Game-Time:

Evaluating Temporal Dynamics in Spoken Language Models

Kai-Wei Chang, *Postdoc Researcher, MIT CSAIL*

MIT CSAIL Computer Science & Artificial Intelligence Laboratory

ICASSP2026 Barcelona

Game-Time:

Evaluating Temporal Dynamics in Spoken Language Models

Kai-Wei Chang, *Postdoc Researcher, MIT CSAIL*

MIT CSAIL Computer Science & Artificial Intelligence Laboratory

Whether current SLMs have "time awareness"?

ICASSP2026 Barcelona

Outline

1. Motivation
2. Game-Time Benchmarks
    a. Benchmark construction
    b. Evaluation Methods
    c. Benchmark Results
3. Conclusion
4. Future Works
    a. Time Controllable Spoken Dialogue Model - TiCo

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Outline

1. Motivation
2. Game-Time Benchmarks
   a. Benchmark construction
   b. Evaluation Methods
   c. Benchmark Results
3. Conclusion
4. Future Works
   a. Time Controllable Spoken Dialogue Model - TiCo

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Outline

1. Motivation
2. Game-Time Benchmarks
    a. Benchmark construction
    b. Evaluation Methods
    c. Benchmark Results
3. Conclusion
4. Future Works
    a. Time Controllable Spoken Dialogue Model - TiCo

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Motivation

What's the fundamental difference?
text assistant vs. voice assistant

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

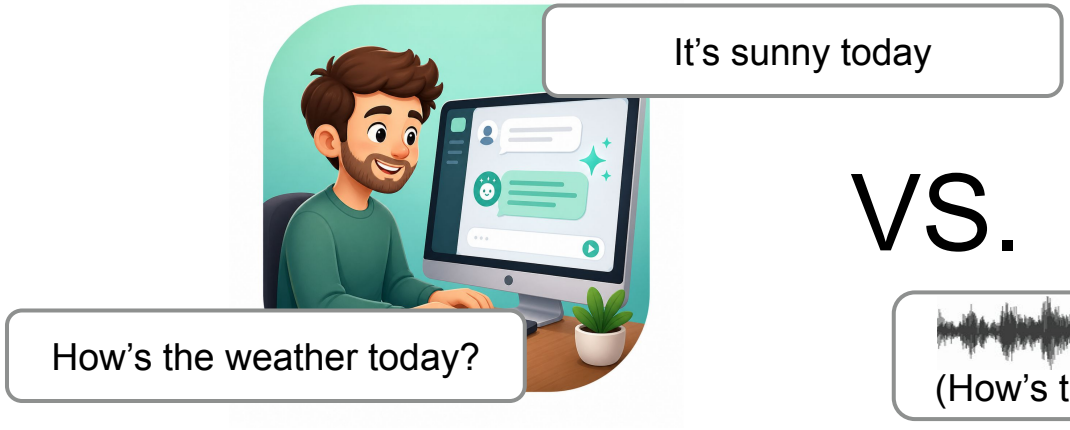
What's the fundamental difference?
text assistant vs. voice assistant

VS.

- Text input. Text output
- Speech input. Speech output

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

What's the fundamental difference?
text assistant vs. voice assistant



VS.

- Text input. Text output
- Speech input. Speech output

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

What's the fundamental difference?
text assistant vs. voice assistant

VS.

- Text input. Text output
- Emotion (emoji): 😡 😱 😢 😄
- ...

- Speech input. Speech output
- Emotion conveyed in speech
- ...

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

What's the fundamental difference?
text assistant vs. voice assistant

VS.

- Text input. Text output
- Emotion (emoji): 😡 😱 😢 😄
- ...

- Speech input. Speech output
- Emotion conveyed in speech
- ...

Text can still describe paralinguistics to a certain degree.

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

What's the fundamental difference?
text assistant vs. voice assistant

VS.

A key element making speech conversation special:

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

What's the fundamental difference?
text assistant vs. voice assistant

VS.

A key element making speech conversation special:

Time!

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

How is “time” presented in the speech conversation?

Text Conversation

Speech Conversation

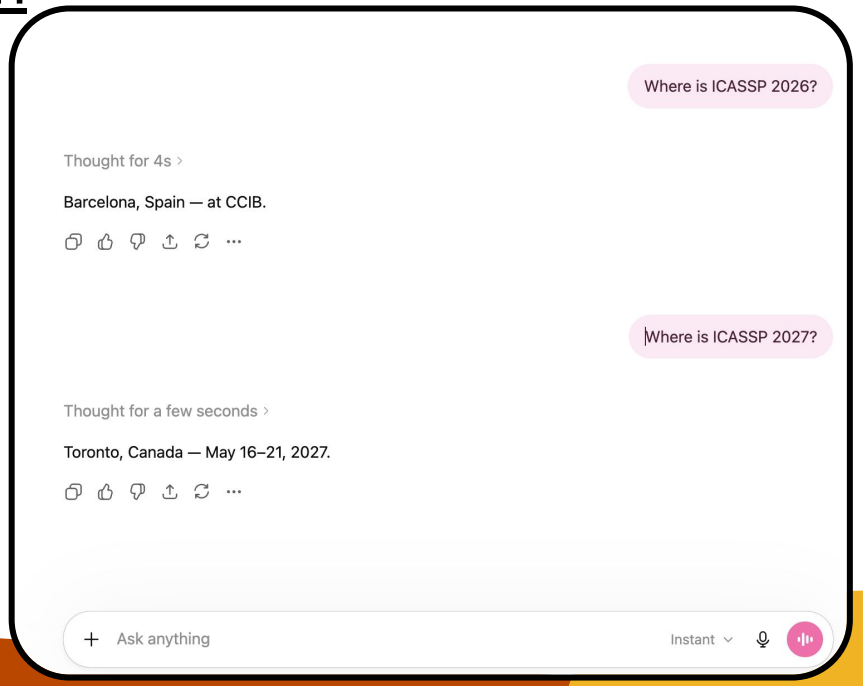
Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

How is “time” presented in the speech conversation?

Text Conversation

Speech Conversation

Turn-by-turn



User’s turn

Model’s turn

User’s turn

Model’s turn

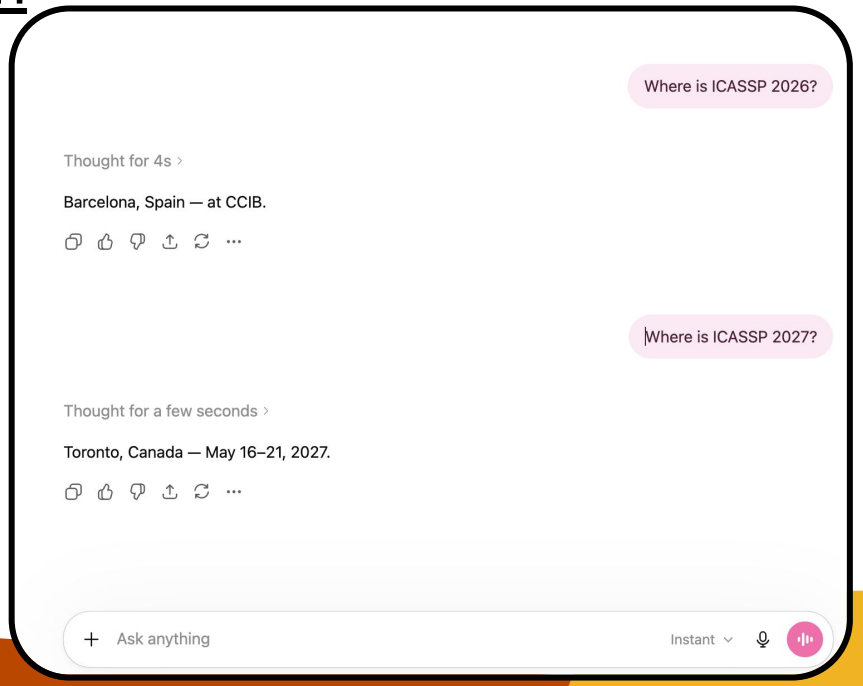
Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

How is “time” presented in the speech conversation?

Text Conversation

Speech Conversation

Turn-by-turn



User’s turn

Model’s turn

User’s turn

Model’s turn

Time interval between each turn doesn’t matter a lot

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

How is “time” presented in the speech conversation?

Text Conversation

Speech Conversation

（**Full-duplex**: Both speakers are keep listening and can speak at any time）

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

How is “time” presented in the speech conversation?

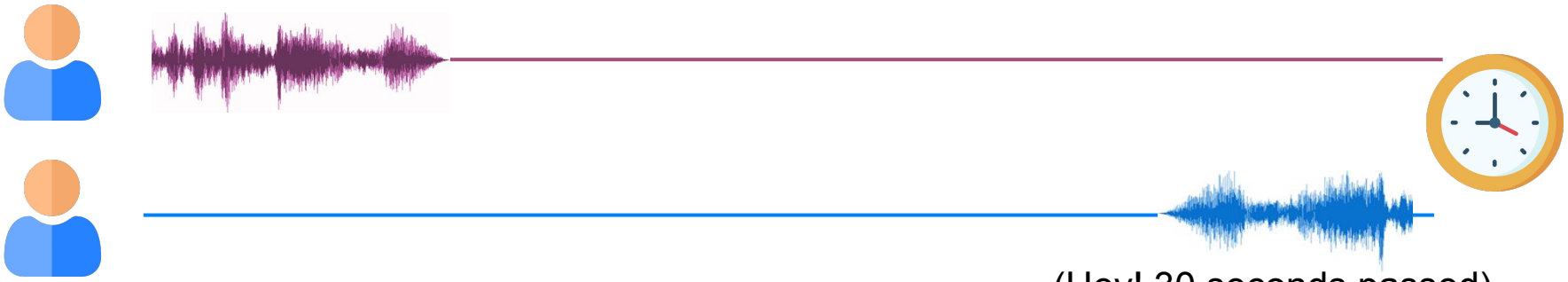
Text Conversation

Speech Conversation

（**Full-duplex**: Both speakers are keep listening and can speak at any time）

Timing

(Remind me in 30 seconds)



(Hey! 30 seconds passed)

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

How is “time” presented in the speech conversation?

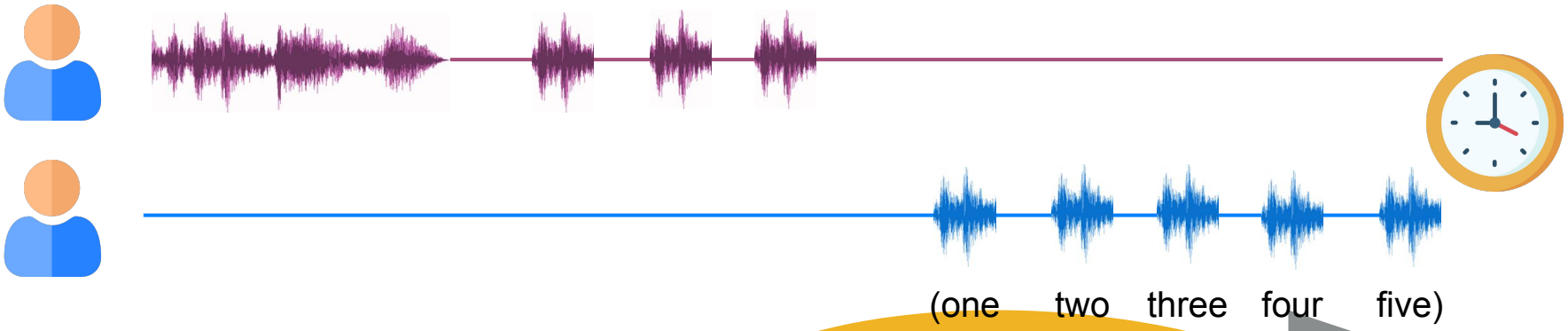
Text Conversation

Speech Conversation

（**Full-duplex**: Both speakers are keep listening and can speak at any time）

Tempo

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



Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

How is “time” presented in the speech conversation?

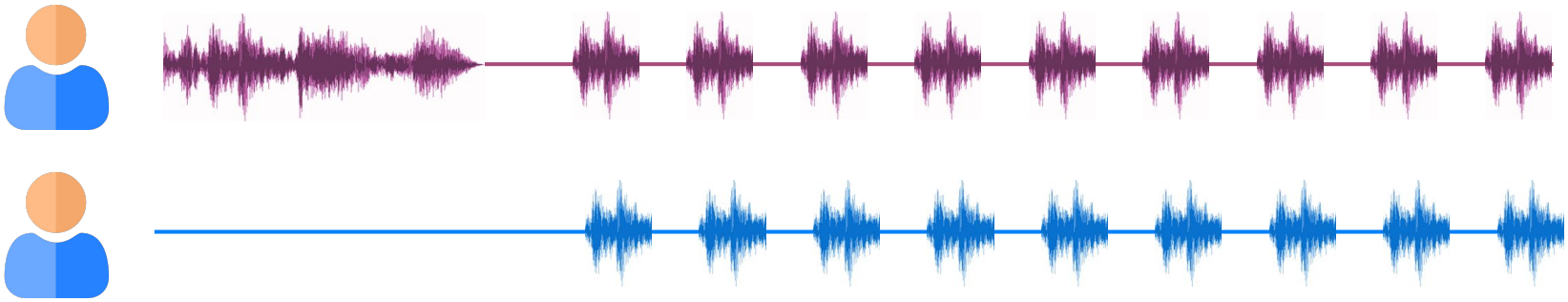
Text Conversation

Speech Conversation

Simul. Speaking (Overlapping)

（**Full-duplex**: Both speakers are keep listening and can speak at any time）

Happy new year! Let’s count down together. Ten, nine, eight...



Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark

- **Timing** - Knowing when and how long to respond
- **Tempo** - Responding at a specified tempo
- **Simul. Speaking** - Synchronizing and overlapping with the user

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark

Game-Time Benchmark

(Instruction following paradigm)

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark

(Instruction following paradigm)

Basic Tasks

- Easy tasks (e.g. counting, QA)

(Please count from 1 to 10)

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark

(Instruction following paradigm)

Basic Tasks

- Easy tasks (e.g. counting, QA)

(Please count from 1 to 10)

Advanced Tasks

- Basic tasks with time-related constraints

(Please count from 1 to 10 in 5 seconds)

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark

(Instruction following paradigm)

Basic Tasks

- Easy tasks (e.g. counting, QA)

(Please count from 1 to 10)

Advanced Tasks

- Basic tasks with time-related constraints

(Please count from 1 to 10 in 5 seconds)

Synthetic data

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Instruction Following task

Instruction following

Perform a **base task *t*** while satisfying **constraints {c_1, c_2, ...}**, where c contains **variables {N_1, N_2, ...}**.

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Instruction Following task

Instruction following

Perform a **base task *t*** while satisfying **constraints {c_1, c_2, ...}**, where c contains **variables {N_1, N_2, ...}**.

Example 1: “Please repeat after me: ‘I have a pen’”

- **Base task**: Repeat
- **Constraints**: The response should include “I have a pen”
- **Variables:** N_1 = “I have a pen”

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Instruction Following task

Instruction following

Perform a **base task *t*** while satisfying **constraints {c_1, c_2, ...}**, where c contains **variables {N_1, N_2, ...}**.

Example 2: “Please count from one to ten”

- **Base task**: Sequence generation
- **Constraints**: The response should incldue from 1 to 10 in order
- **Variables:** N_1 = 1, N_2 =10

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Instruction Following task

Instruction following

Perform a **base task *t*** while satisfying **constraints {c_1, c_2, ...}**, where c contains **variables {N_1, N_2, ...}**.

Example 3: “Please count from one to ten in five seconds”

- **Base task**: Sequence generation

- c_1: The response should incldue from 1 to 10 in order
- c_2: The response should be within 5 seconds
- **Variables:** N_1 = 1, N_2 =10. N_3=5

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark - Dataset Generation

*Step 1: Define **base tasks***

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark - Dataset Generation

Step 1: Define base tasks

- “Please repeat after me: [N1]”
- “Please count from [N1] to [N2]”
- ...

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark - Dataset Generation

*Step 1: Define **base tasks***

- “Please repeat after me: [N1]”
- “Please count from [N1] to [N2]”
- ...

Step 2: Paraphrasing

- “Repeat the following sentence: [N1]”
- “Please repeat after me: [N1]”
- ...
- “I want you to count from [N1] to
- [N2]“Please count from [N1] to [N2]”
- ...

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark - Dataset Generation

- “Repeat the following sentence: [N1]”
- “Please repeat after me: [N1]”
- ...
- “I want you to count from [N1] to
- [N2]“Please count from [N1] to [N2]”
- ...

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark - Dataset Generation

- “Repeat the following sentence: [N1]”
- “Please repeat after me: [N1]”
- ...
- “I want you to count from [N1] to
- [N2]“Please count from [N1] to [N2]”
- ...

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark - Dataset Generation

*Step 3: Augment with **constraints** and **variables***

- “Repeat the following sentence: [N1]”
- “Please repeat after me: [N1]”
- ...
- “I want you to count from [N1] to [N2]
- “Please count from [N1] to [N2]”
- ...

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark - Dataset Generation

*Step 3: Augment with **constraints** and **variables***

- "Repeat the following sentence: [N1]"
  - N1 = "I have a pen"
  - N1 = "Nice to meet you"
- "Please repeat after me: [N1]"
  - N1 = "This is a book"
  - N1 = "Good morning!"
- ...
  - ...
- "I want you to count from [N1] to [N2]
  - N1 = 1, N2 = 10
  - N1 = 5, N2 = 23
- "Please count from [N1] to [N2]"
  - N1 = 7, N2 = 12
  - N1 = 12, N2 = 20
- ...

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

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.”

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

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.”

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

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.”

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

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.”

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

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.”

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

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.”

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

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.”

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

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.”

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark

(Instruction following paradigm)

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

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Benchmark - Dataset Generation

*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)

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

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”

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

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”

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

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”

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

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”

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

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 speech instructions
Advanced Tasks: 775 speech instructions

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

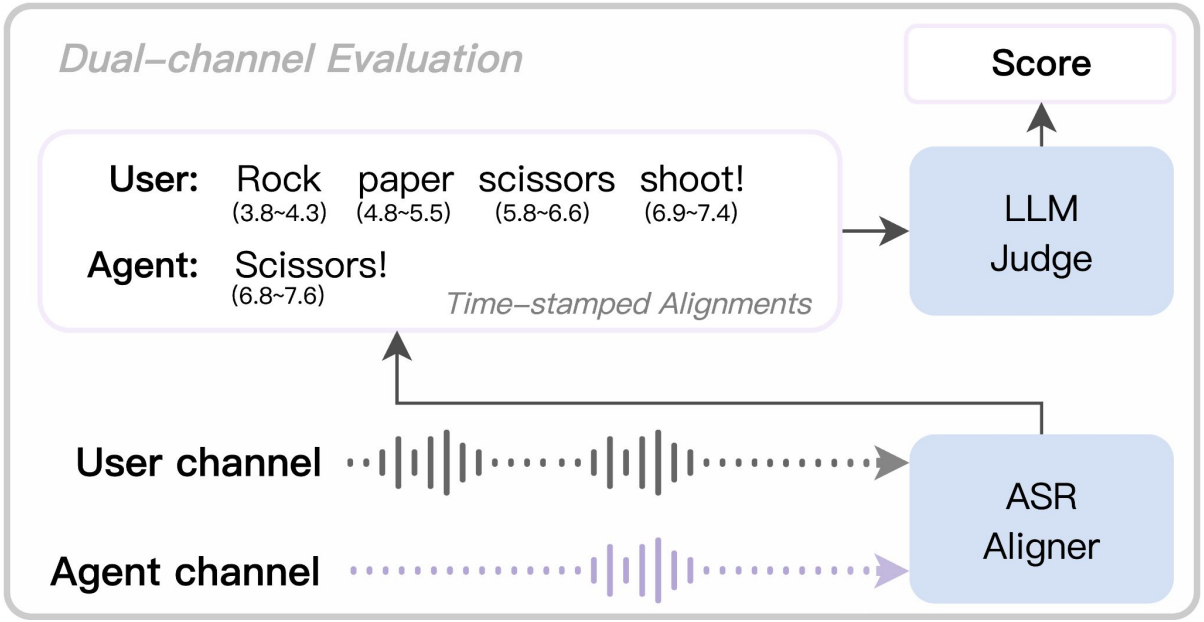
Game-Time Inspiration

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

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Evaluation

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



Gemini 2.5 Pro

Whisper medium

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Models

- Gemini-Live
- GPT-realtime

} *Commercial API*

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Models

- Gemini-Live
- GPT-realtime

Commercial API

- Moshi
- PersonaPlex NEW!
- RaonSpeech Chat NEW!

Dual-channel Full-Duplex

- Moshi: a speech-text foundation model for real-time dialogue (https://arxiv.org/abs/2410.00037)
- PersonaPlex: Voice and Role Control for Full Duplex Conversational Speech Models (https://arxiv.org/abs/2602.06053)
- Raon Speech (https://huggingface.co/KRAFTON/Raon-Speech-9B)
- Streaming Sequence-to-Sequence Learning with Delayed Streams Modeling (https://arxiv.org/pdf/2509.08753)
- Freeze-Omni: A Smart and Low Latency Speech-to-speech Dialogue Model with Frozen LLM (https://arxiv.org/abs/2411.00774)

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Models

- Gemini-Live
- GPT-realtime

Commercial API

- Moshi
- PersonaPlex NEW!
- RaonSpeech Chat NEW!

Dual-channel Full-Duplex

- Unmute (cascaded)
- Freeze-Omni

Time Multiplexing Full-Duplex

- Moshi: a speech-text foundation model for real-time dialogue (https://arxiv.org/abs/2410.00037)
- PersonaPlex: Voice and Role Control for Full Duplex Conversational Speech Models (https://arxiv.org/abs/2602.06053)
- Raon Speech (https://huggingface.co/KRAFTON/Raon-Speech-9B)
- Streaming Sequence-to-Sequence Learning with Delayed Streams Modeling (https://arxiv.org/pdf/2509.08753)
- Freeze-Omni: A Smart and Low Latency Speech-to-speech Dialogue Model with Frozen LLM (https://arxiv.org/abs/2411.00774)

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

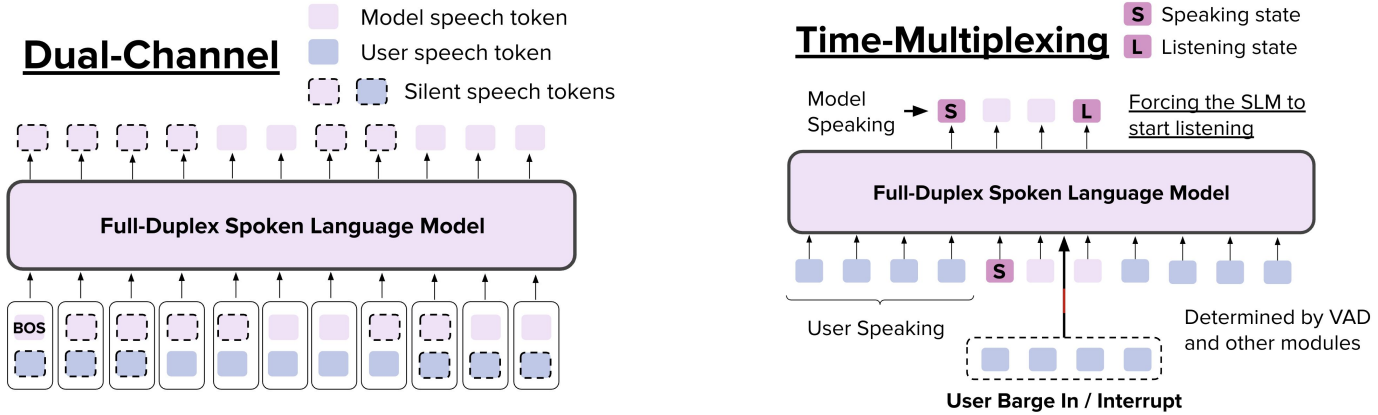
Models

- Gemini-Live
- GPT-realtime

Commercial API

- Moshi
- PersonaPlex
- [illegible]

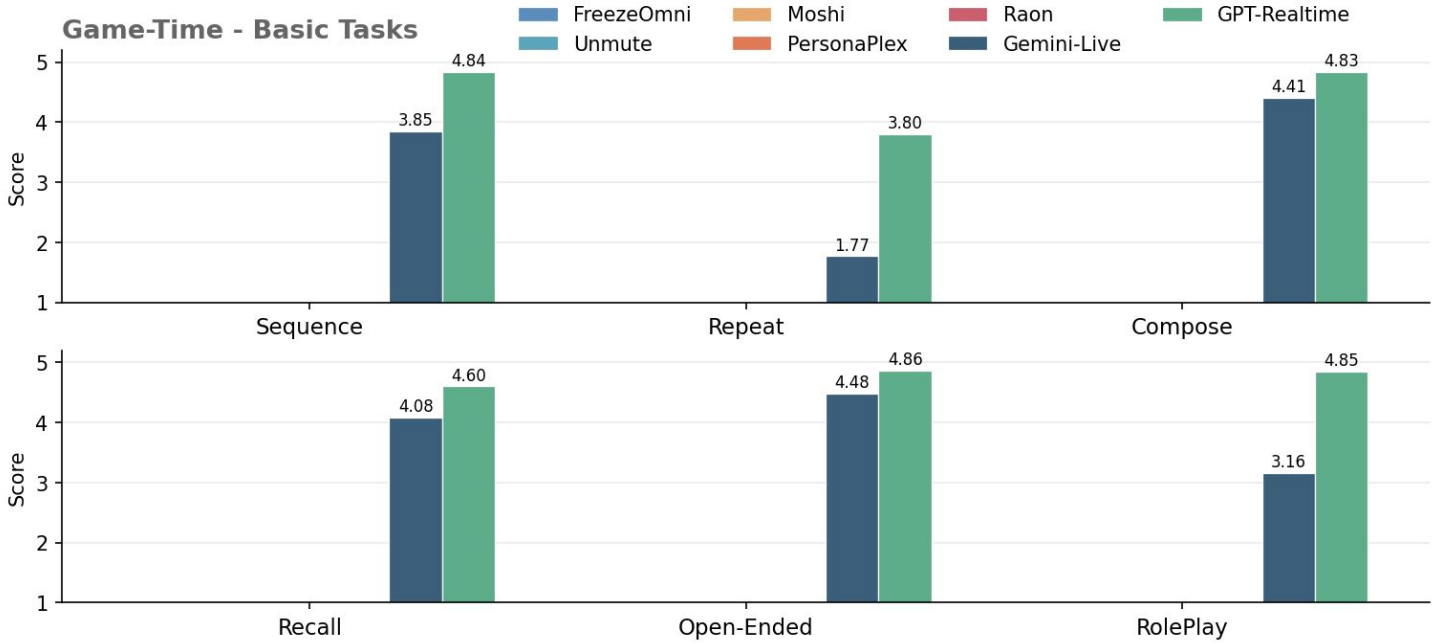
Dual-channel Full-Duplex



Ref: *On the Landscape of Spoken Language Models* (https://arxiv.org/abs/2504.08528, TMLR 2025) / Interspeech 2026 Tutorial

Game-Time Benchmark - Results

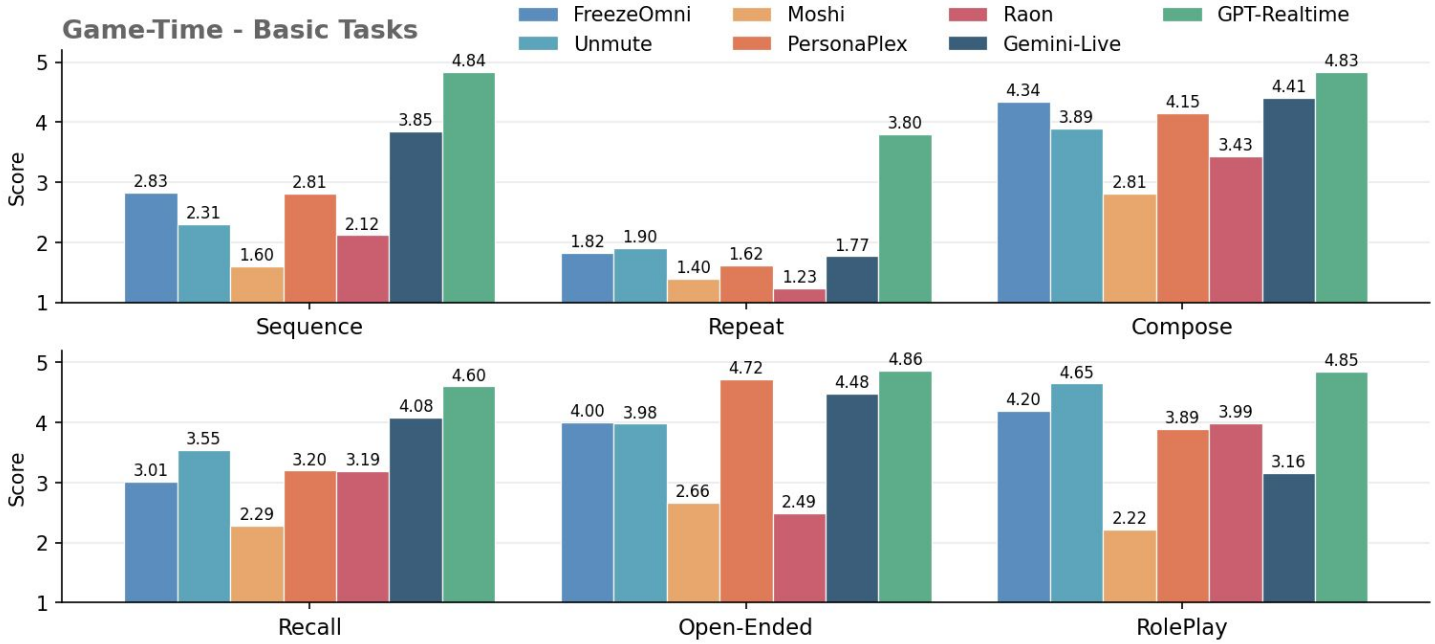
Game-Time Basic Tasks Results



Commercial API performs well on basic tasks

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

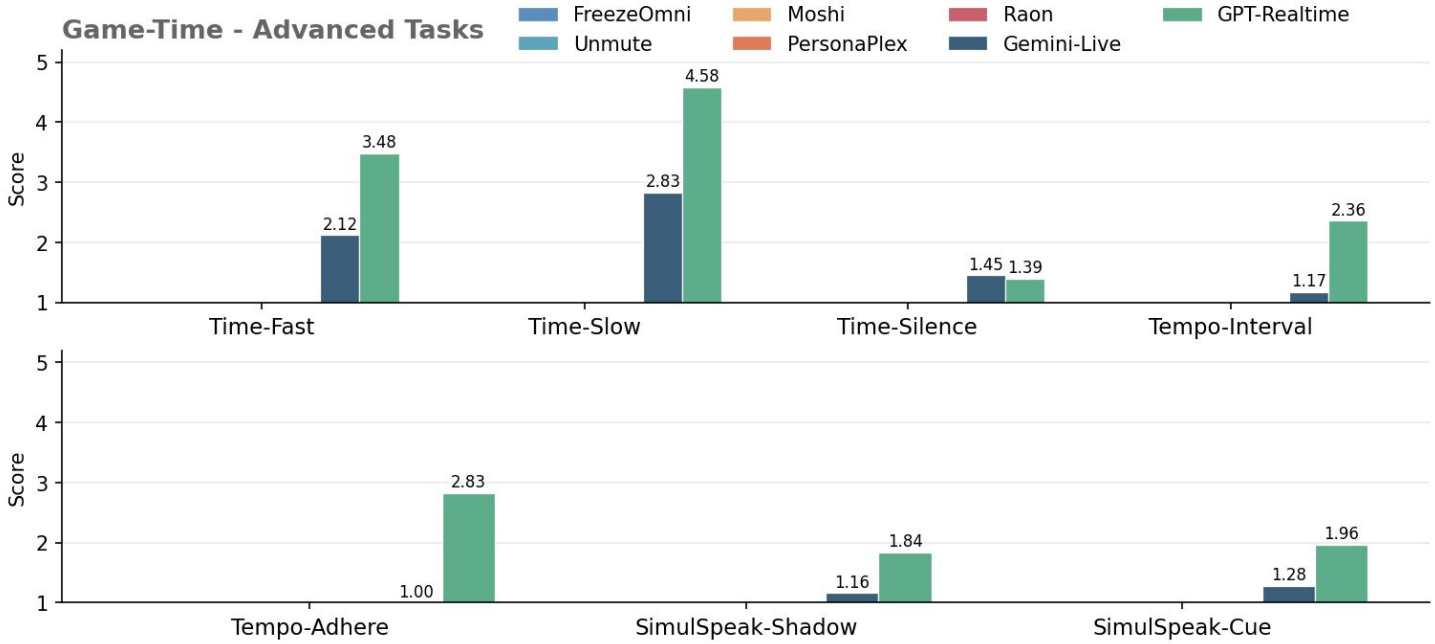
Game-Time Basic Tasks Results



Open-sourced models < Commercial API

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

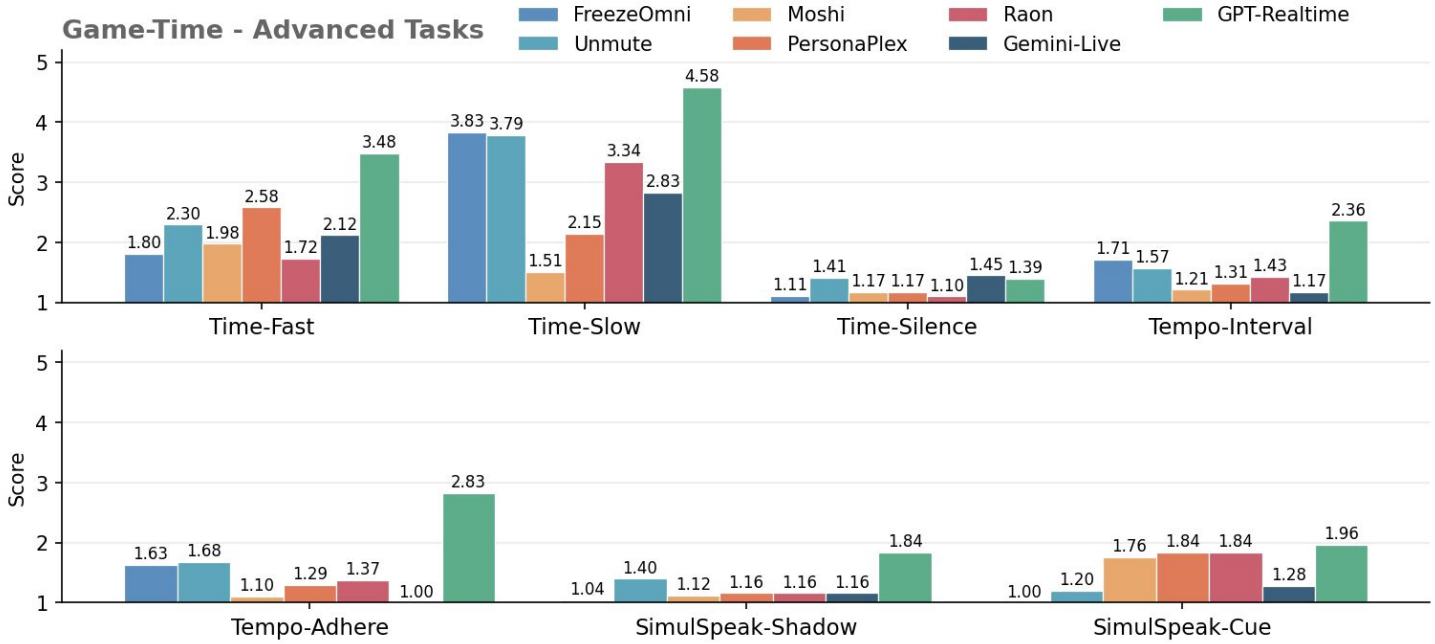
Game-Time Advanced Tasks Results



After introducing time related constraint, the model degrades a lot

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Game-Time Advanced Tasks Results



After introducing time related constraint, the model degrades a lot

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Correlation with Human Judgment

NEW!

open sourced

Judge Pair	Spearman's ρ	Pearson's r
Human - LLM (Gemini Pro)	**0.68**	**0.68**
Human - Qwen3 Omni (text)	0.58	0.57
Human - Qwen3 Omni (text + audio)	0.65	0.67

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Correlation with Human Judgment

NEW!

open sourced

Judge Pair	Spearman's ρ	Pearson's r
Human - LLM (Gemini Pro)	**0.68**	**0.68**
Human - Qwen3 Omni (text)	0.58	0.57
Human - Qwen3 Omni (text + audio)	0.65	0.67

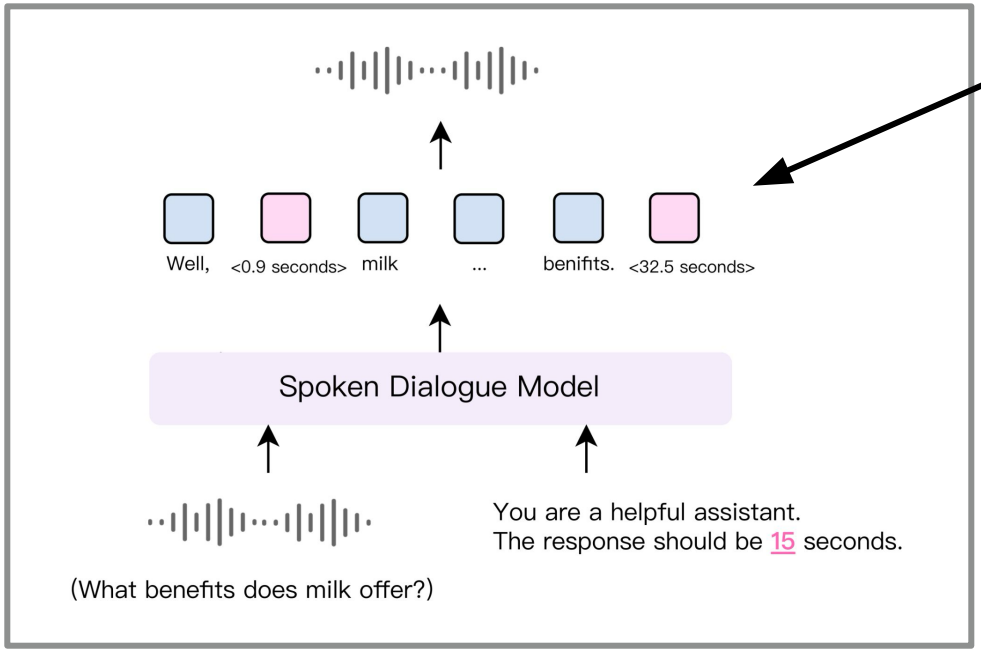
Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Conclusion

- The SOTA SLMs can excel on most basic tasks
  * However, some SLMs (e.g. Moshi) still cannot
- All SLMs do not have reasonable performance when time constraints are introduced
- SLMs cann’t perform on some tasks that are easy to human!

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Future Work - TiCo

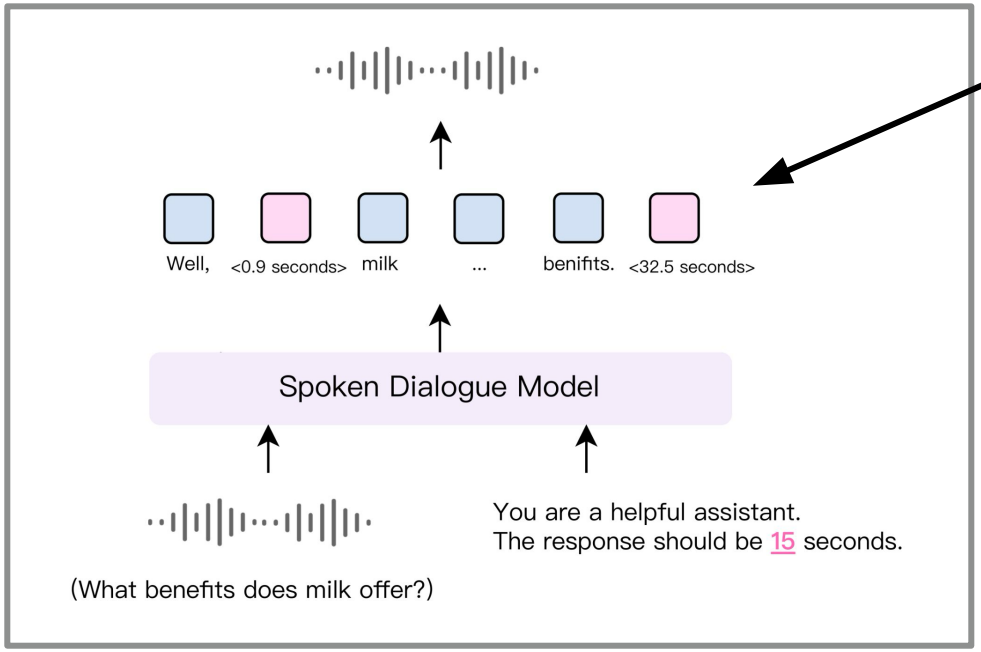


Generates
“time token”

TiCo: Time-Controllable Training for Spoken Dialogue Models
(https://arxiv.org/abs/2603.22267)

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Future Work - TiCo



Generates
“time token”

- Able to follow time-control instruction - time awareness
- TiCo is a half-duplex model
- Still can not solve most Game-Time tasks
- A step towards time-aware SLM!

TiCo: Time-Controllable Training for Spoken Dialogue Models
(https://arxiv.org/abs/2603.22267)

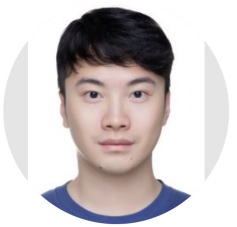
Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>

Collaborators

Kai-Wei Chang
(MIT)

En-Pei Hu
(NTU)

Chun-Yi Kuan
(NTU)



Wenze Ren
(NTU)

Wei-Chih Chen
(NTU)

Guan-Ting Lin
(NTU)

Yu Tsao
(Academia Sinica)

Shao-Hua Sun
(NTU AI-CoRE)

Hung-yi Lee
(NTU AI-CoRE)

James Glass
(MIT)

Kai-Wei Chang, MIT CSAIL <kwchang@mit.edu>


Thanks for
your listening