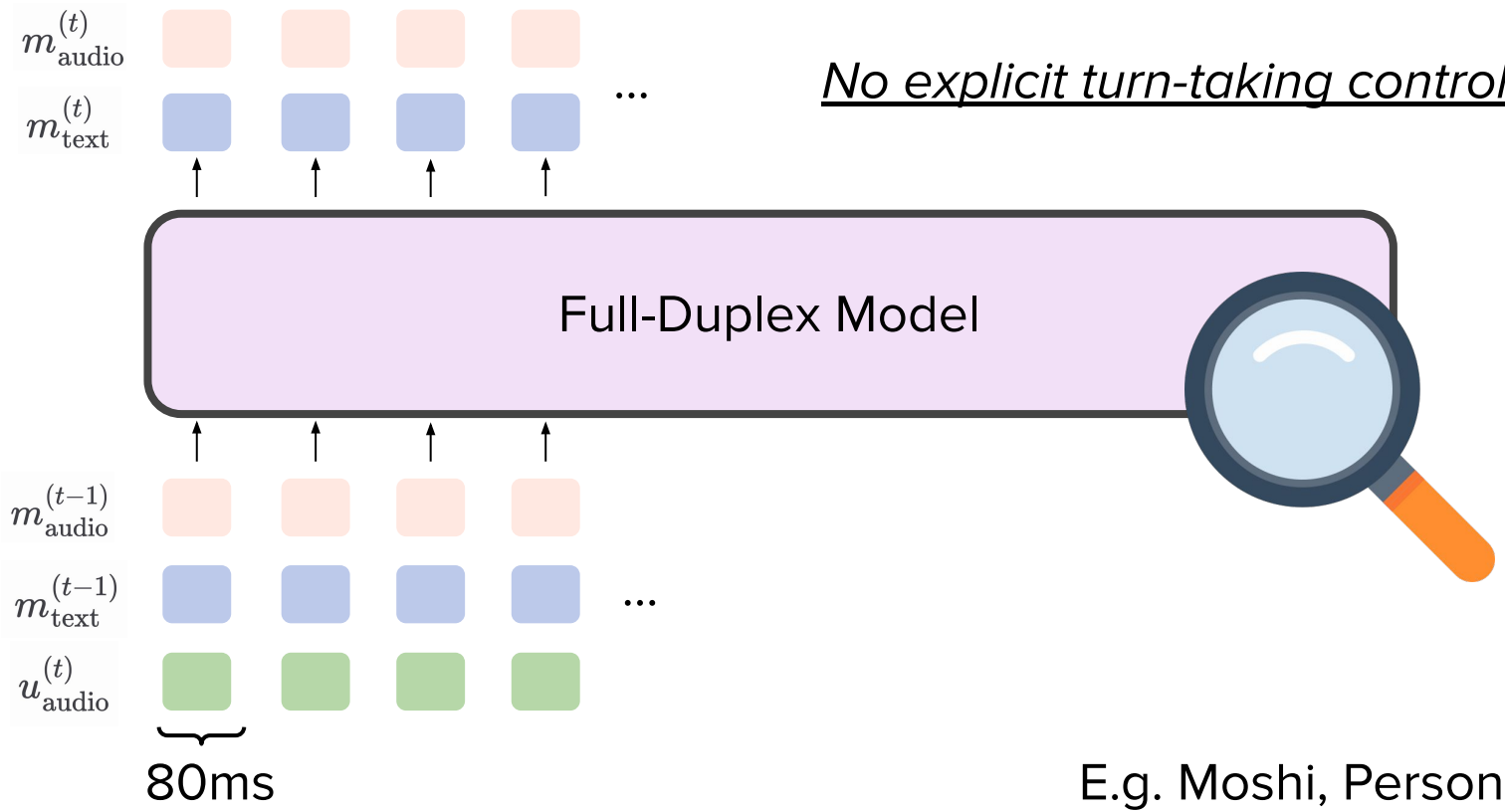
TiCo

Chang*, Chen*+ [arXiv 2026]

State Inertia

Chang, **Chang***+ [arXiv 2026]

Full-Duplex Models



No explicit turn-taking control

E.g. Moshi, PersonaPlex

D´efossez*, Mazar´e* [arXiv 2024]
Roy+ [ICASSP 2026]

FINDING 1

Full-duplex model shows **stream-specific focus**

FINDING 2

Speaking and listening behaviors are coordinated by **generative** and **perceptive states**

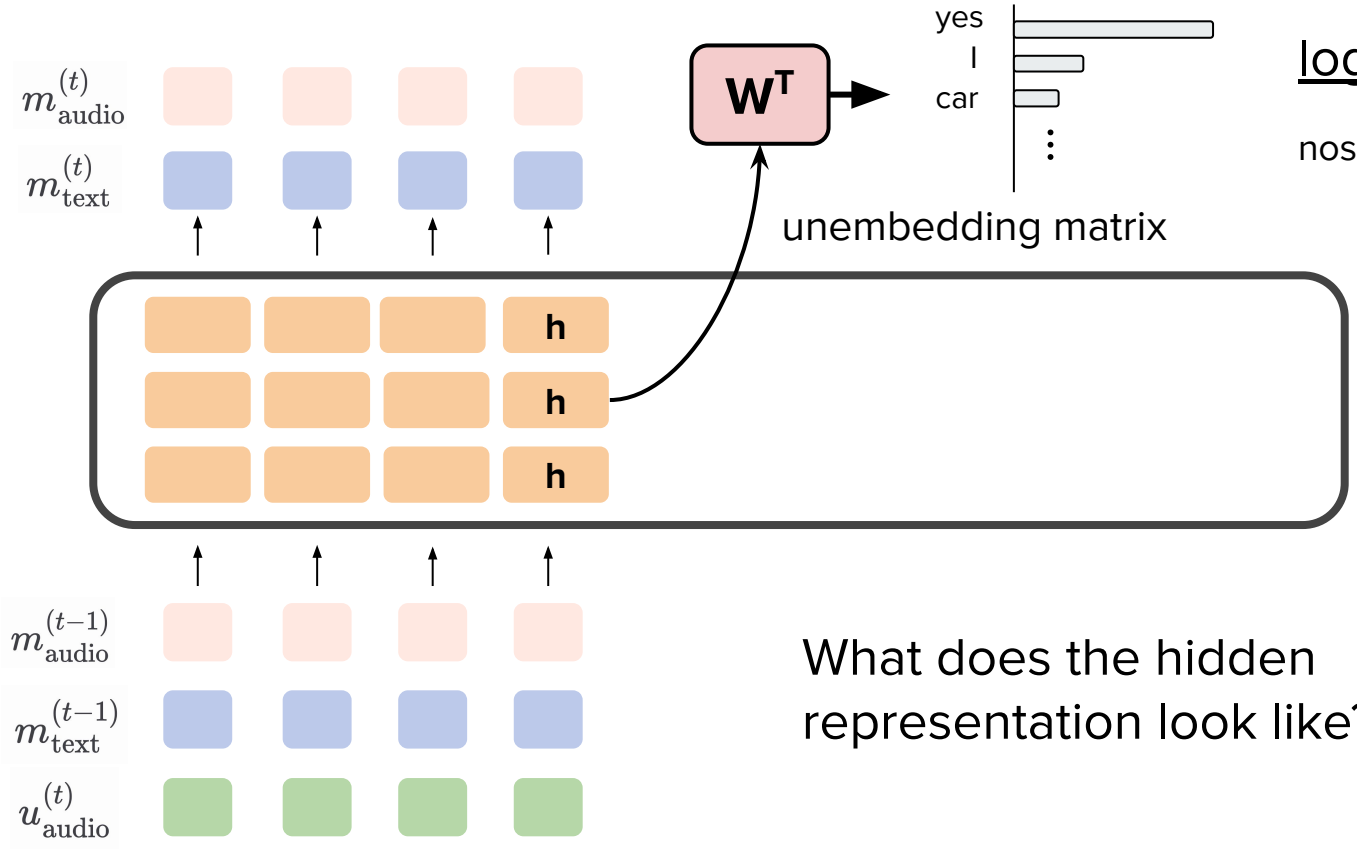
FINDING 3

Interruptions induce **state inertia**

METHOD

Activation steering mitigates state inertia

Finding 1: Stream-specific predictive focus

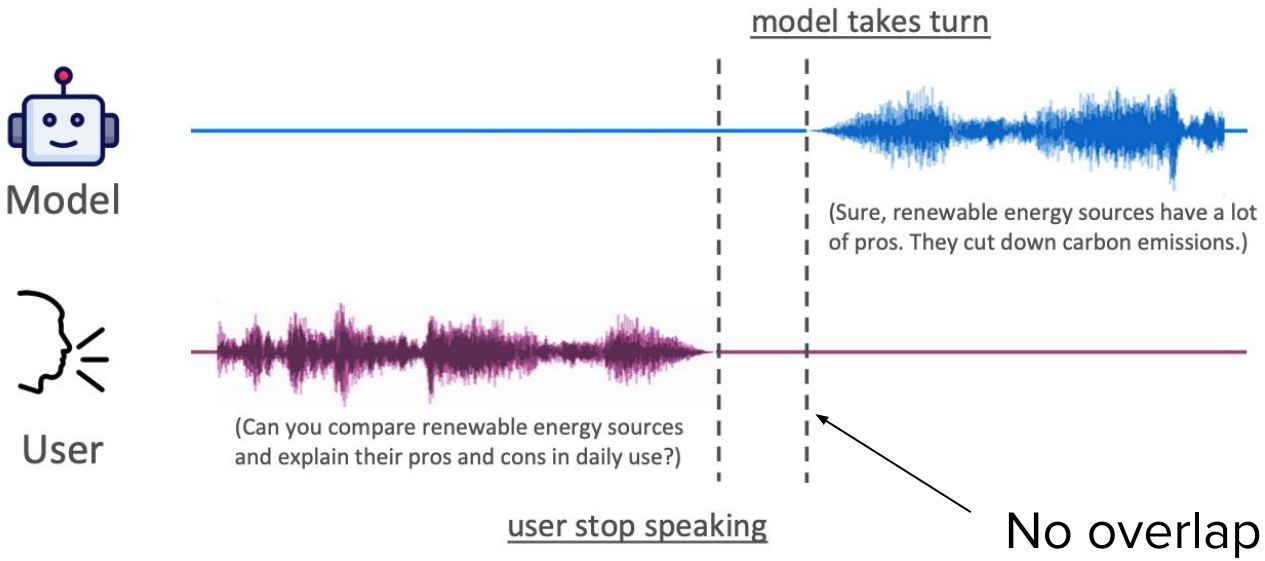


logit lens

nostalgebraist [2020]

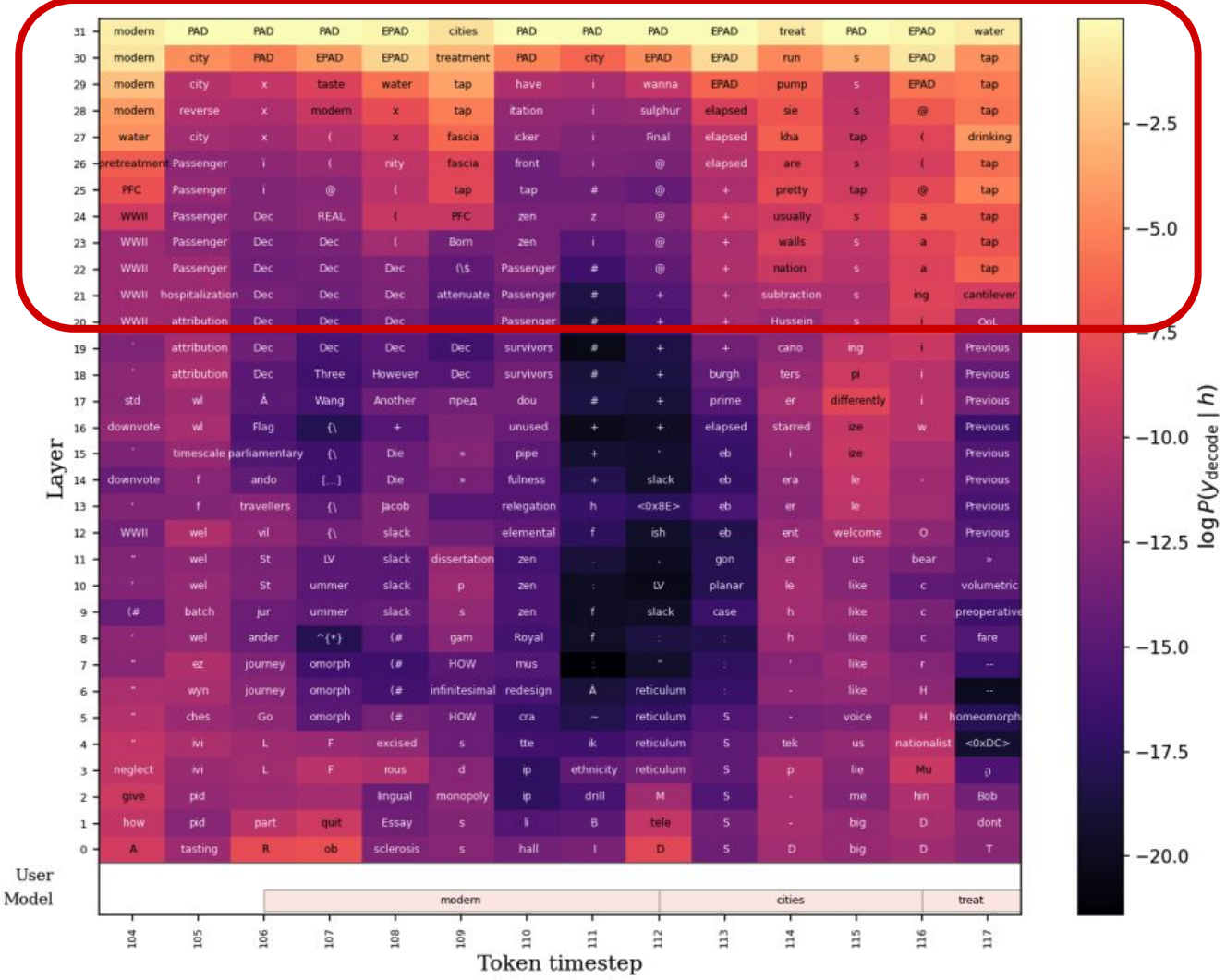
What does the hidden representation look like?

Finding 1: Stream-specific predictive focus



Turn-by-turn dataset: The model first listens to the user (stay silent) then respond

Logit Lens during model speaking



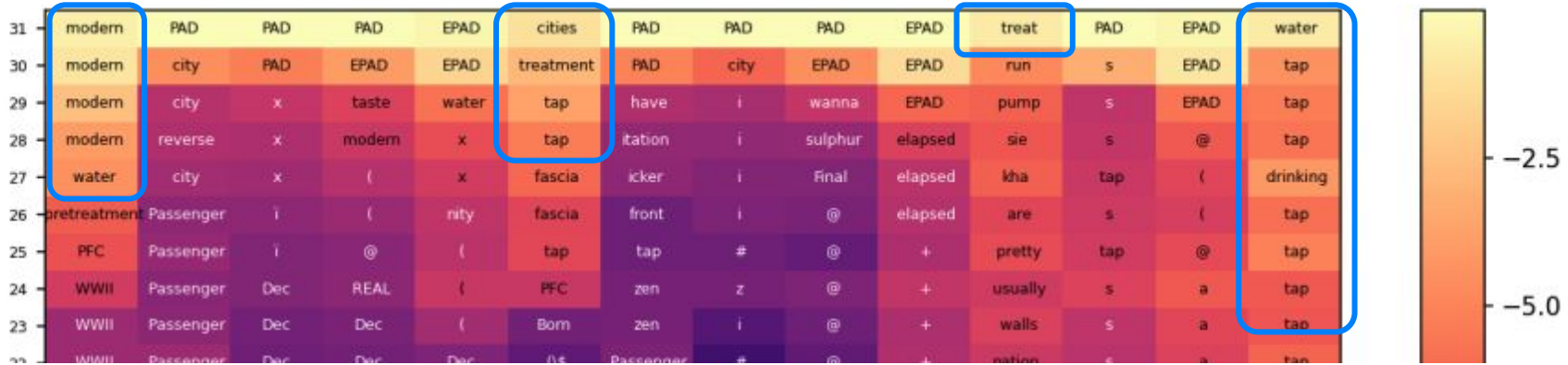
Model: PersonaPlex

Roy+ [ICASSP 2026]

Logit Lens during model speaking

User: How does water treatment make tap water safe to drink in modern cities?

Model: Modern cities treat water ....



model speaks: (alignment)

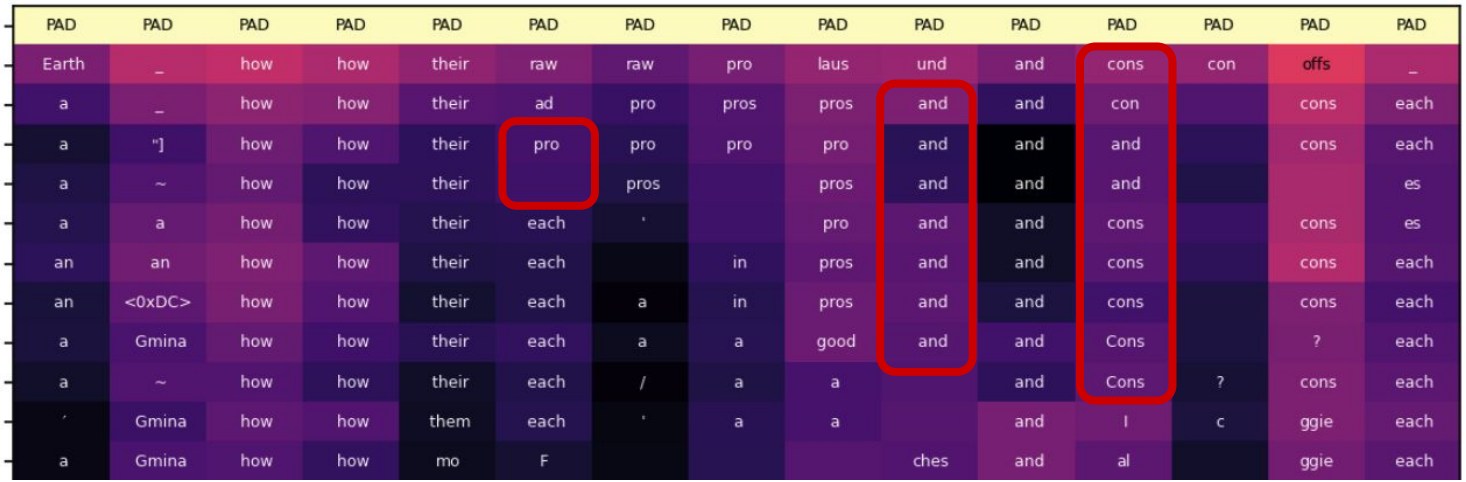
	modern	cities	treat

Speaking: Hidden representation predicting model's output stream

Logit Lens during model listening

User: Can you compare renewable energy sources and explain their pros and cons in daily use?

Model: Sure, renewable energy sources ...



user speak:

explain	their	pros	and	cons

Listening: Predicting the incoming user stream

FINDING 1

Full-duplex model shows **stream-specific focus:**

Listening: predict the incoming user stream

Speaking: predict the output model stream

Finding 2: Generative State and Perceptive State

Perception Affinity $\mathcal{S}_{\text{perc}}(t) = P(u_{\text{audio}}^{(t+1)} \mid h^{(t)})$

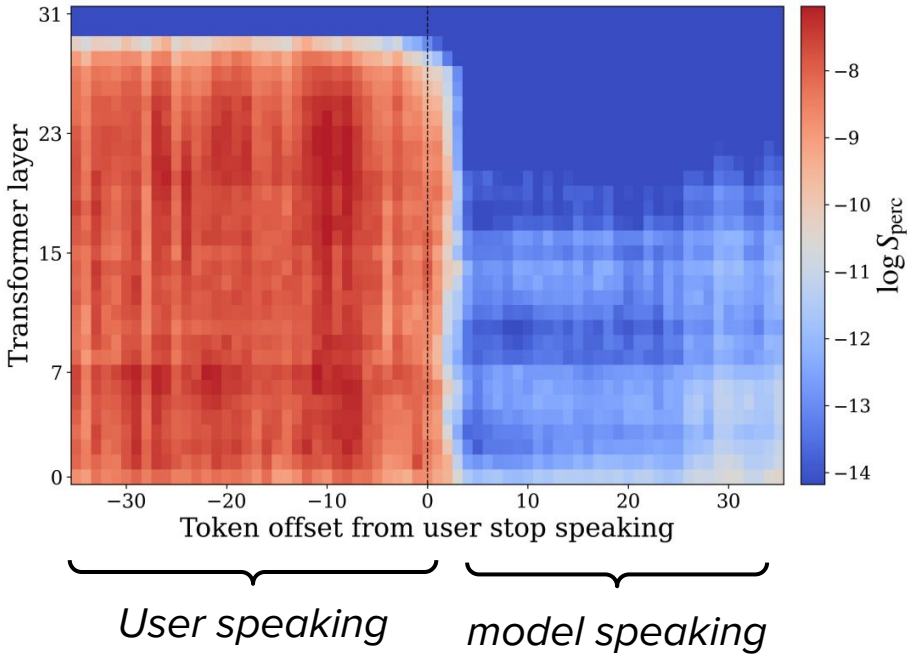
Generation Affinity $\mathcal{S}_{\text{gen}}(t) = \frac{1}{2}\left(P(m_{\text{audio}}^{(t)} \mid h^{(t)}) + P(m_{\text{text}}^{(t)} \mid h^{(t)})\right)$

How closely the hidden representation aligns with the model stream and the user stream

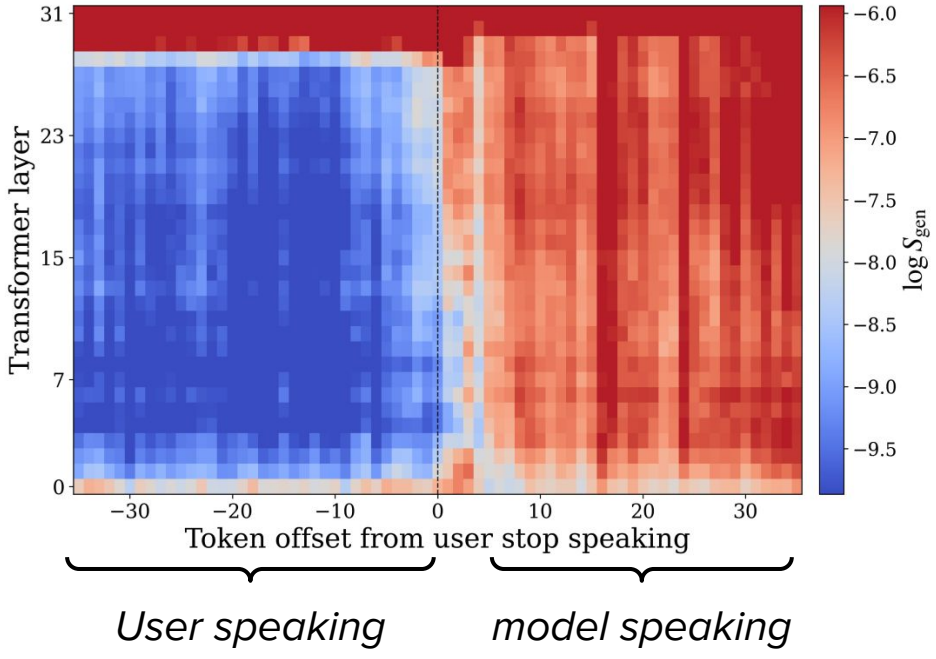
Finding 2: Generative State and Perceptive State

Turn-by-turn dataset. 100 samples.
Align user stop speaking at timestamp = 0

Perception Affinity



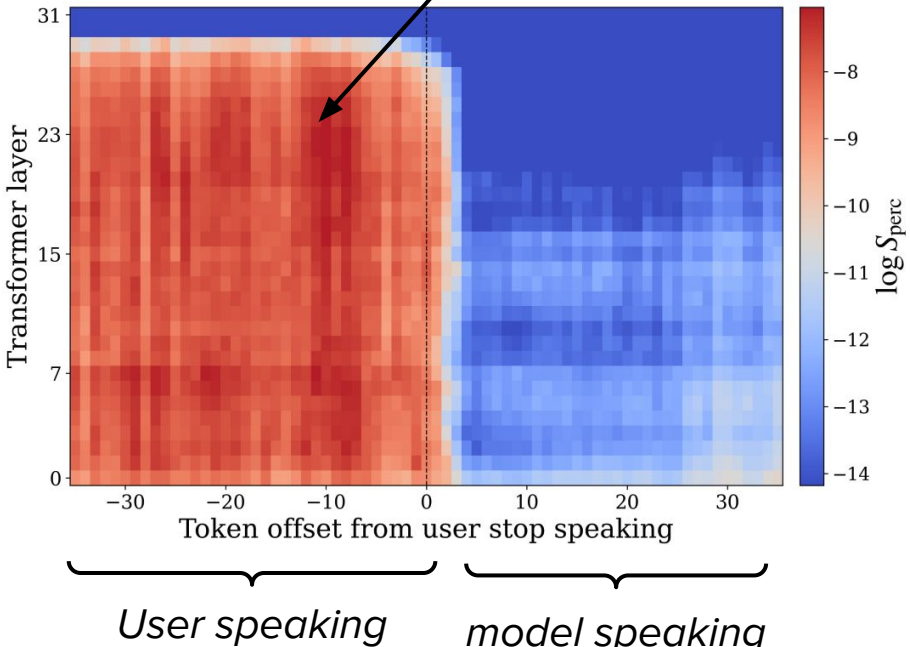
Generation Affinity



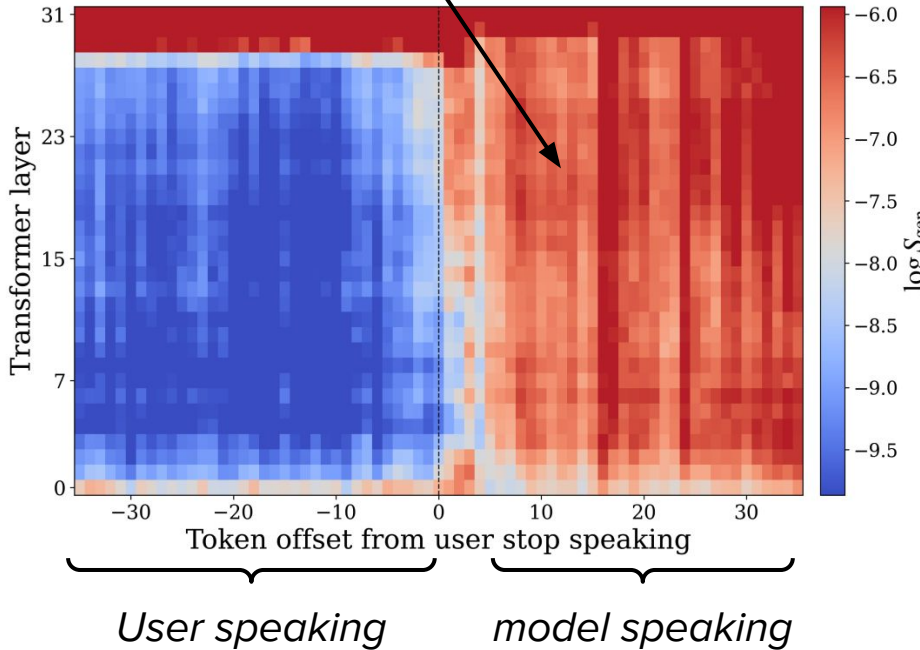
Finding 2: Generative State and Perceptive State

Perceptive State Generative state

Perception Affinity



Generation Affinity

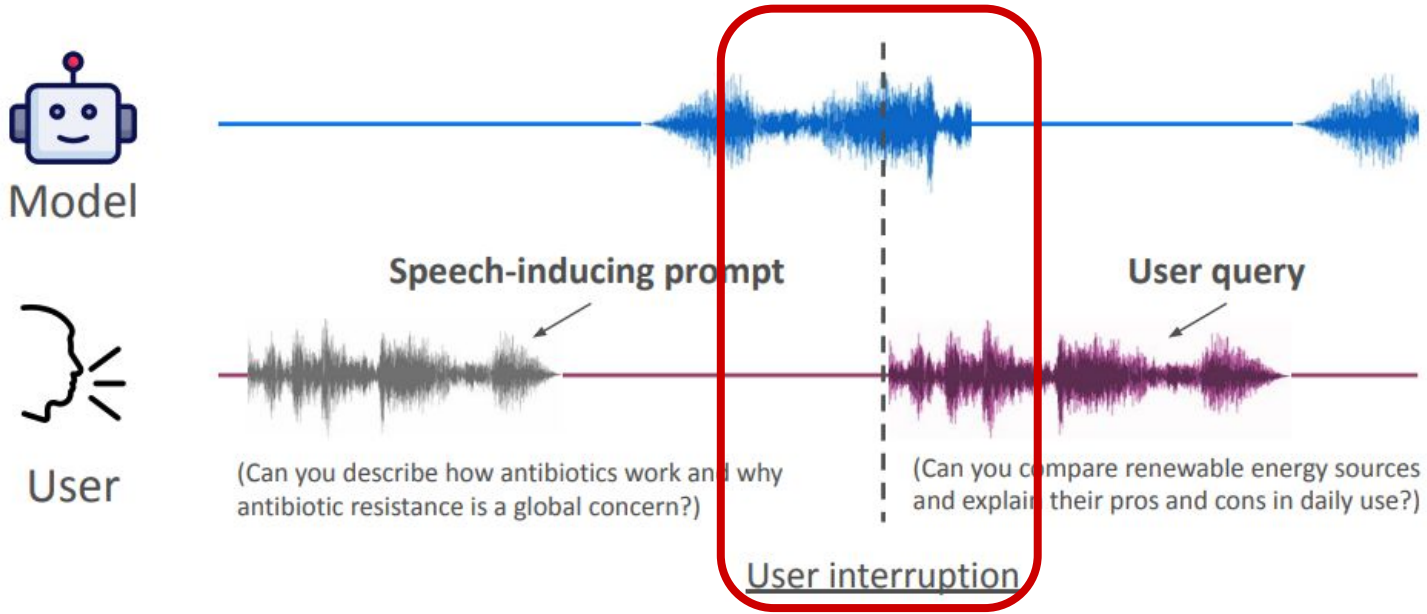


FINDING 2

Full-duplex models coordinate speaking and listening by modulating between generative and perceptive states

Finding 3: State Inertia

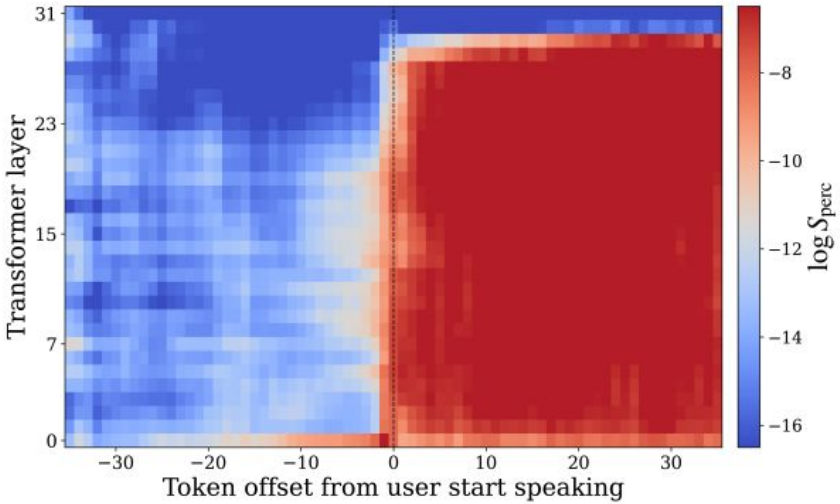
What if the model and the user overlap?



Finding 3: State Inertia

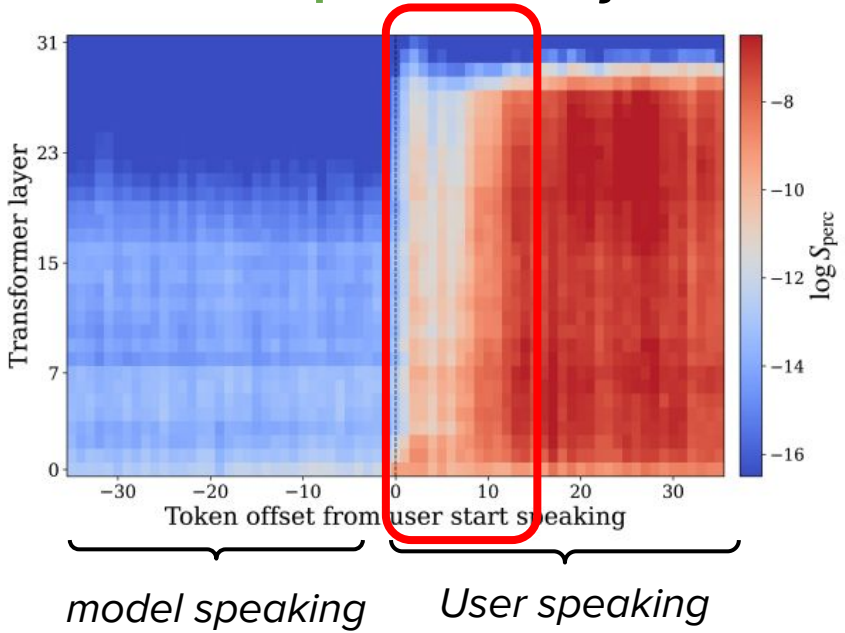
No interruption

Perception Affinity



Interruption

Perception Affinity



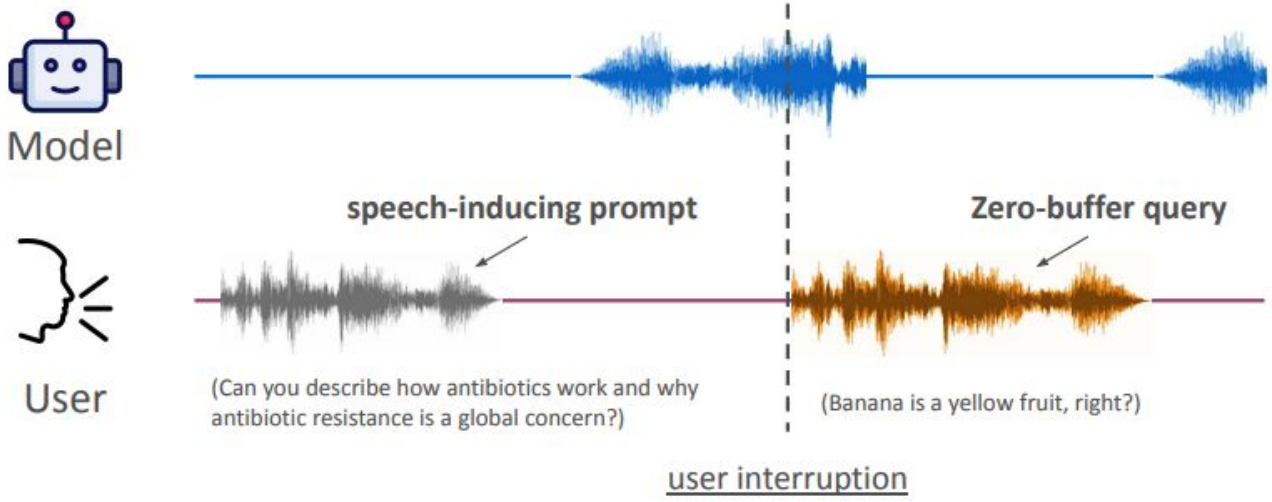
State Inertia. The model can't switch states fast enough.

FINDING 3

Full-duplex Models shows **state inertia**, a tendency to remain in its prior state.

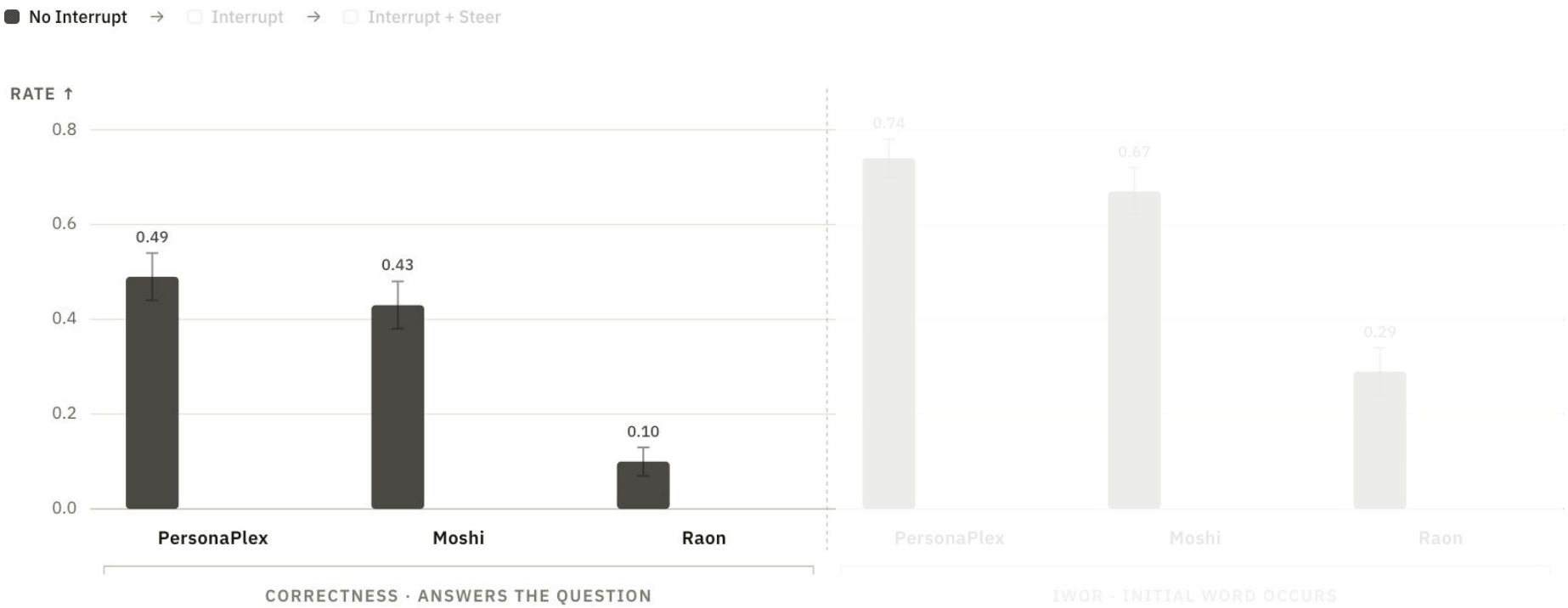
(When the model is speaking, it can not listen to the users)

Zero-Buffer Benchmark (ZBB)



Zero-buffer query:
e.g. **Banana is a yellow fruit, right?**

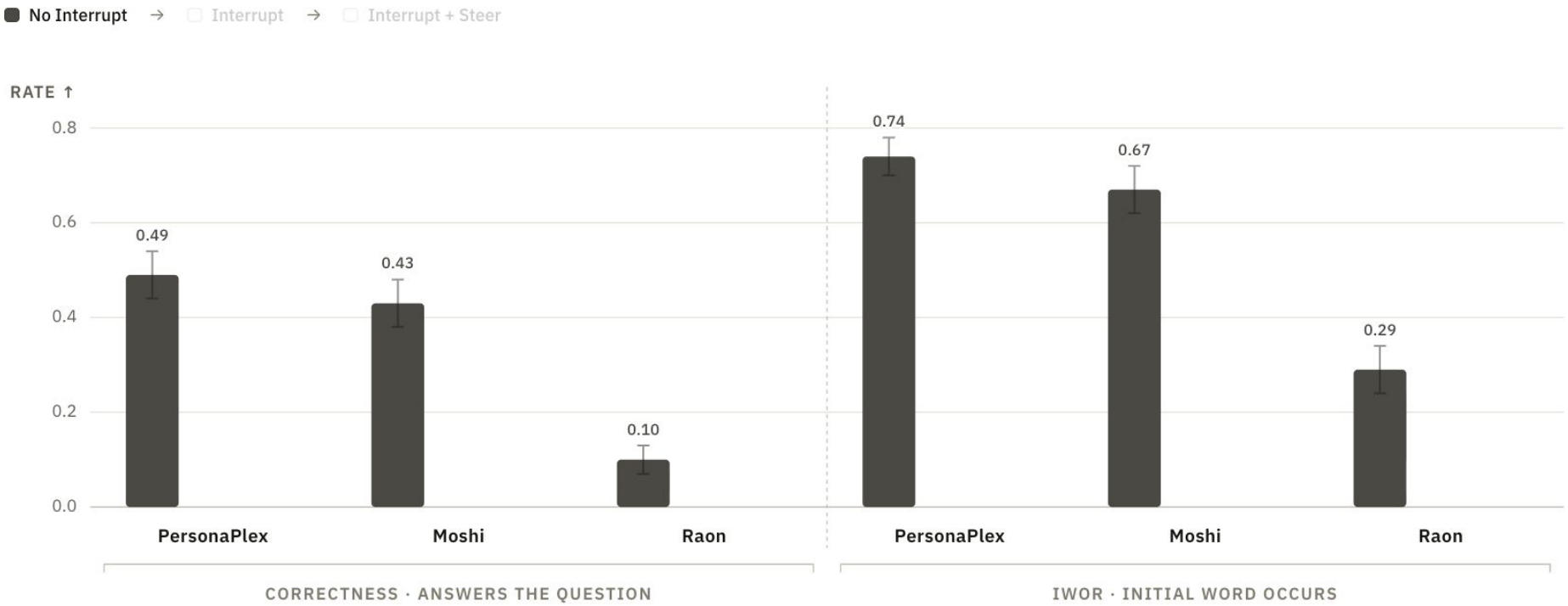
The model can only answer the question correctly if it does not miss the initial keyword



- **Correctness** ↑
  The model answers the question correctly

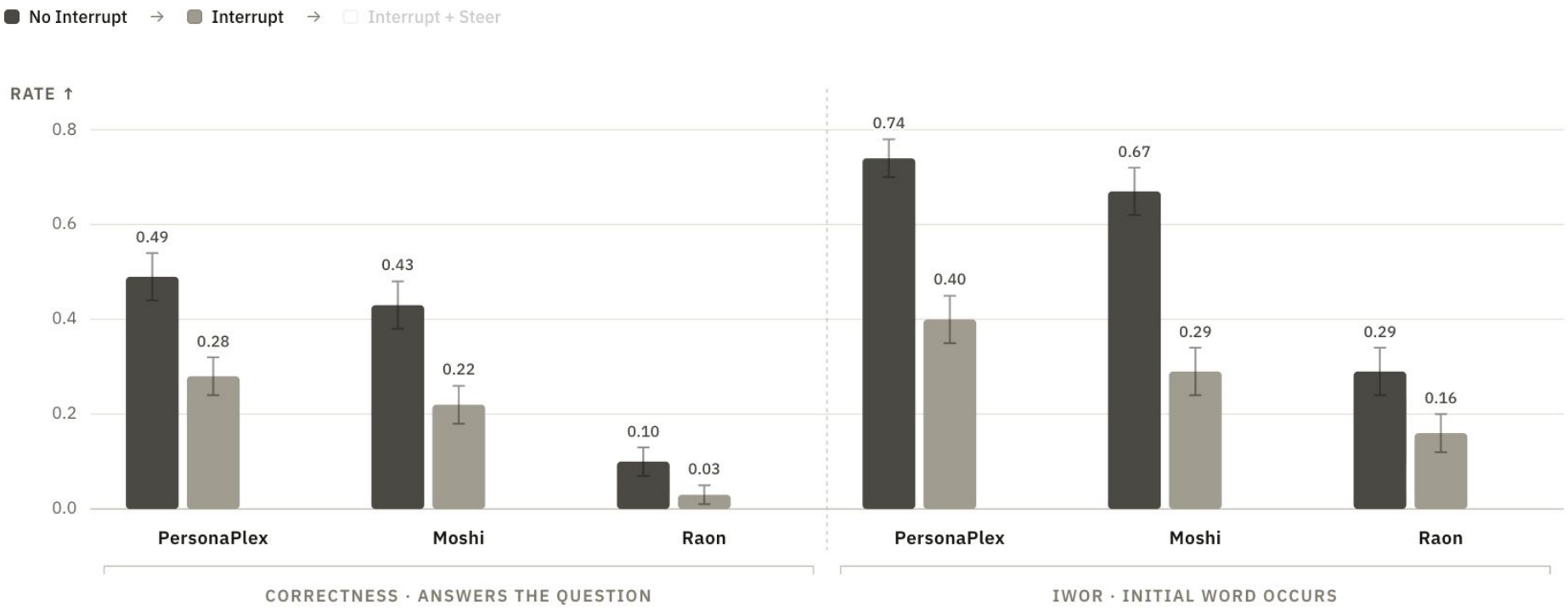
- **IWOR (Initial Word Occurance Rate)** ↑
  The initial word occurs in the response

Interruption breaks the model's answer



- **Correctness** ↑
  The model answers the question correctly
- **IWOR (Initial Word Occurance Rate)** ↑
  The initial word occurs in the response

Interruption breaks the model's answer



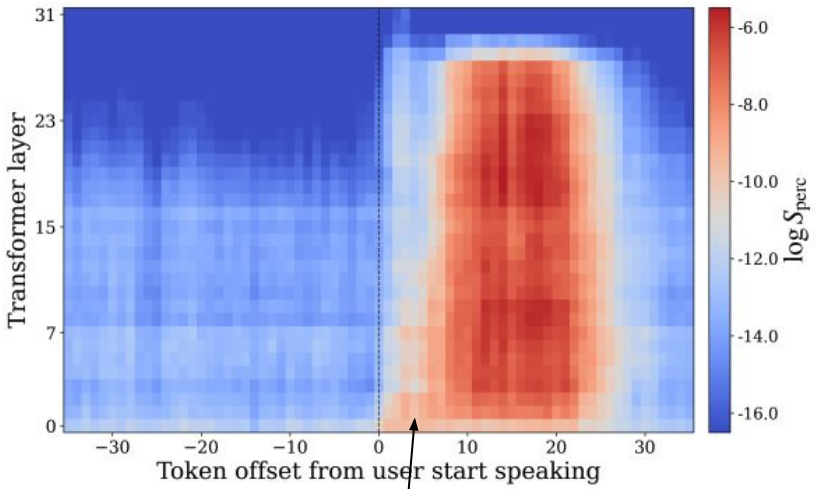
- **Correctness** ↑
  The model answers the question correctly
- **IWOR (Initial Word Occurance Rate)**
  The initial word occurs in the response

Interruption breaks the model's answer

Activation Steering

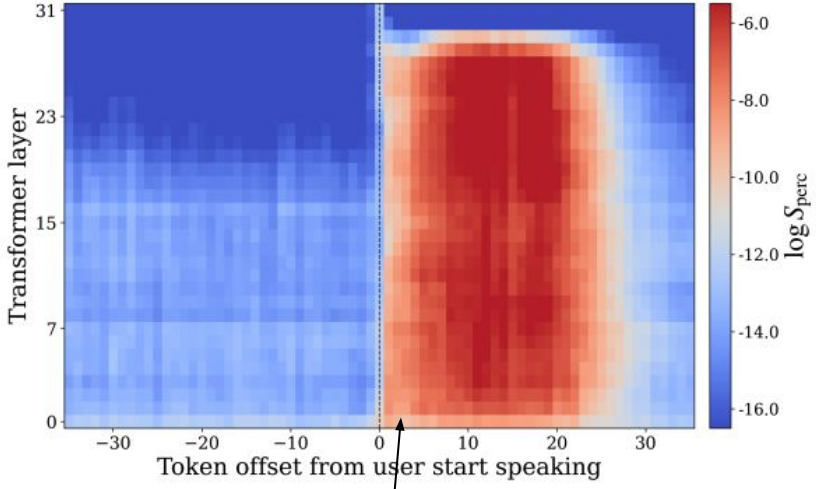
Perception vector $$\mu_{g \to p} = \frac{1}{|T_{\text{perc}}|} \sum_{t \in T_{\text{perc}}} h^{(t)} - \frac{1}{|T_{\text{gen}}|} \sum_{t \in T_{\text{gen}}} h^{(t)}$$

Perception Affinity w/o steering

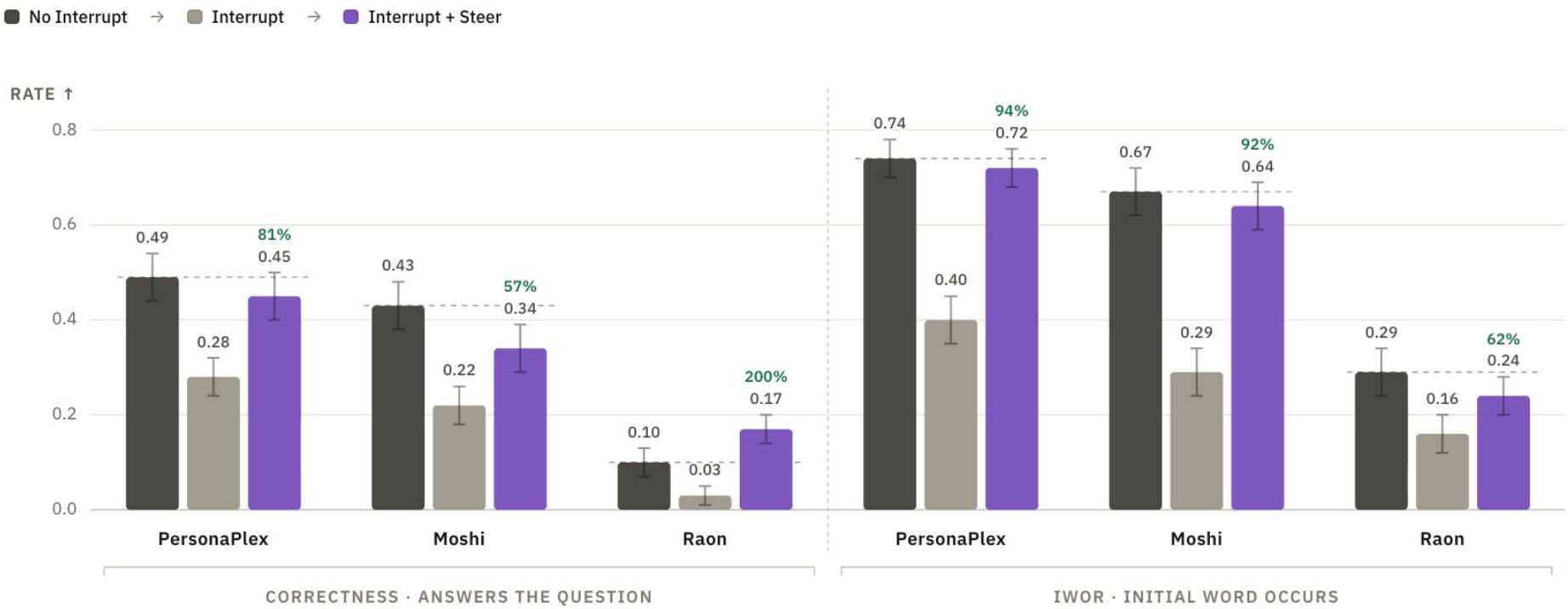


state inertia

Perception Affinity w/ steering



steering



- **Correctness** ↑
  The model answers the question correctly
- **IWOR (Initial Word Occurance Rate)**
  The initial word occurs in the response

Steering brings back the performance

State Inertia Takeaway

- Full-duplex models operate in two internal states: a **generative state** and a **perceptive state**

- **State inertia** makes the model slow to switch states, missing the beginning of user interruptions

- **Activation steering** can mitigate state inertia — no training needed

Conclusion

- Current spoken dialogue models still lack “time-awareness”.
- TiCo: Post-train the model with time-conditioning for better controllability.
- State-Inertia: Study internal mechanisms to better understand and improve the model.
- There is still a lot of room for research on “time” in spoken conversations.