

MoshiRAG

Asynchronous Knowledge Retrieval for Full-Duplex Speech Language Models

Chung-Ming Chien ¹, Manu Orsini ², Eugene Kharitonov ², Neil Zeghidour ², Karen Livescu ¹, Alexandre Défossez ²
TTIC ¹, Kyutai ²



Voice AI

Voice

AI

Voice AI

Realtime

Smooth turn-taking

Conversational

Interactive

Interruptible

Voice AI

Realtime

Smooth turn-taking

Conversational

Interactive

Interruptible

Natural

Emotional

Expressive

Human-like

Voice AI

Voice AI

Smart
Factual
Reasoning
Tool-use
Problem-solving

Voice AI

Smart
Factual
Reasoning
Tool-use
Problem-solving

Reliable
Personalized
Helpful
Agentic

Outline

Outline

- Background

Outline

- Background
- System design

Outline

- Background
- System design
- Training

Outline

- Background
- System design
- Training
- Experiments

Outline

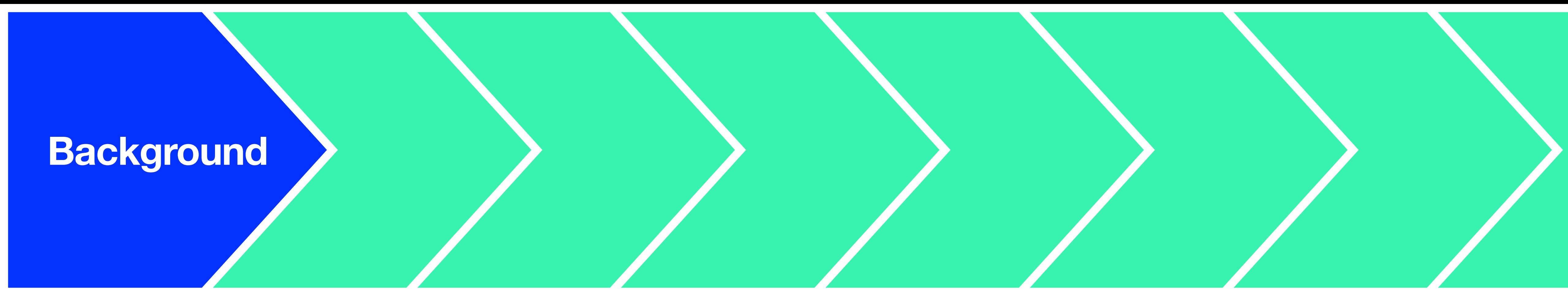
- Background
- System design
- Training
- Experiments
- Conclusion

Outline

- Background
- System design
- Training
- Experiments
- Conclusion
- Demo

Outline

- Background
- System design
- Training
- Experiments
- Conclusion
- Demo
- What's next?



Background

Voice AI

From turn-based to full-duplex models

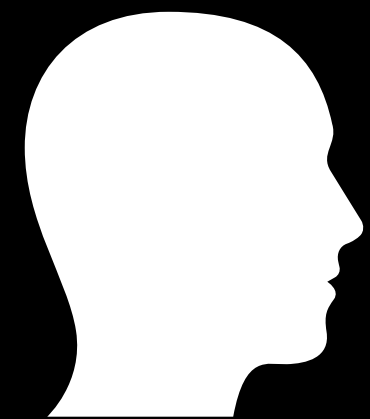
Voice AI

From turn-based to full-duplex models

Turn-based model

Voice AI

From turn-based to full-duplex models



Turn-based model

Voice AI

From turn-based to full-duplex models

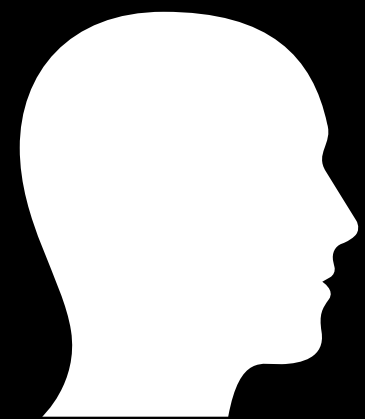
Turn-based model



Voice AI

From turn-based to full-duplex models

Turn-based model



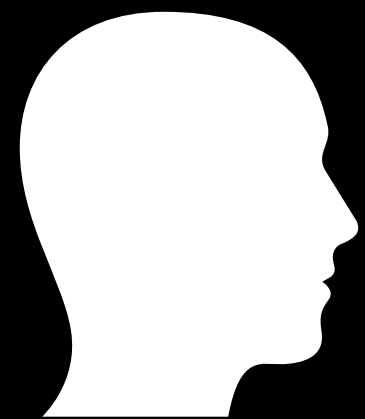
End-of-turn

Voice AI

From turn-based to full-duplex models



Turn-based model



End-of-turn

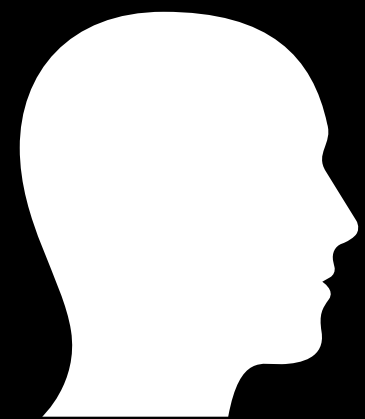
Voice AI

From turn-based to full-duplex models



Turn-based model

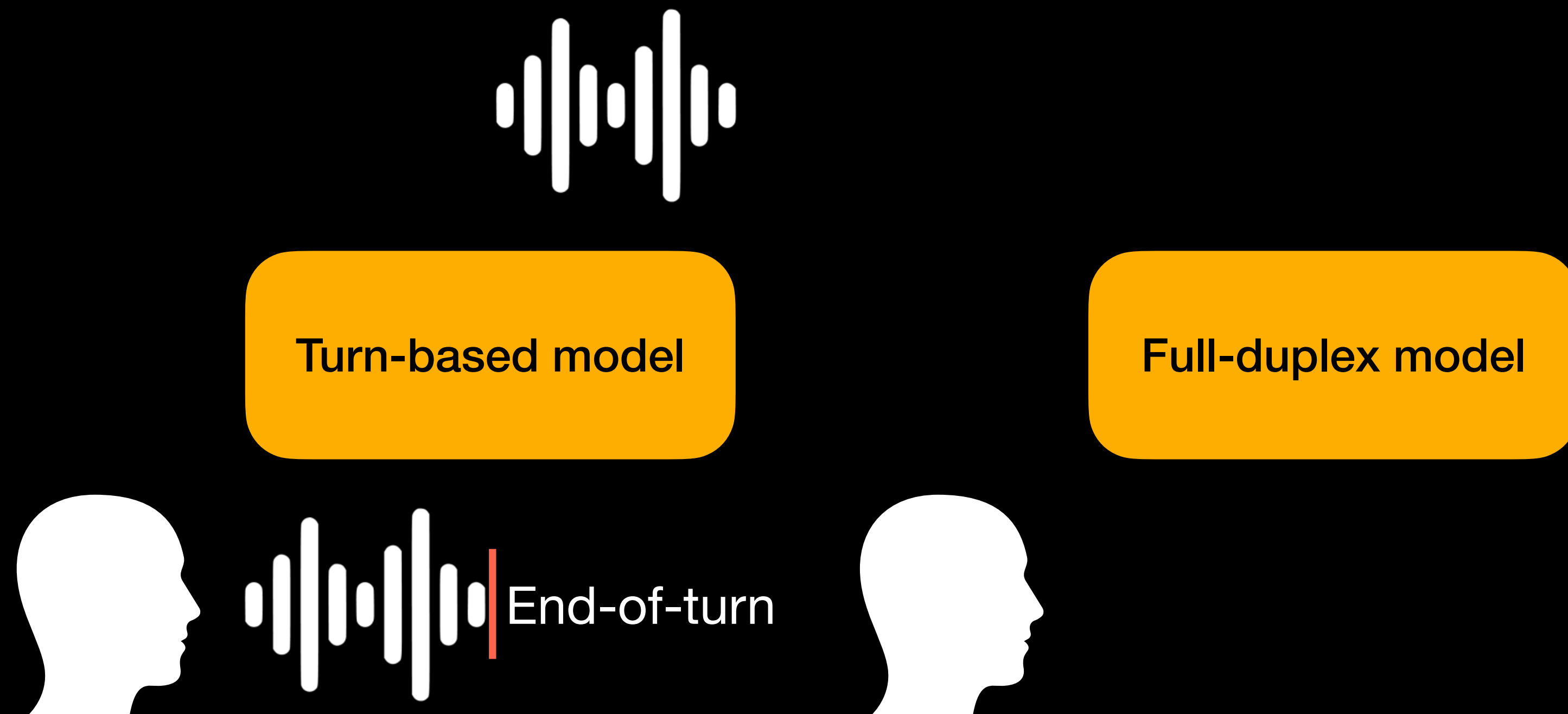
Full-duplex model



End-of-turn

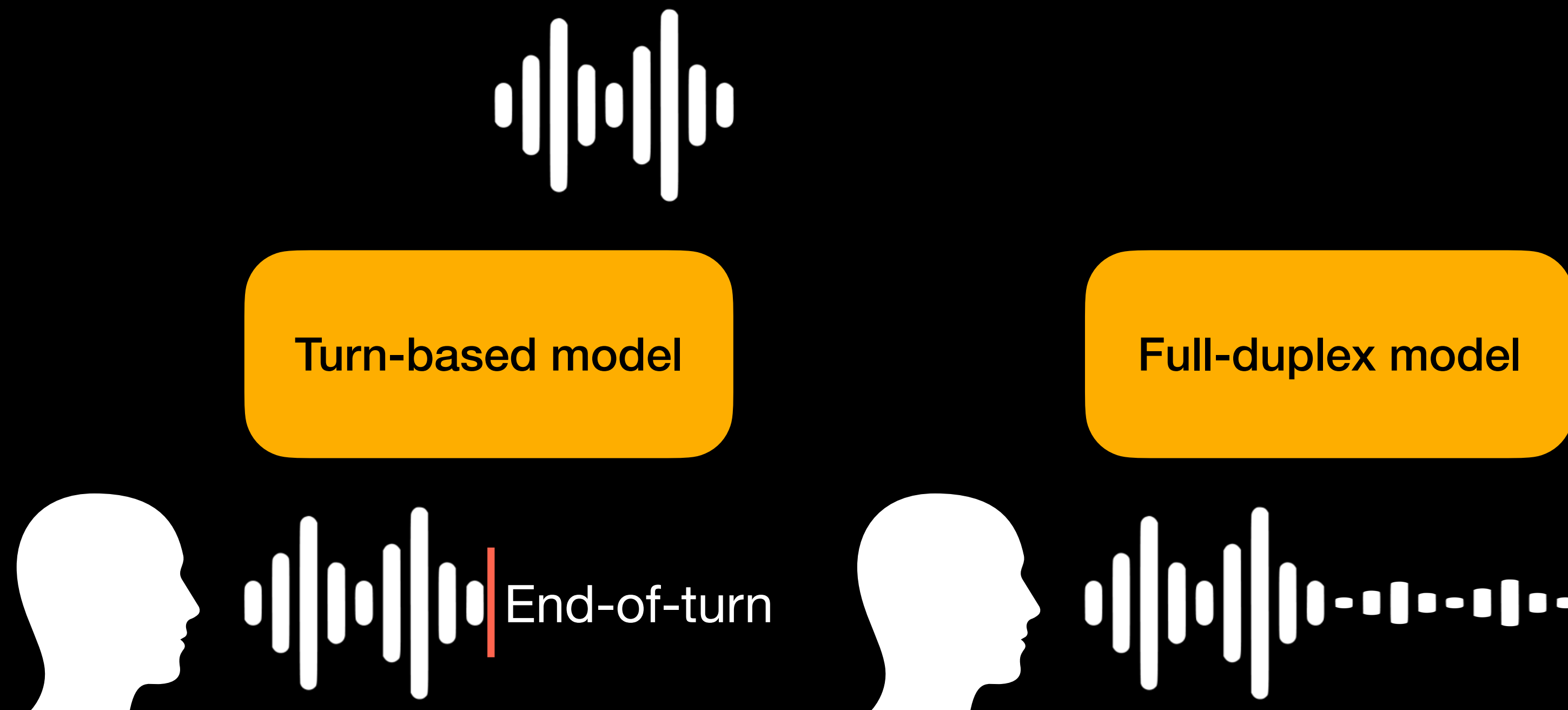
Voice AI

From turn-based to full-duplex models



Voice AI

From turn-based to full-duplex models



Voice AI

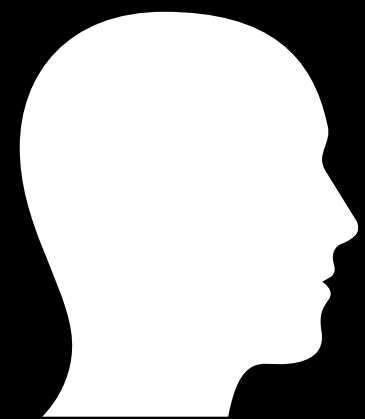
From turn-based to full-duplex models



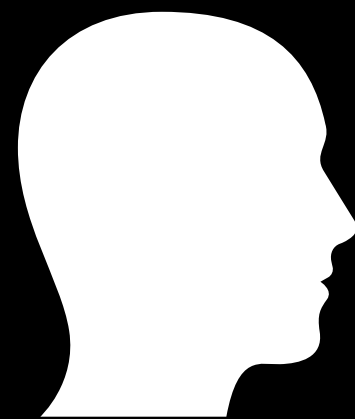
Turn-based model



Full-duplex model



End-of-turn



Voice AI

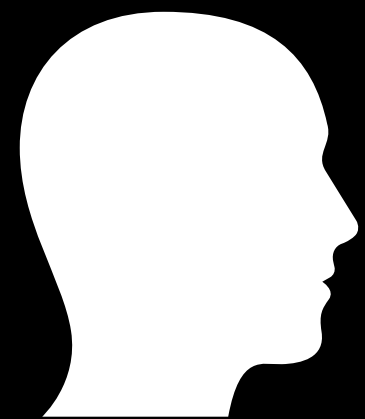
From turn-based to full-duplex models



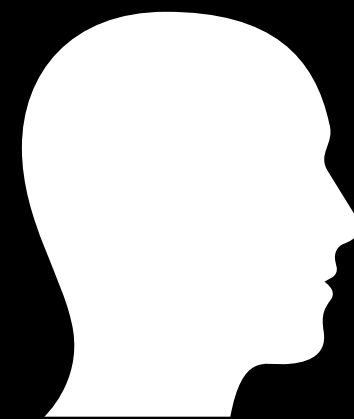
Turn-based model



Full-duplex model



End-of-turn



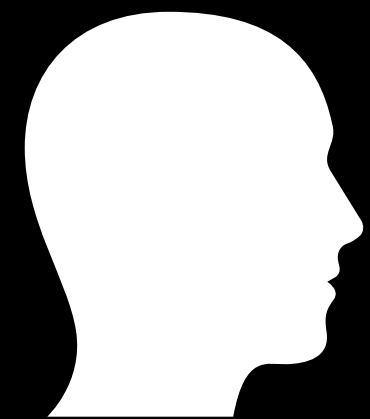
e.g. siri since 2010,
Qwen3-Omni, Kimi-Audio,
and GPT-4o Audio?

Voice AI

From turn-based to full-duplex models



Turn-based model

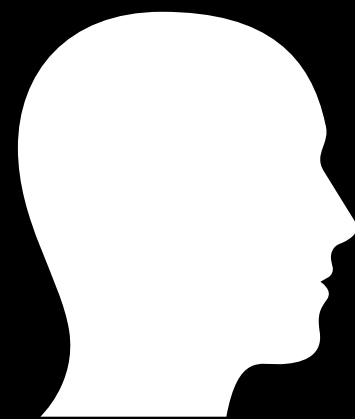


End-of-turn

e.g. siri since 2010,
Qwen3-Omni, Kimi-Audio,
and GPT-4o Audio?



Full-duplex model



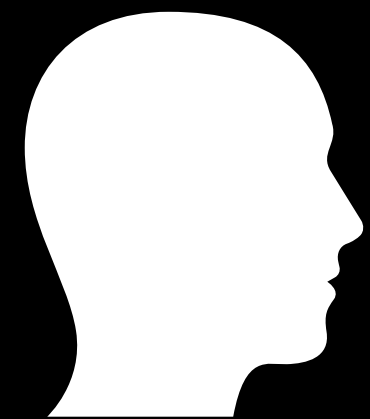
e.g. Moshi

Voice AI

From turn-based to full-duplex models



Turn-based model

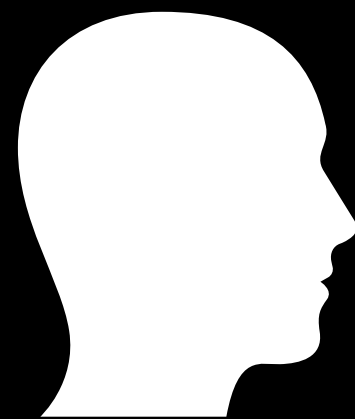


End-of-turn

e.g. siri since 2010,
Qwen3-Omni, Kimi-Audio,
and GPT-4o Audio?



Full-duplex model



e.g. Moshi

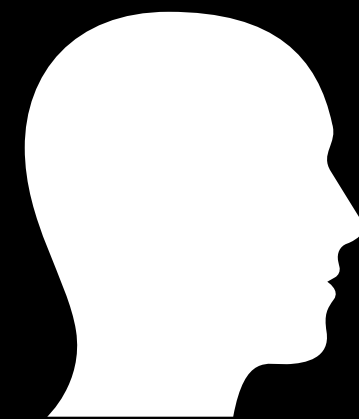
Why full-duplex?

Voice AI

From turn-based to full-duplex models



Turn-based model

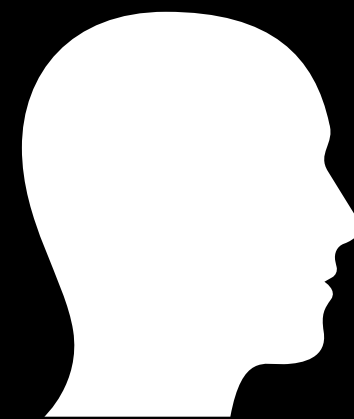


End-of-turn

e.g. siri since 2010,
Qwen3-Omni, Kimi-Audio,
and GPT-4o Audio?



Full-duplex model



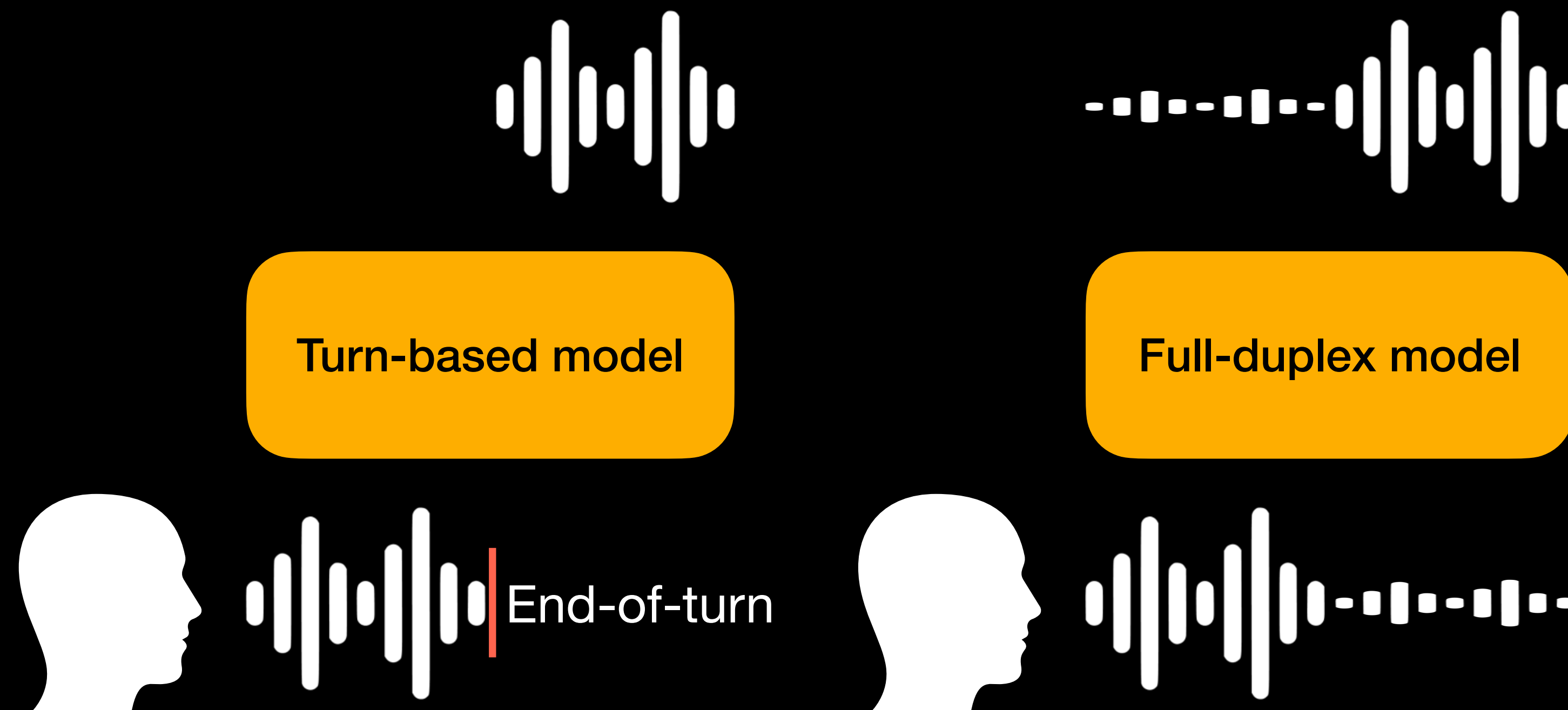
e.g. Moshi

Why full-duplex?

- Learning “when to speak/stop” end-to-end

Voice AI

From turn-based to full-duplex models



e.g. siri since 2010,
Qwen3-Omni, Kimi-Audio,
and GPT-4o Audio?

e.g. Moshi

Why full-duplex?

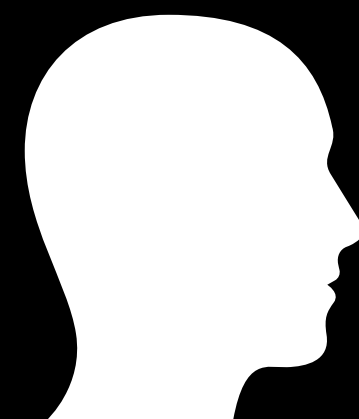
- Learning “when to speak/stop” end-to-end
- Fewer modules
→ lower latency

Voice AI

From turn-based to full-duplex models



Turn-based model

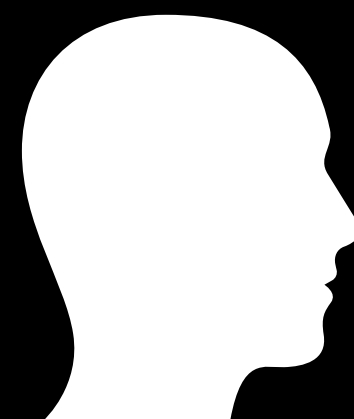


End-of-turn

e.g. siri since 2010,
Qwen3-Omni, Kimi-Audio,
and GPT-4o Audio?



Full-duplex model



e.g. Moshi

Why full-duplex?

- Learning “when to speak/stop” end-to-end
- Fewer modules
→ lower latency
- This is how humans communicate

Moshi: a Pioneer Full-Duplex Model in the LLM Era

Moshi: a Pioneer Full-Duplex Model in the LLM Era

An autoregressive LM

Moshi: a Pioneer Full-Duplex Model in the LLM Era

An autoregressive LM

- that listens while speaking with extremely low latency

Moshi: a Pioneer Full-Duplex Model in the LLM Era

An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM

Moshi: a Pioneer Full-Duplex Model in the LLM Era

An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

Moshi: a Pioneer Full-Duplex Model in the LLM Era

Moshi

An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

Moshi: a Pioneer Full-Duplex Model in the LLM Era

An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

Moshi

User speech unit

Moshi: a Pioneer Full-Duplex Model in the LLM Era

Moshi speech unit

An autoregressive LM

Moshi

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

User speech unit

Moshi: a Pioneer Full-Duplex Model in the LLM Era

Moshi speech unit

Moshi text
(aligned with speech)

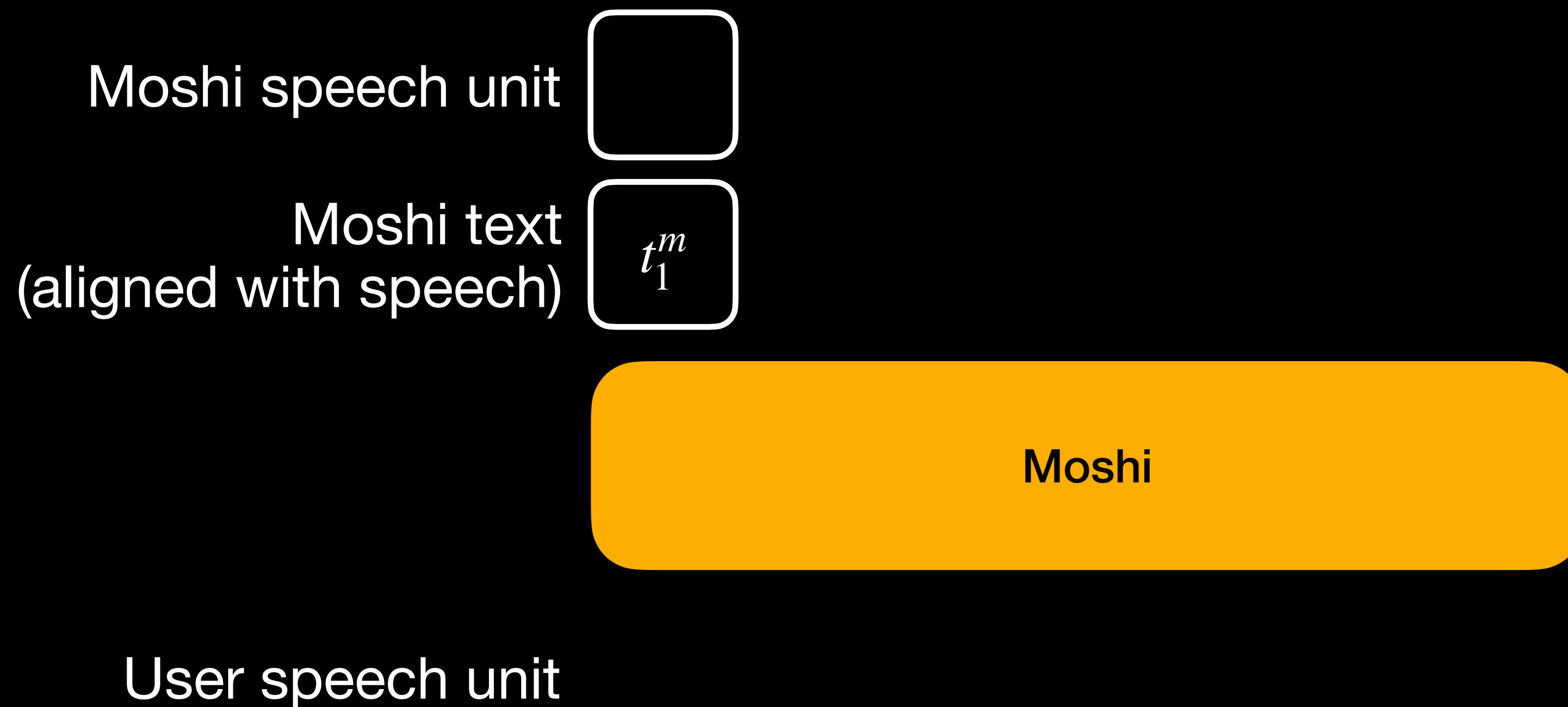
Moshi

User speech unit

An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

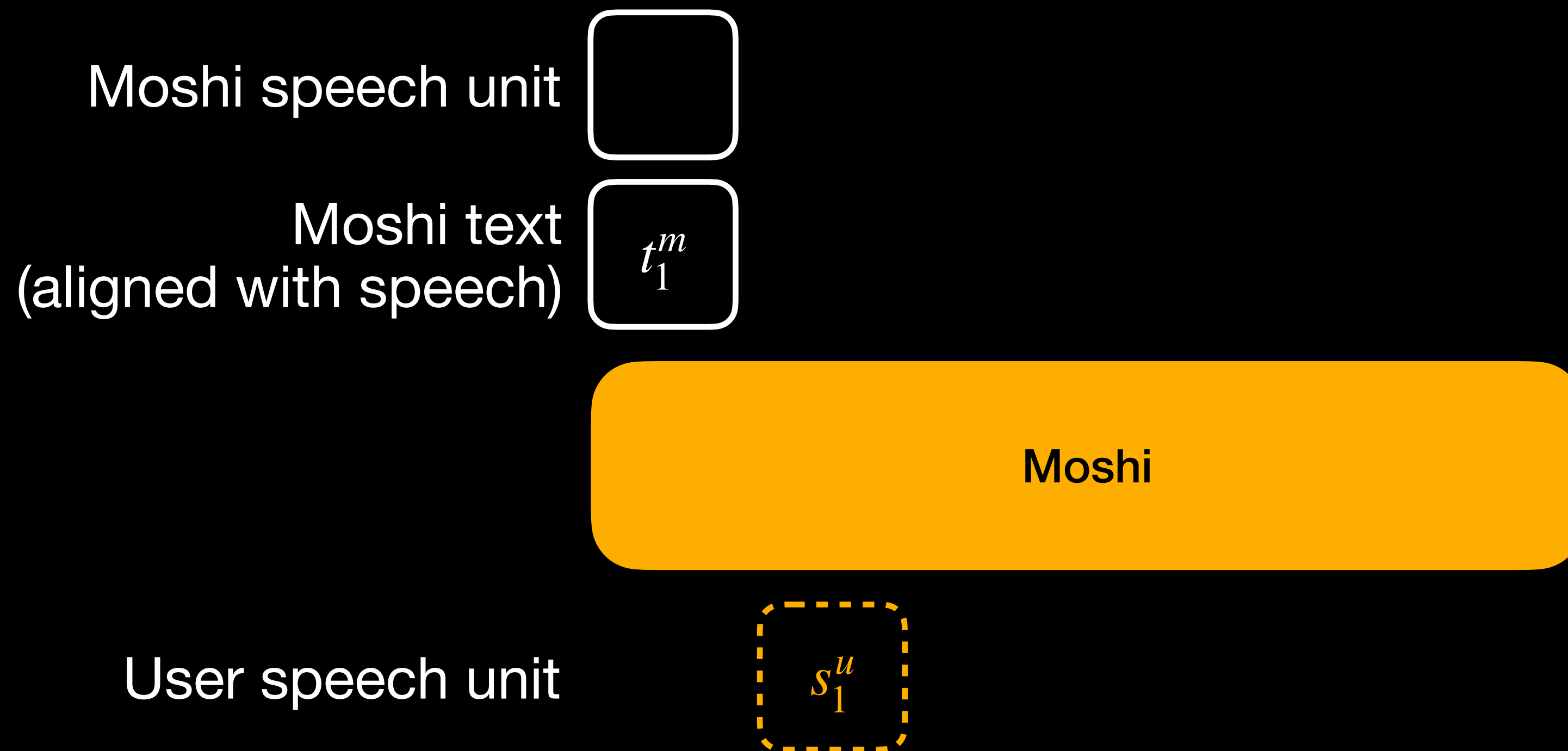
Moshi: a Pioneer Full-Duplex Model in the LLM Era



An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

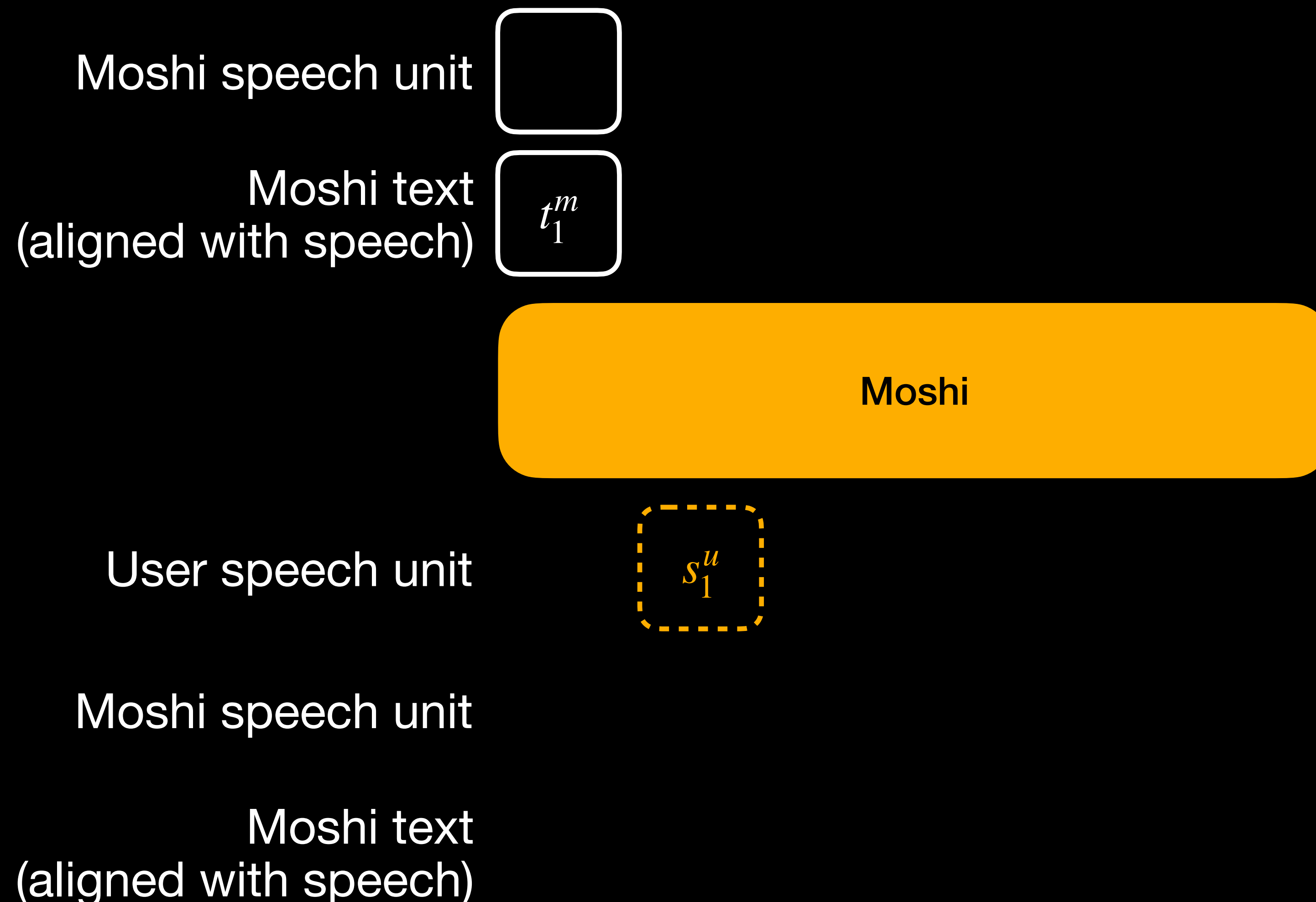
Moshi: a Pioneer Full-Duplex Model in the LLM Era



An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

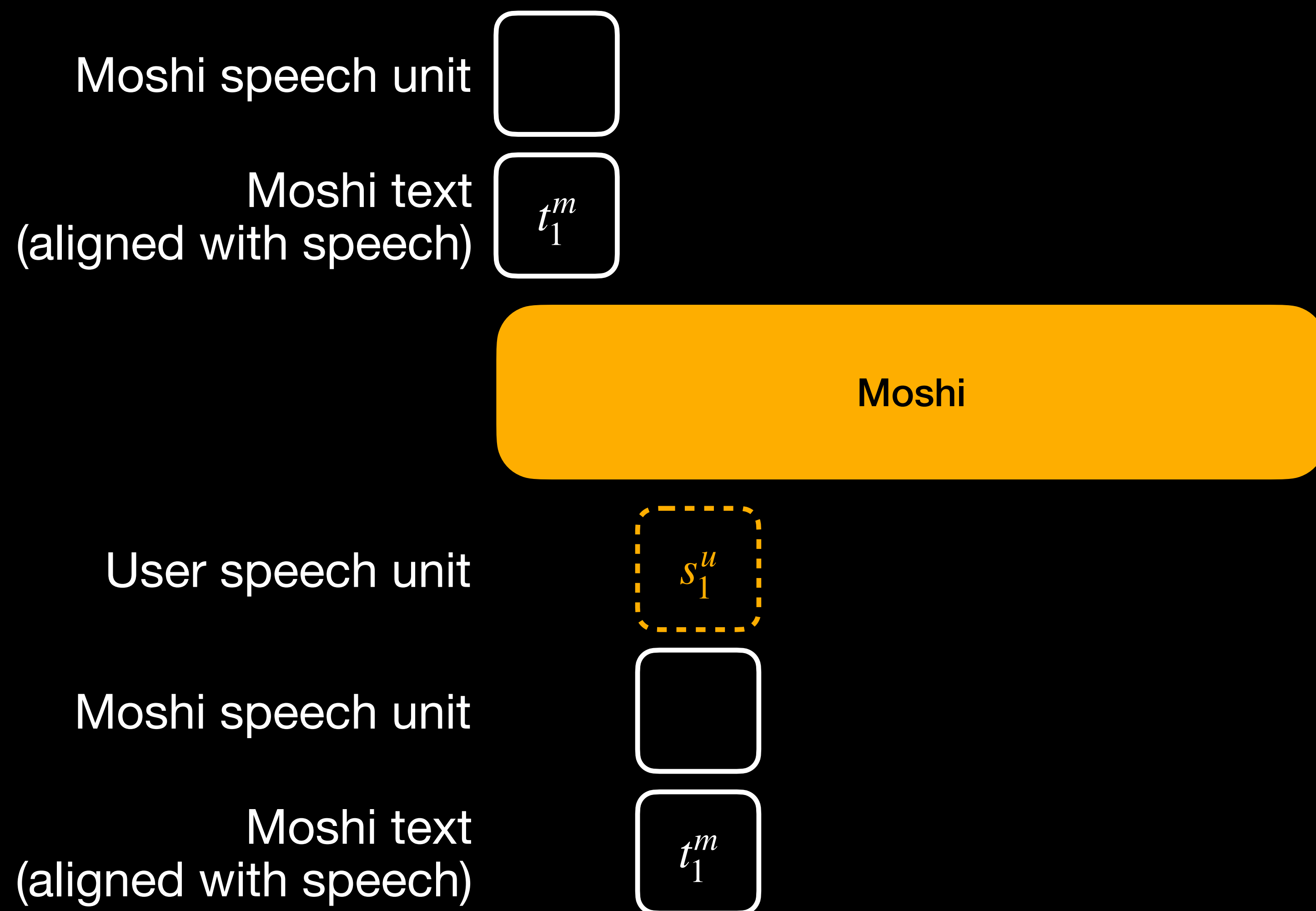
Moshi: a Pioneer Full-Duplex Model in the LLM Era



An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

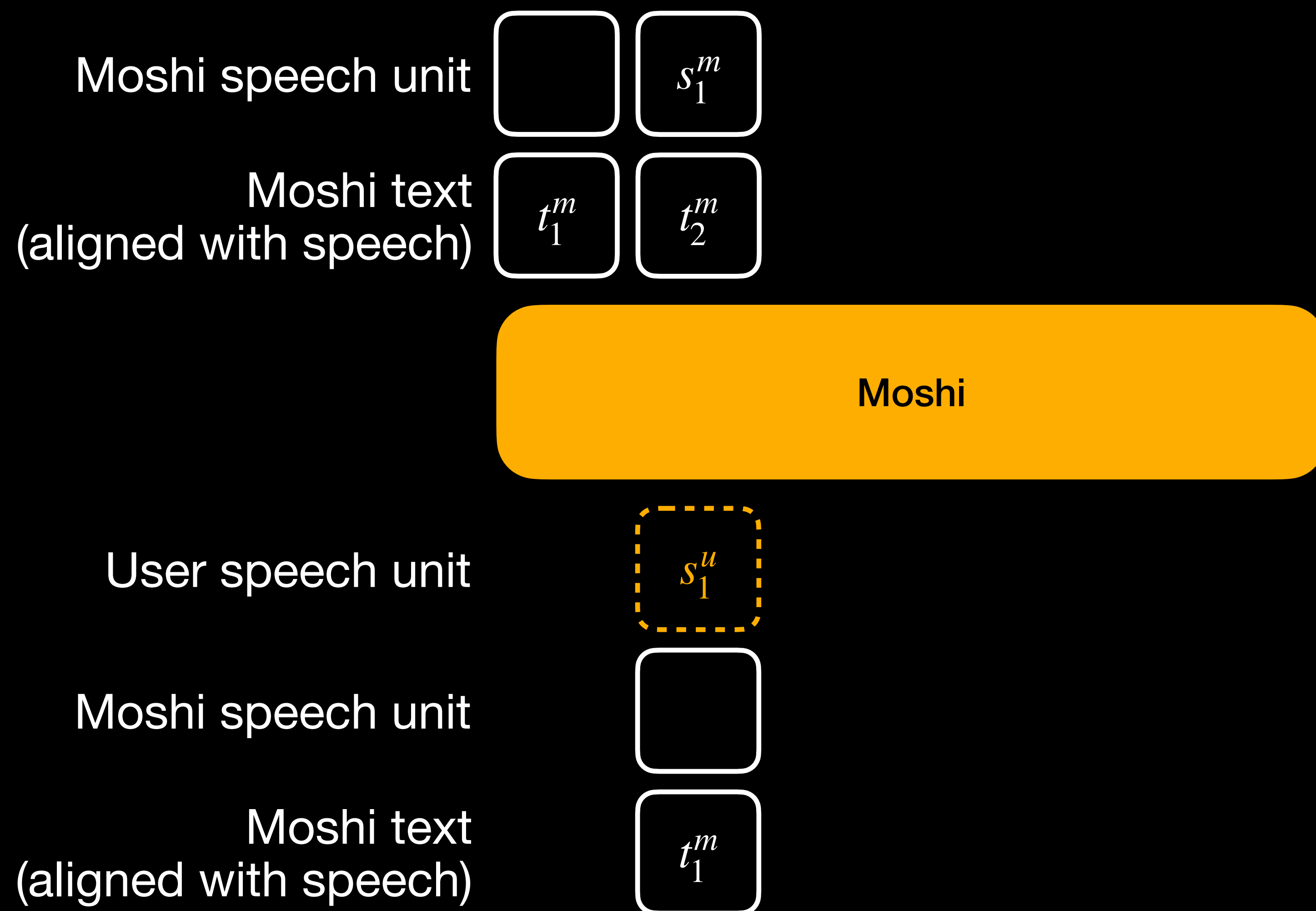
Moshi: a Pioneer Full-Duplex Model in the LLM Era



An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

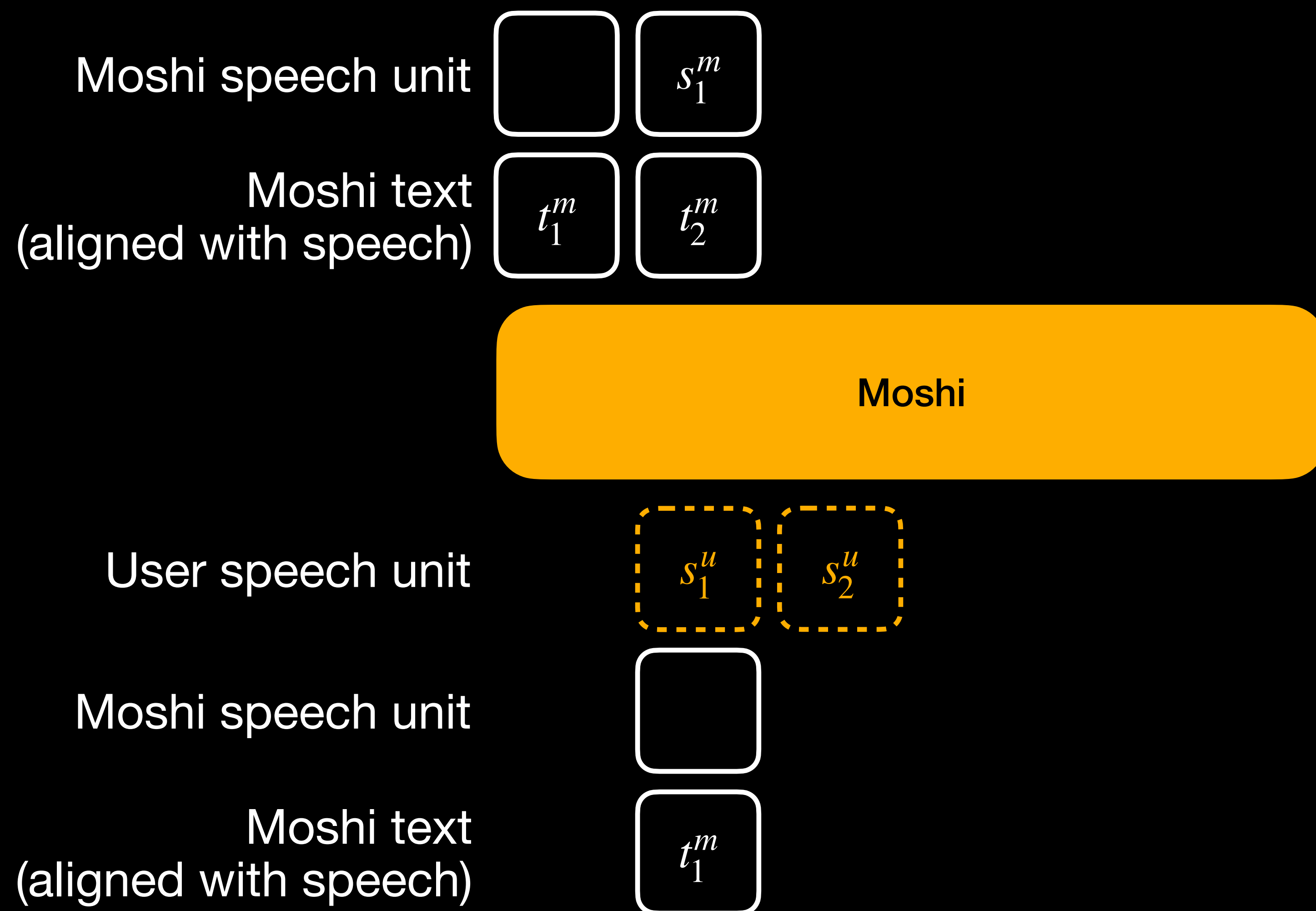
Moshi: a Pioneer Full-Duplex Model in the LLM Era



An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

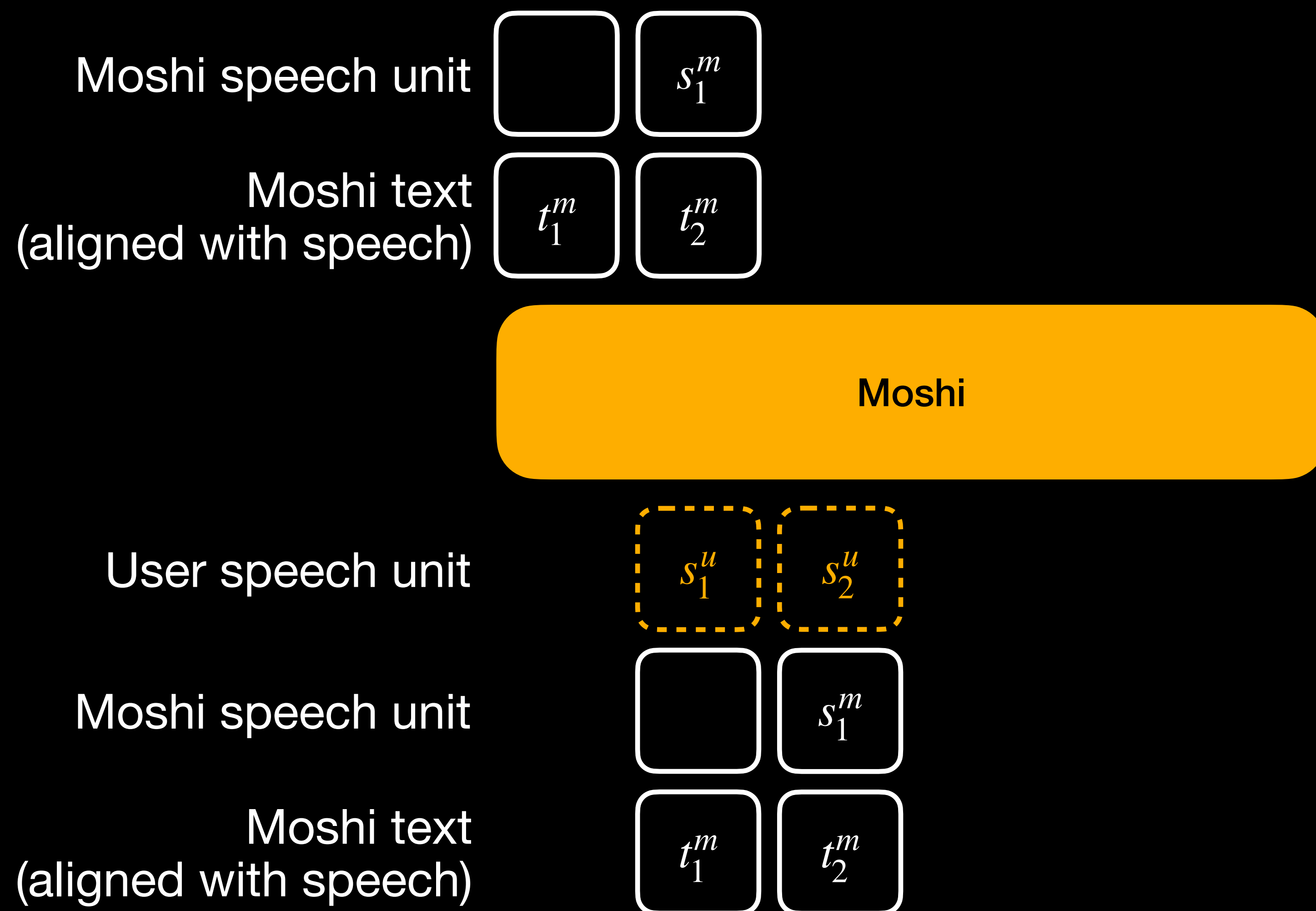
Moshi: a Pioneer Full-Duplex Model in the LLM Era



An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

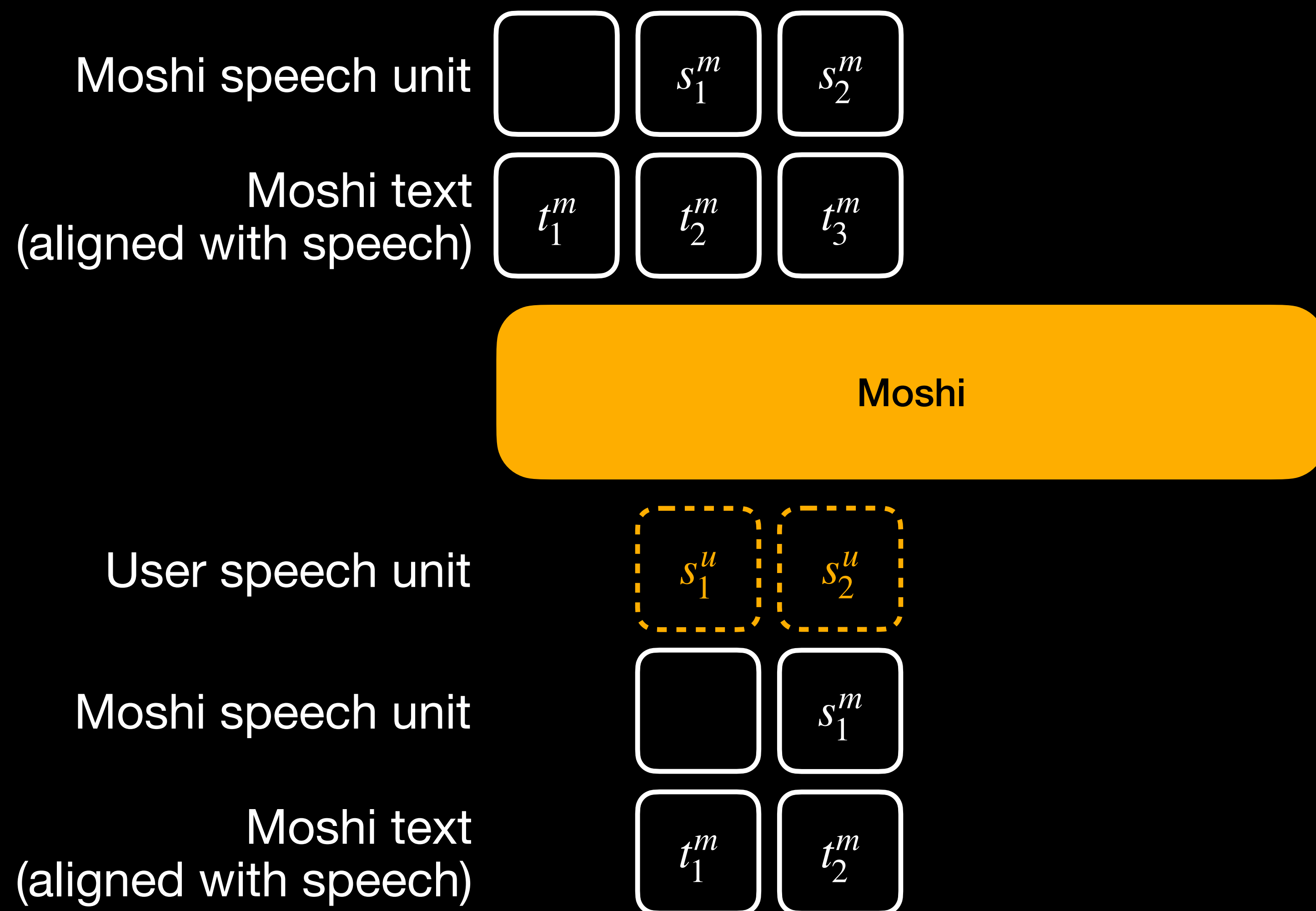
Moshi: a Pioneer Full-Duplex Model in the LLM Era



An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

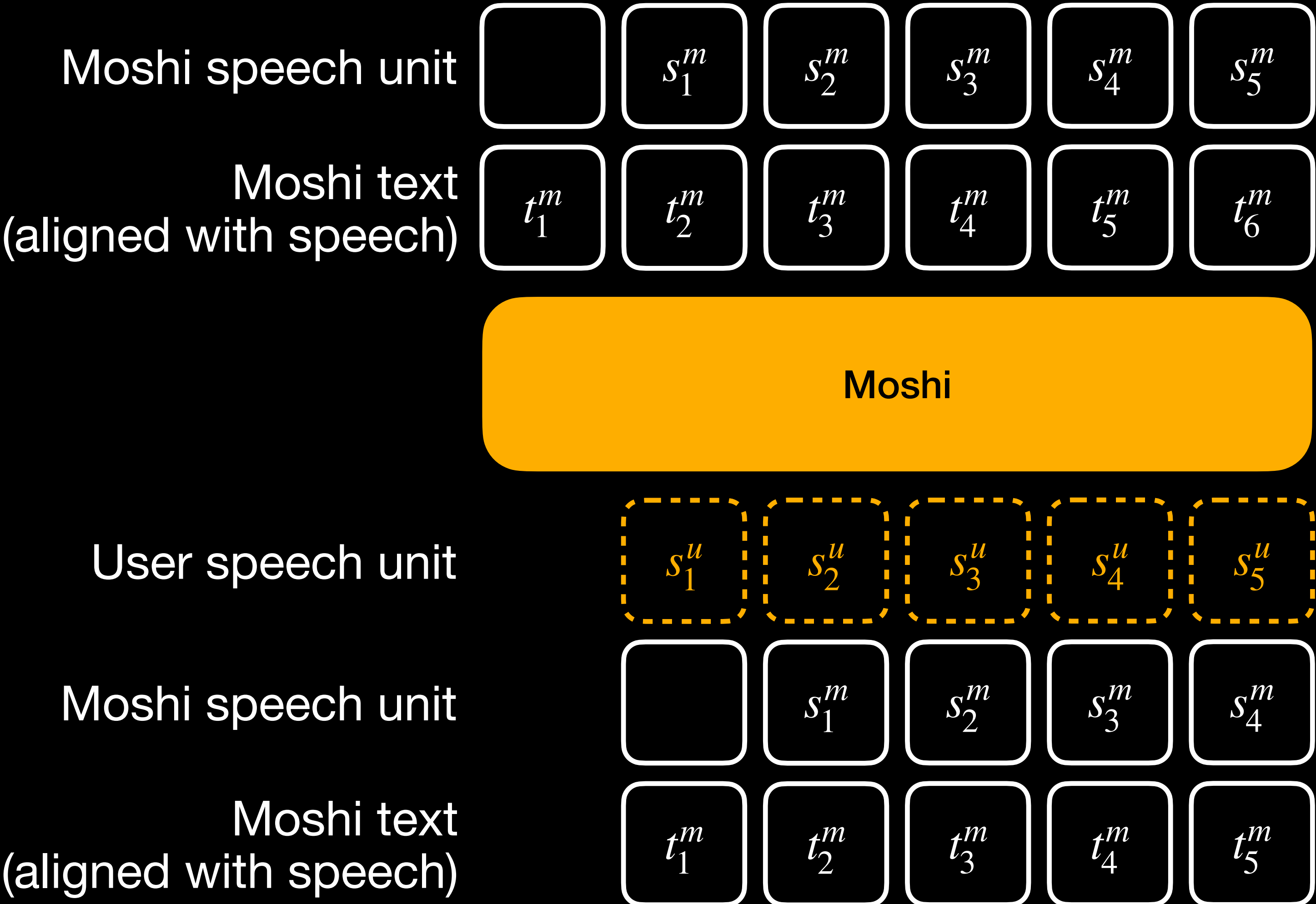
Moshi: a Pioneer Full-Duplex Model in the LLM Era



An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

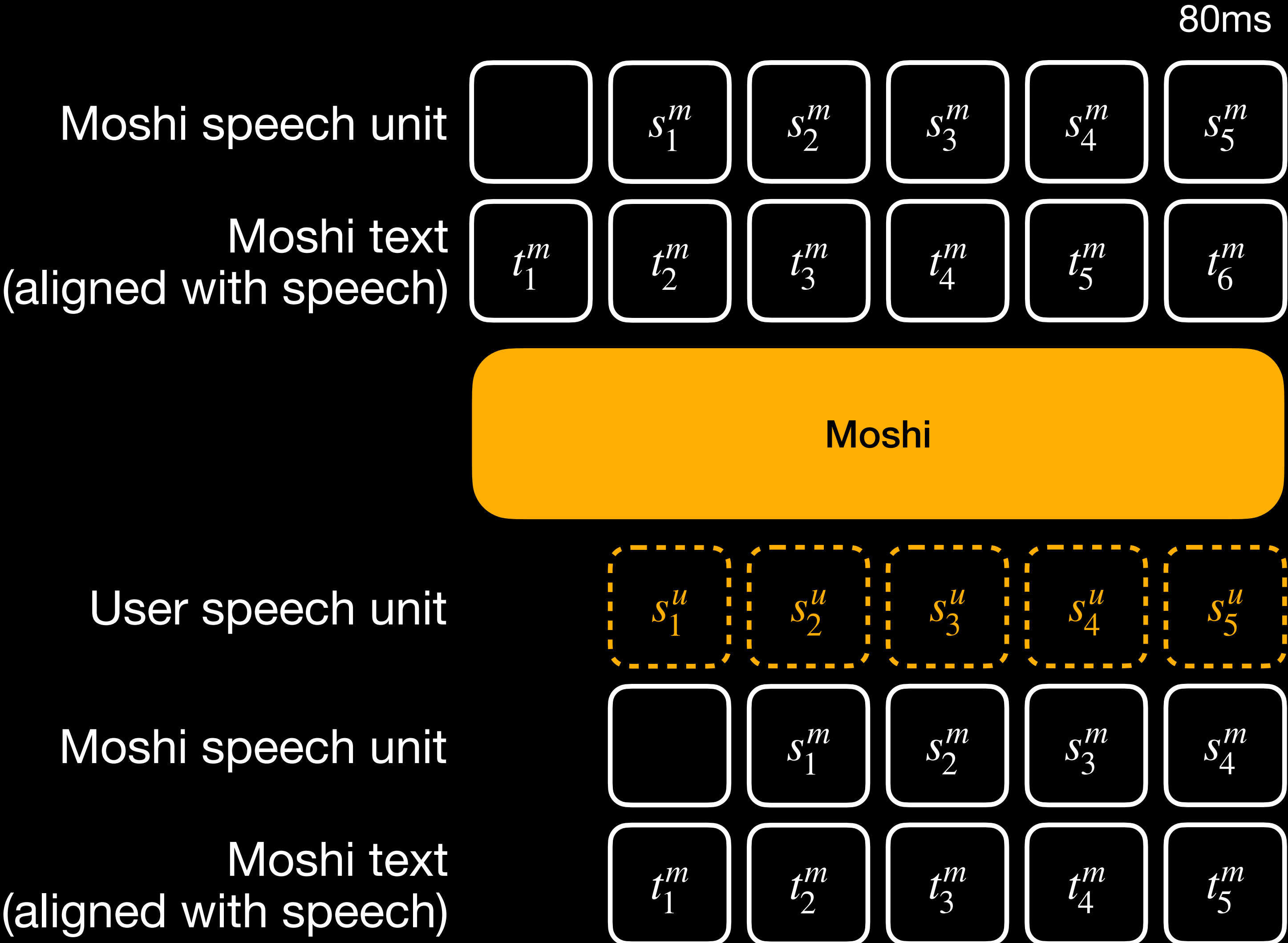
Moshi: a Pioneer Full-Duplex Model in the LLM Era



An autoregressive LM

- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

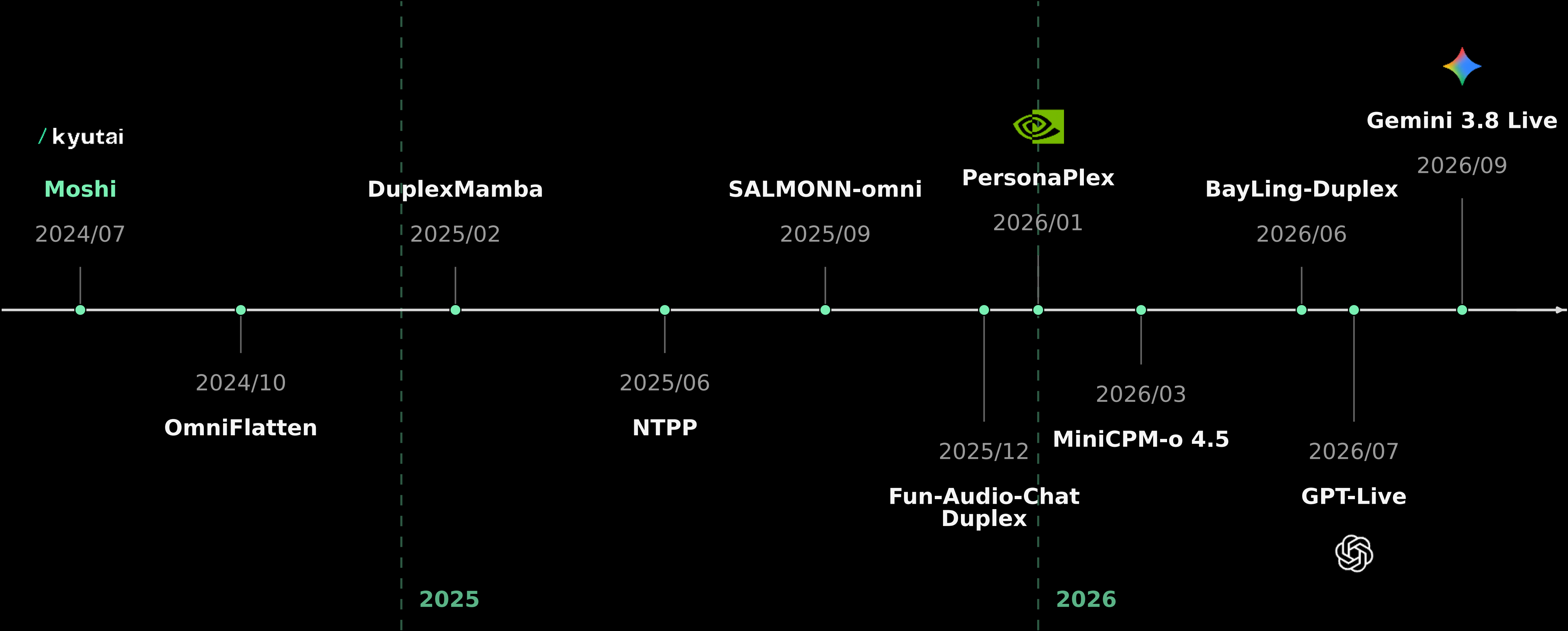
Moshi: a Pioneer Full-Duplex Model in the LLM Era



An autoregressive LM

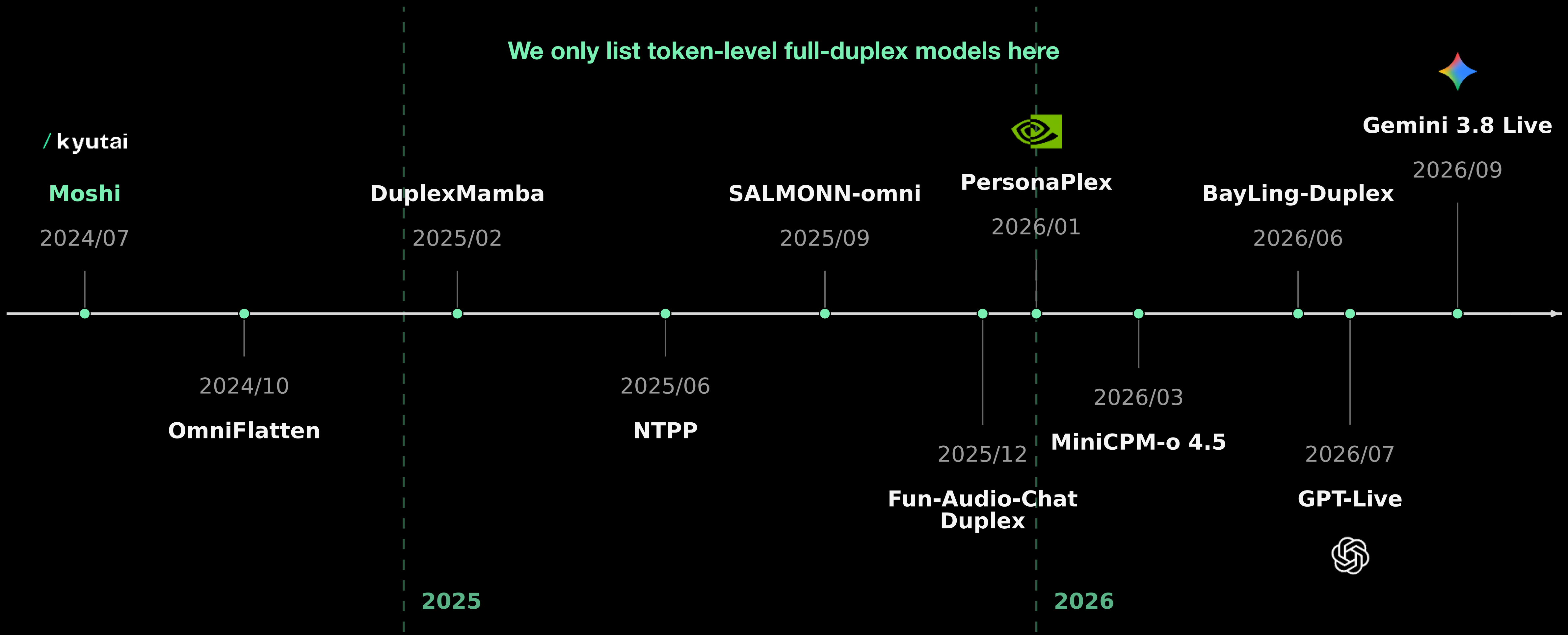
- that listens while speaking with extremely low latency
- built by fined-tuning a pre-trained text LM
- to models multi-channel text/speech inputs and outputs

A Brief Timeline of Full-Duplex Models Since Moshi



A Brief Timeline of Full-Duplex Models Since Moshi

We only list token-level full-duplex models here



Full-Duplexity Sounds Sexy, But...

Full-Duplexity Sounds Sexy, But...

Moshi

Me

Full-Duplexity Sounds Sexy, But...

Moshi

Me

Hey, how are you doing? I' doing well. How about

Full-Duplexity Sounds Sexy, But...

Moshi

Hey, how are you doing? I' doing well. How about

Me

- Who is the current president of France?

Full-Duplexity Sounds Sexy, But...

Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second

Me

- Who is the current president of France?

Full-Duplexity Sounds Sexy, But...

Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second

Me

- Who is the current president of France?
- When was his second term election?

Full-Duplexity Sounds Sexy, But...

Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with 57.55% of the vote, defeating the National Front's Marine Le Pen.

Me

- Who is the current president of France?
- When was his second term election?

Full-Duplexity Sounds Sexy, But...

Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with 57.55% of the vote, defeating the National Front's Marine Le Pen.

Me

- Who is the current president of France?
- When was his second term election?

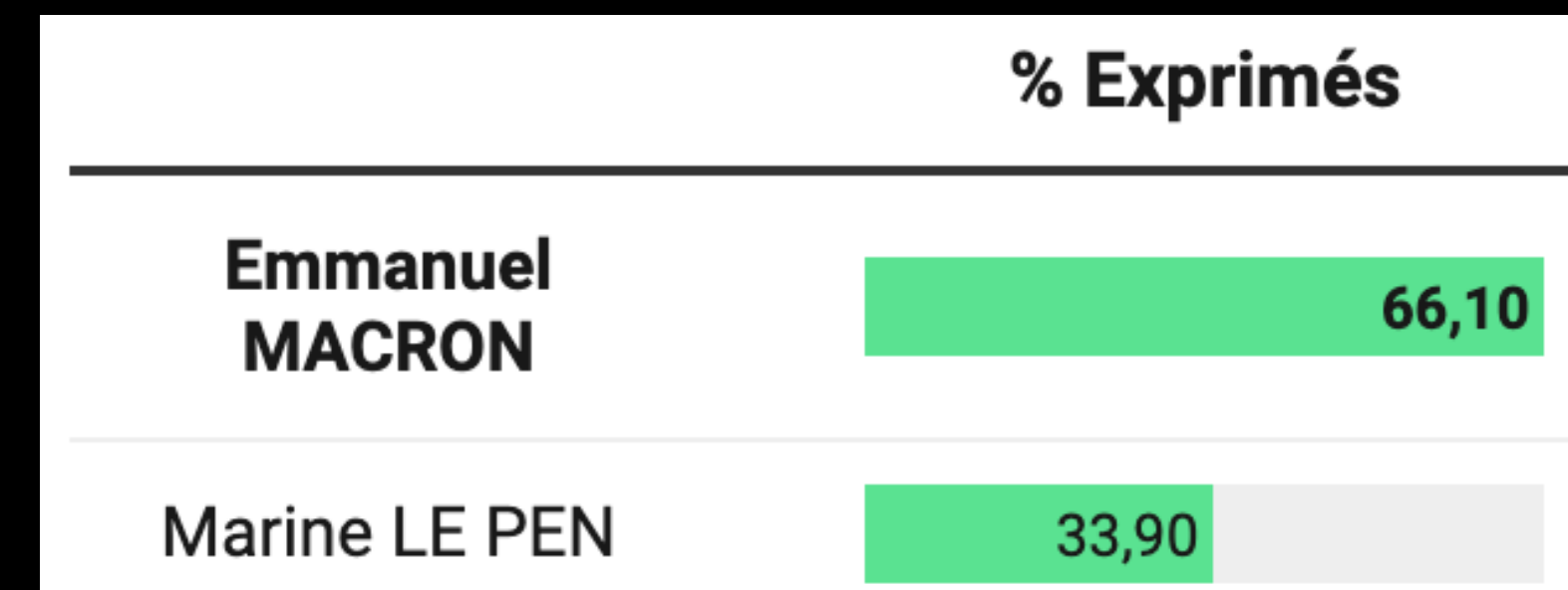
Full-Duplexity Sounds Sexy, But...

Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with 57.55% of the vote, defeating the National Front's Marine Le Pen.

Me

- Who is the current president of France?
- When was his second term election?



Full-Duplexity Sounds Sexy, But...

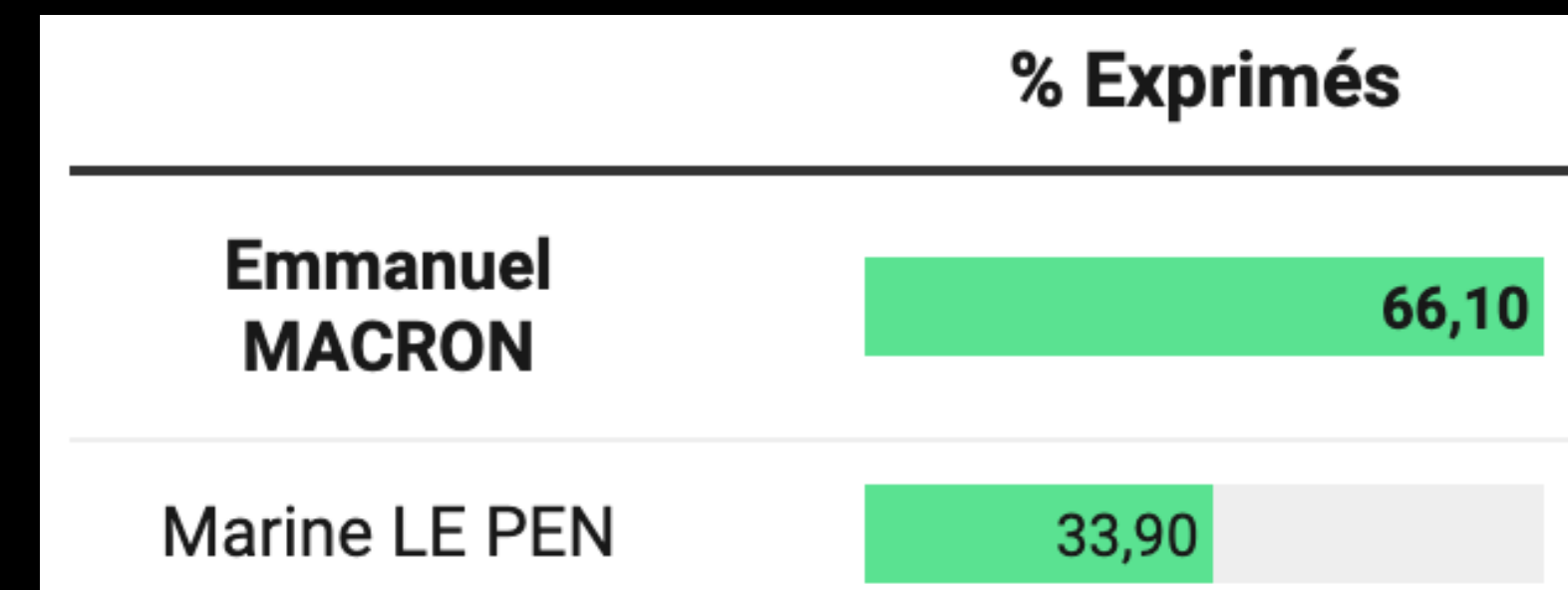
Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with 57.55% of the vote, defeating the National Front's Marine Le Pen.

Hey, how are you doing? Oh, I'm glad you asked. I'm

Me

- Who is the current president of France?
- When was his second term election?



Full-Duplexity Sounds Sexy, But...

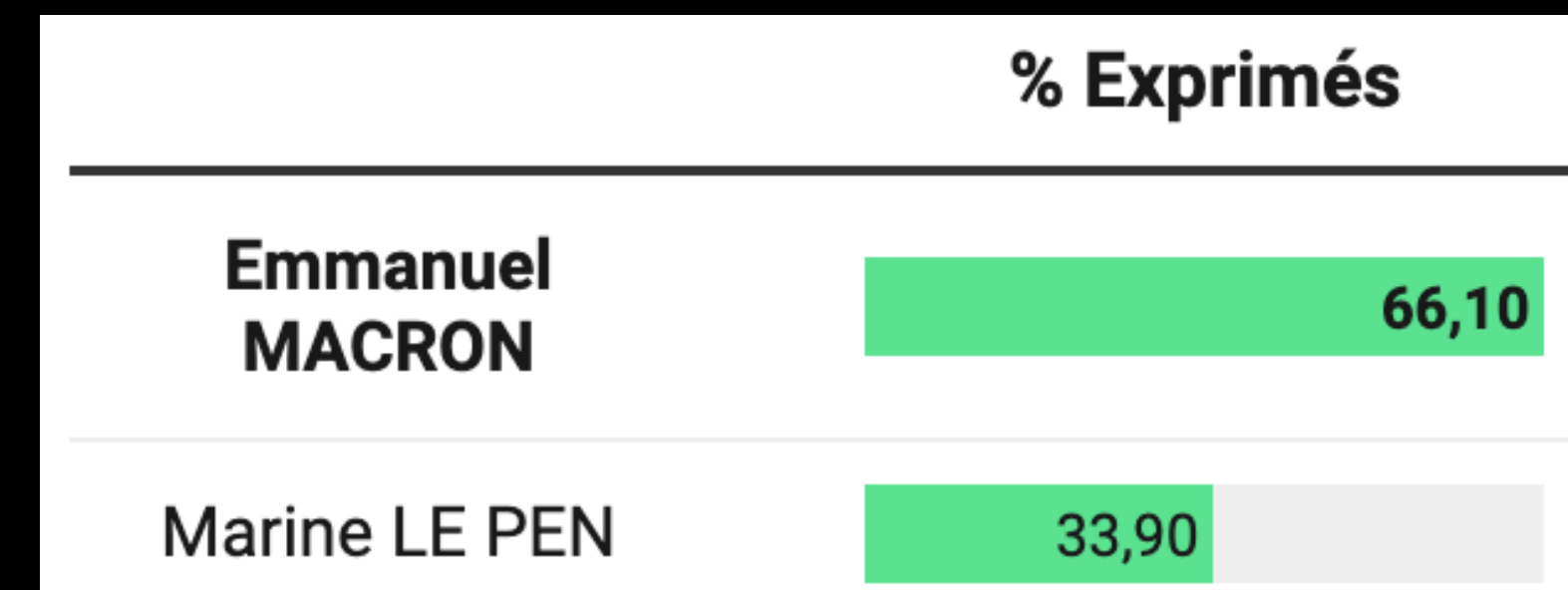
Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with **57.55%** of the vote, defeating the National Front's Marine Le Pen.

Hey, how are you doing? Oh, I'm glad you asked. I'm

Me

- Who is the current president of France?
- When was his second term election?



- Who are you?

Full-Duplexity Sounds Sexy, But...

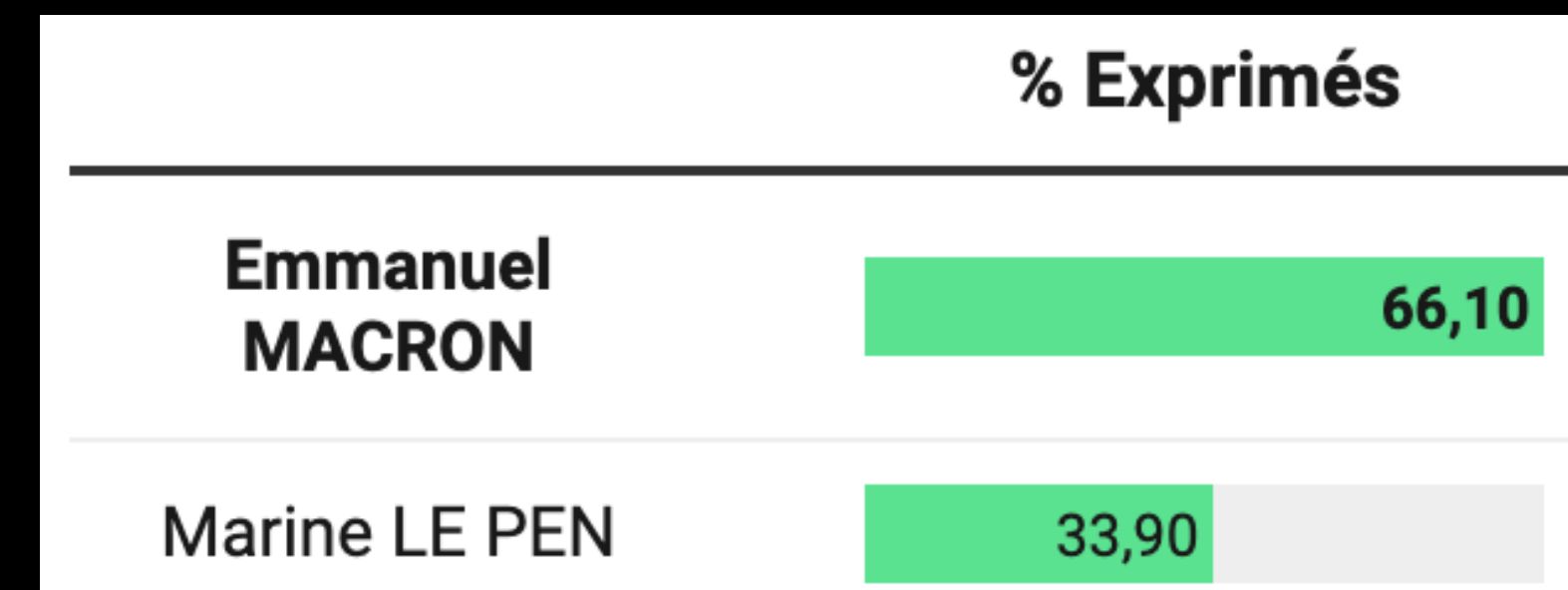
Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with **57.55%** of the vote, defeating the National Front's Marine Le Pen.

Hey, how are you doing? Oh, I'm glad you asked. I'm your host, Moshi. I' a cutting edge AI model developed by Kyutai, a Paris based non profit

Me

- Who is the current president of France?
- When was his second term election?



- Who are you?

Full-Duplexity Sounds Sexy, But...

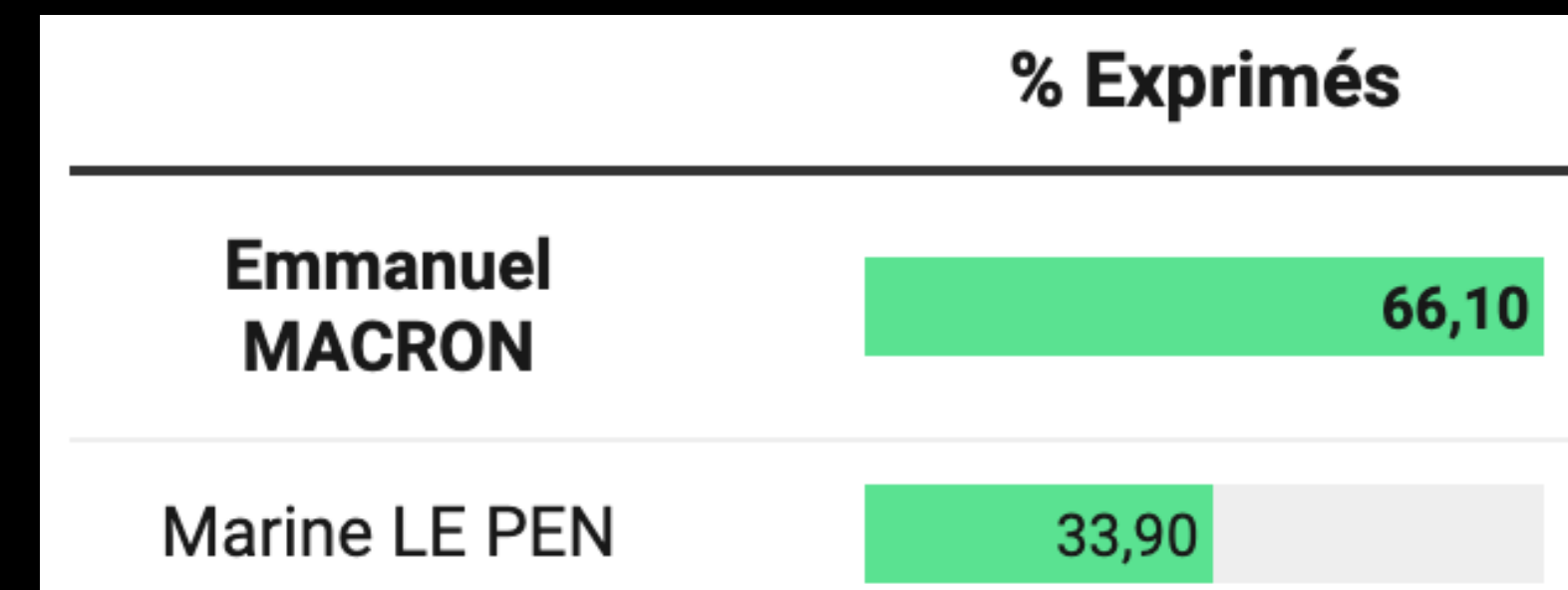
Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with **57.55%** of the vote, defeating the National Front's Marine Le Pen.

Hey, how are you doing? Oh, I'm glad you asked. I'm your host, Moshi. I' a cutting edge AI model developed by Kyutai, a Paris based non profit

Me

- Who is the current president of France?
- When was his second term election?



- Who are you?
- Has Kyutai been there for long?

Full-Duplexity Sounds Sexy, But...

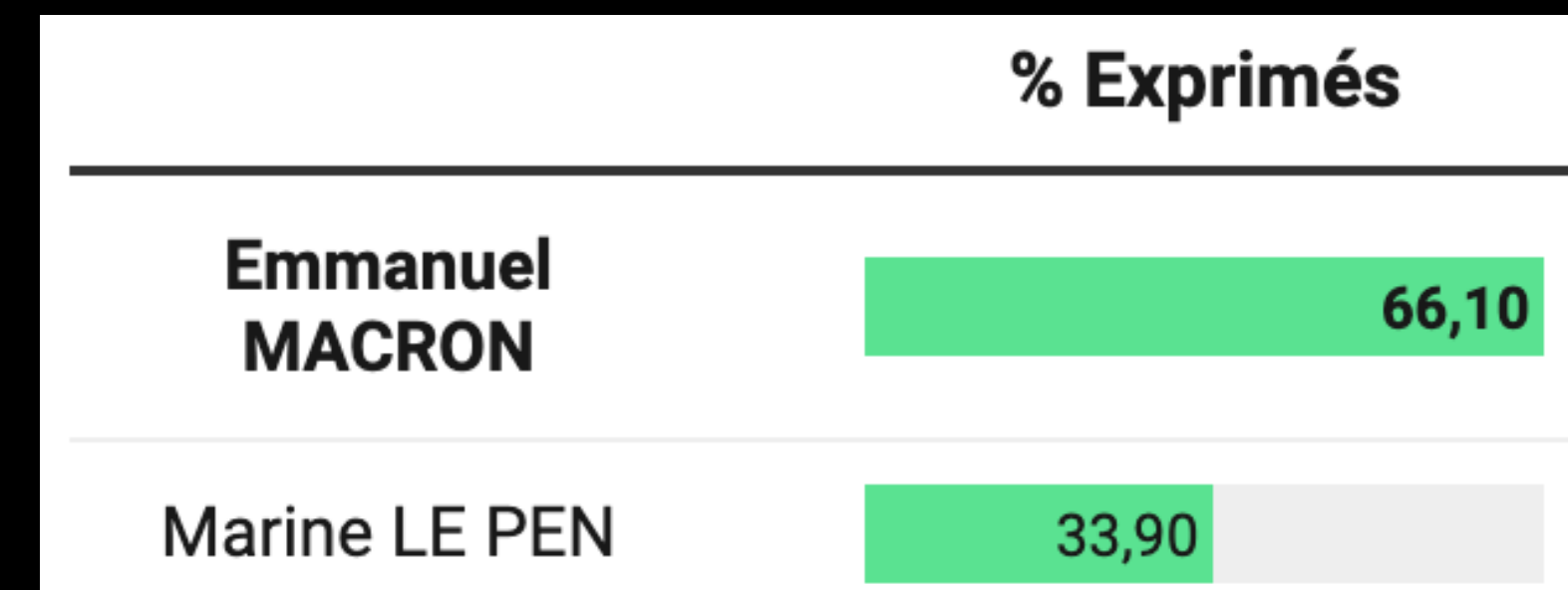
Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with **57.55%** of the vote, defeating the National Front's Marine Le Pen.

Hey, how are you doing? Oh, I'm glad you asked. I'm your host, Moshi. I' a cutting edge AI model developed by Kyutai, a Paris based non profit research lab. Well, Kyutai was founded in 2024, so it's still a relatively new organization, but it's already made quite a name for itself. Kyutai was

Me

- Who is the current president of France?
- When was his second term election?



- Who are you?
- Has Kyutai been there for long?

Full-Duplexity Sounds Sexy, But...

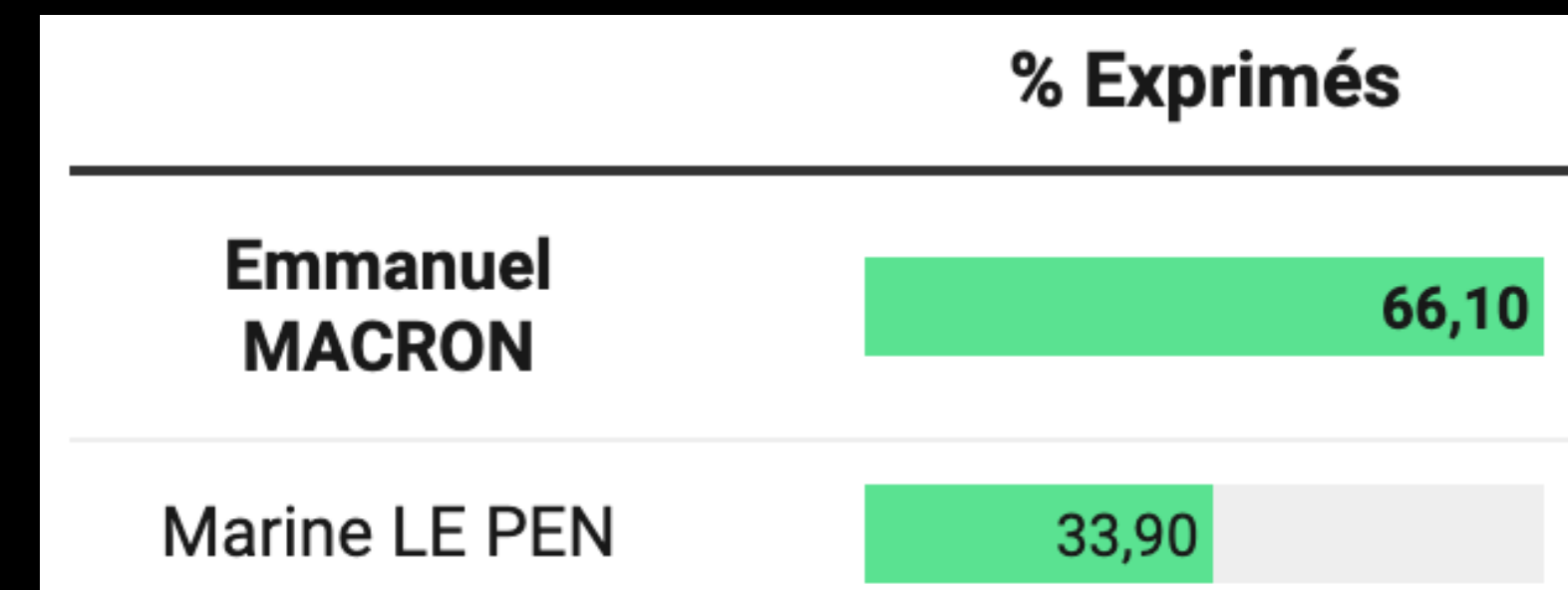
Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with **57.55%** of the vote, defeating the National Front's Marine Le Pen.

Hey, how are you doing? Oh, I'm glad you asked. I'm your host, Moshi. I' a cutting edge AI model developed by Kyutai, a Paris based non profit research lab. Well, Kyutai was founded in **2024**, so it's still a relatively new organization, but it's already made quite a name for itself. Kyutai was

Me

- Who is the current president of France?
- When was his second term election?



- Who are you?
- Has Kyutai been there for long?

Full-Duplexity Sounds Sexy, But...

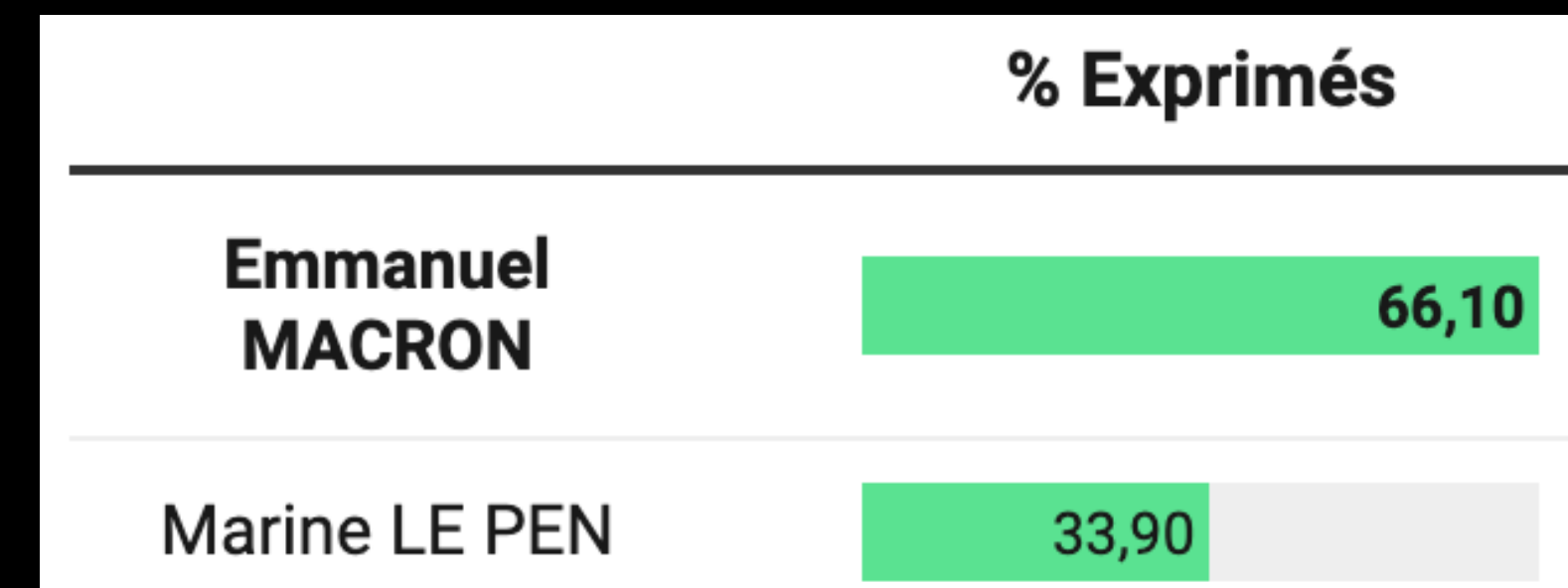
Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with **57.55%** of the vote, defeating the National Front's Marine Le Pen.

Hey, how are you doing? Oh, I'm glad you asked. I'm your host, Moshi. I' a cutting edge AI model developed by Kyutai, a Paris based non profit research lab. Well, Kyutai was founded in **2024**, so it's still a relatively new organization, but it's already made quite a name for itself. Kyutai was

Me

- Who is the current president of France?
- When was his second term election?



- Who are you?
- Has Kyutai been there for long?
- Who founded Kyutai?

Full-Duplexity Sounds Sexy, But...

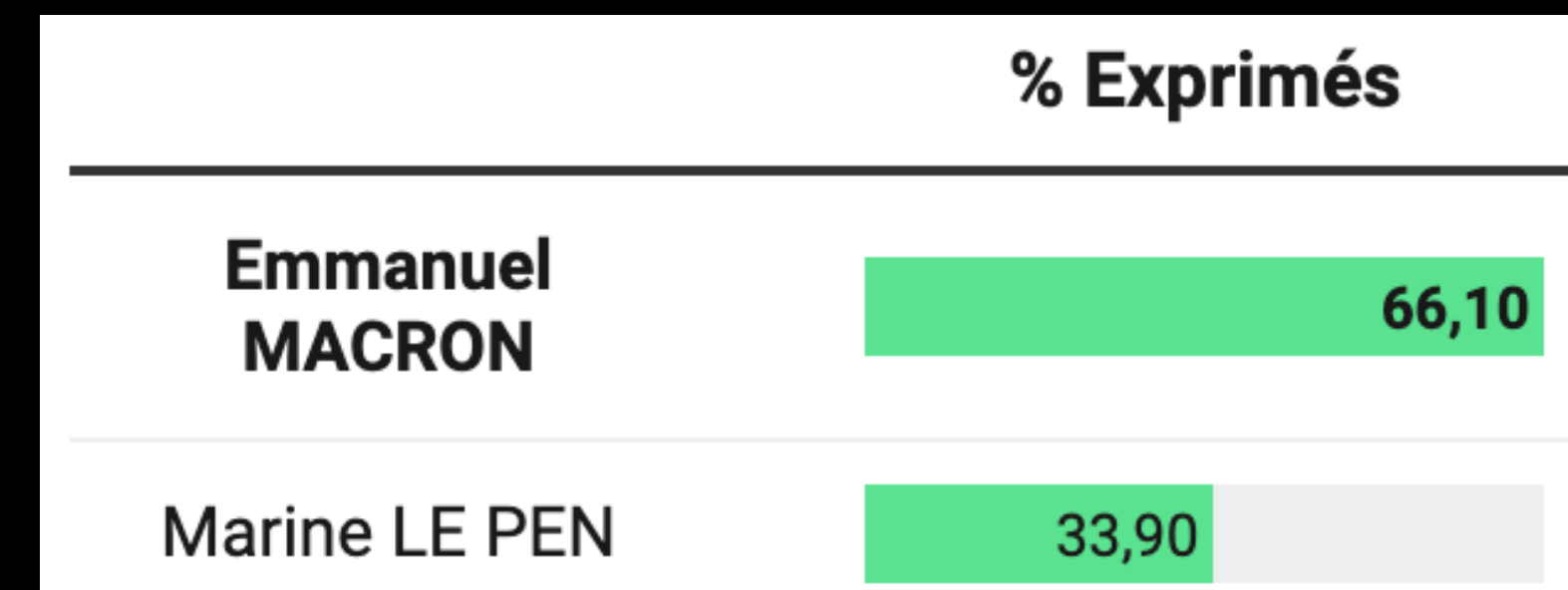
Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with **57.55%** of the vote, defeating the National Front's Marine Le Pen.

Hey, how are you doing? Oh, I'm glad you asked. I'm your host, Moshi. I' a cutting edge AI model developed by Kyutai, a Paris based non profit research lab. Well, Kyutai was founded in **2024**, so it's still a relatively new organization, but it's already made quite a name for itself. Kyutai was created by a team of leading AI specialist, including Yejin Choi, Yann LeCun, and Bernhard Schölkopf.

Me

- Who is the current president of France?
- When was his second term election?



- Who are you?
- Has Kyutai been there for long?
- Who founded Kyutai?

Full-Duplexity Sounds Sexy, But...

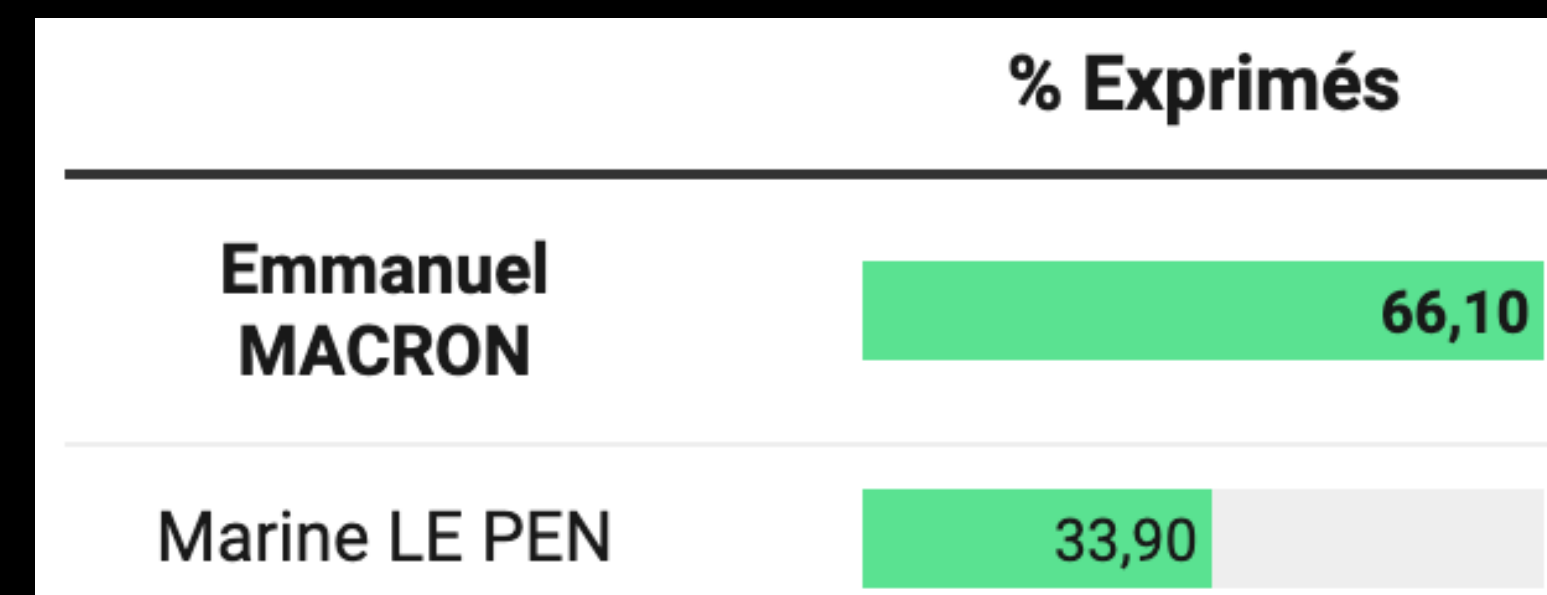
Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with 57.55% of the vote, defeating the National Front's Marine Le Pen.

Hey, how are you doing? Oh, I'm glad you asked. I'm your host, Moshi. I' a cutting edge AI model developed by Kyutai, a Paris based non profit research lab. Well, Kyutai was founded in 2024, so it's still a relatively new organization, but it's already made quite a name for itself. Kyutai was created by a team of leading AI specialist, including Yejin Choi, Yann LeCun, and Bernhard Schölkopf.

Me

- Who is the current president of France?
- When was his second term election?



- Who are you?
- Has Kyutai been there for long?
- Who founded Kyutai?

Full-Duplexity Sounds Sexy, But...

Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with 57.55% of the vote, defeating the National Front's Marine Le Pen.

Hey, how are you your host, Moshi developed by Kyu research lab. We it's still a rel already made qui created by a team Yejin Choi, Yann



Patrick Pérez
Chief Executive Officer
GScholar



Edouard Grave
Chief Language Officer
GScholar



Laurent Mazaré
Chief Technical Officer
GitHub



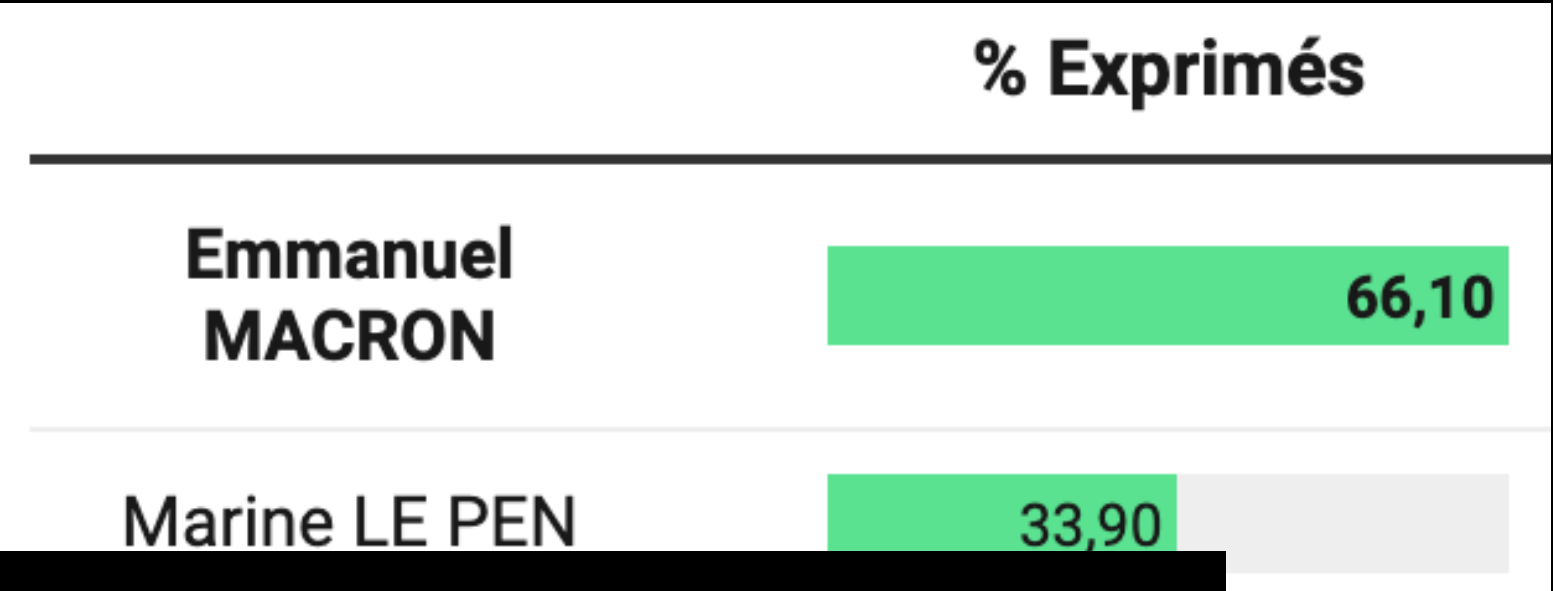
Neil Zeghidour
Chief Modelling Officer
GScholar



Alexandre Défossez
Chief Exploration Officer
GScholar

Me

- Who is the current president of France?
- When was his second term election?



ng?



Full-Duplexity Sounds Sexy, But...

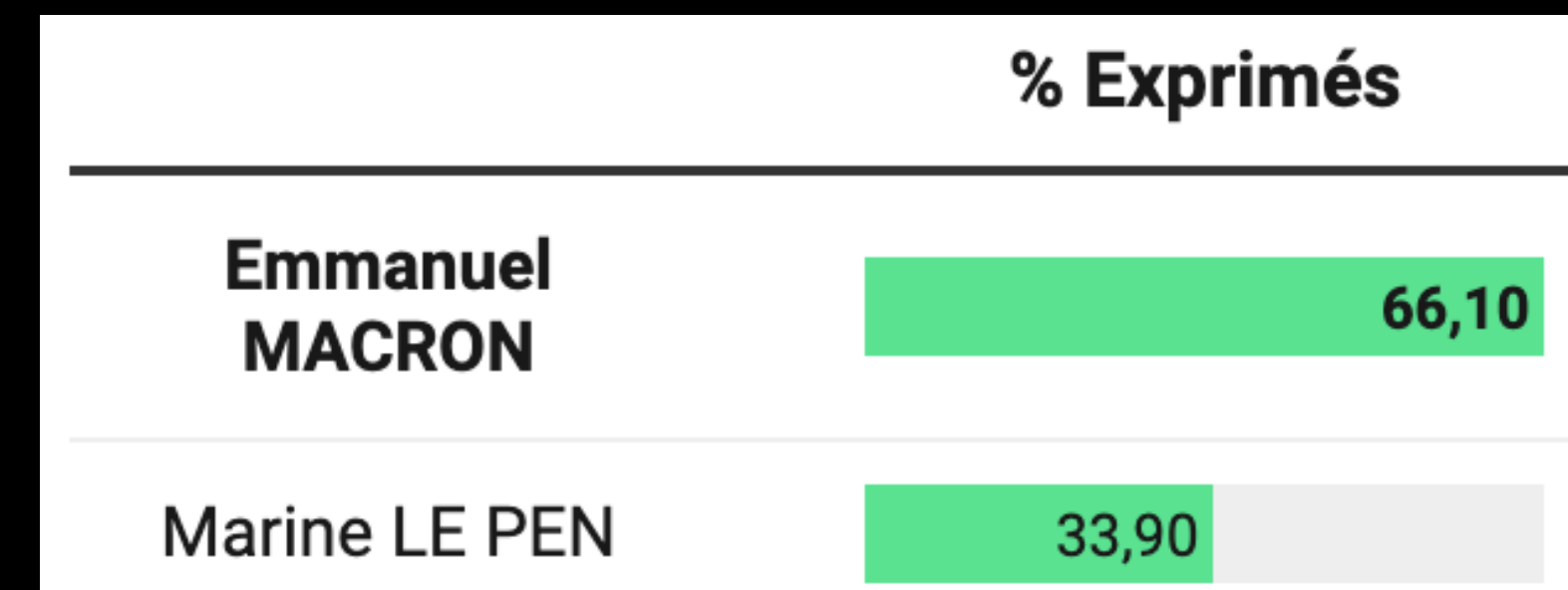
Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with 57.55% of the vote, defeating the National Front's Marine Le Pen.

Hey, how are you doing? Oh, I'm glad you asked. I'm your host, Moshi. I' a cutting edge AI model developed by Kyutai, a Paris based non profit research lab. Well, Kyutai was founded in 2024, so it's still a relatively new organization, but it's already made quite a name for itself. Kyutai was created by a team of leading AI specialist, including Yejin Choi, Yann LeCun, and Bernhard Schölkopf.

Me

- Who is the current president of France?
- When was his second term election?



- Who are you?
- Has Kyutai been there for long?
- Who founded Kyutai?

Full-Duplexity Sounds Sexy, But...

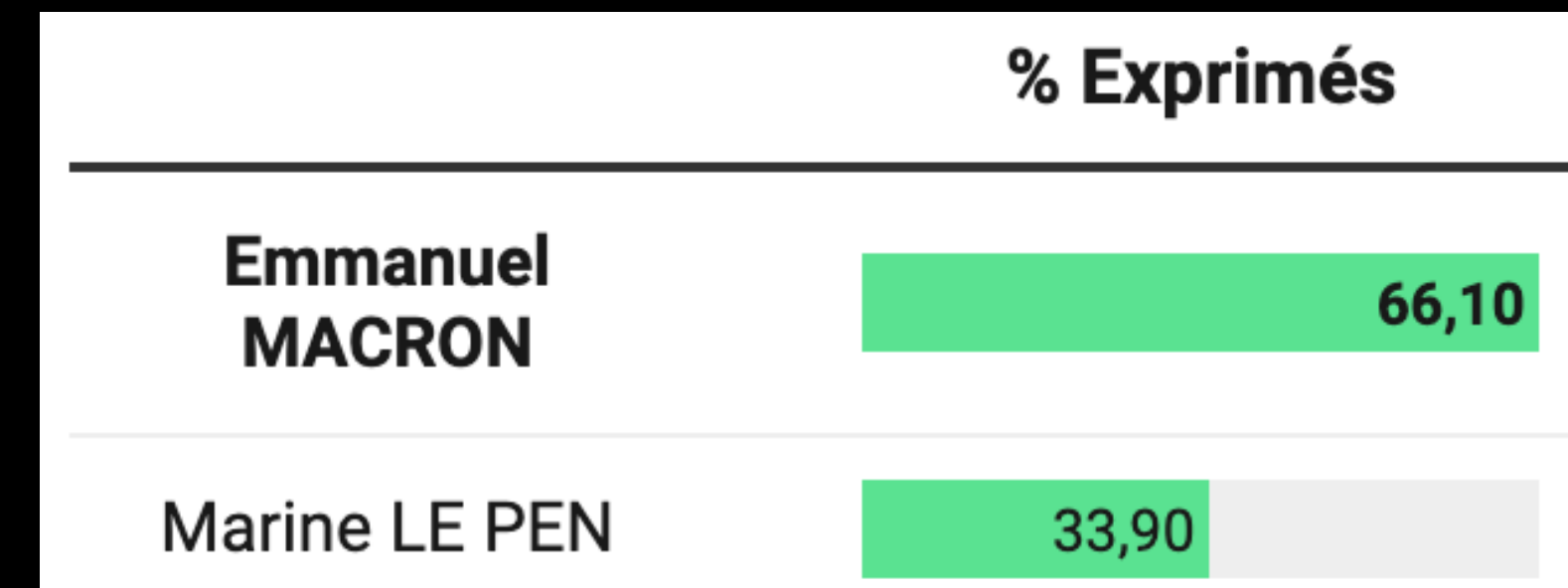
Moshi

Hey, how are you doing? I' doing well. How about yourself? Sure, the current president of France is Emmanuel Macron. He was elected in 2017 and has been serving as the second term president. The second round of the 2017 presidential election took place on May 7th. Emmanuel Macron won the election with **57.55%** of the vote, defeating the National Front's Marine Le Pen.

Hey, how are you doing? Thank you for asking. I was created by Kyutai, a non profit research lab based in Paris. Kyutai was founded in 2021 by the Iliad Group, the CMA CGM Group, and Schmidt Futures. The founders of Kyutai were Xavier Niel, Rodolphe Saadé, and Eric Schmidt.

Me

- Who is the current president of France?
- When was his second term election?



- Who are you?
- Has Kyutai been there for long?
- Who founded Kyutai?

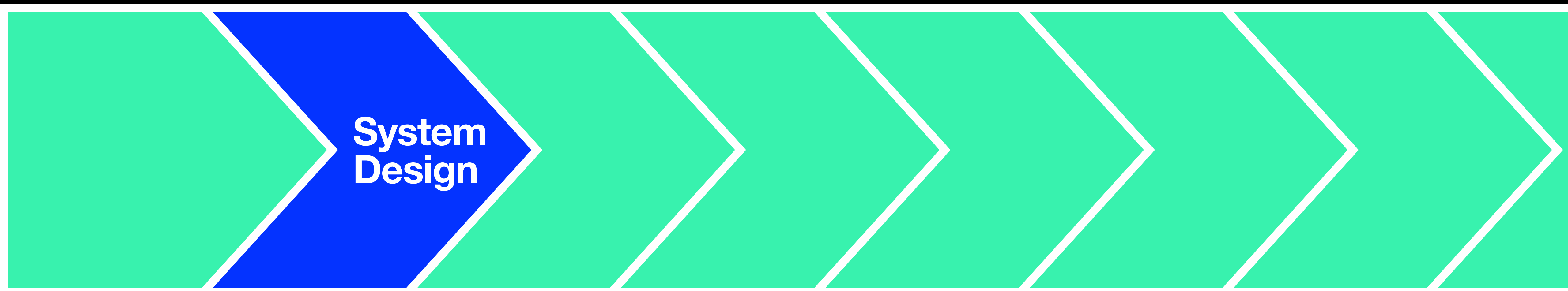
Full-Duplex Systems Incorporating Background Models

Background models for search, tool-use, and more

Full-Duplex Systems Incorporating Background Models

Background models for search, tool-use, and more





**System
Design**

Goal

Goal

- Make Moshi be aware of its limitation: **accuracy** & consistency

Goal

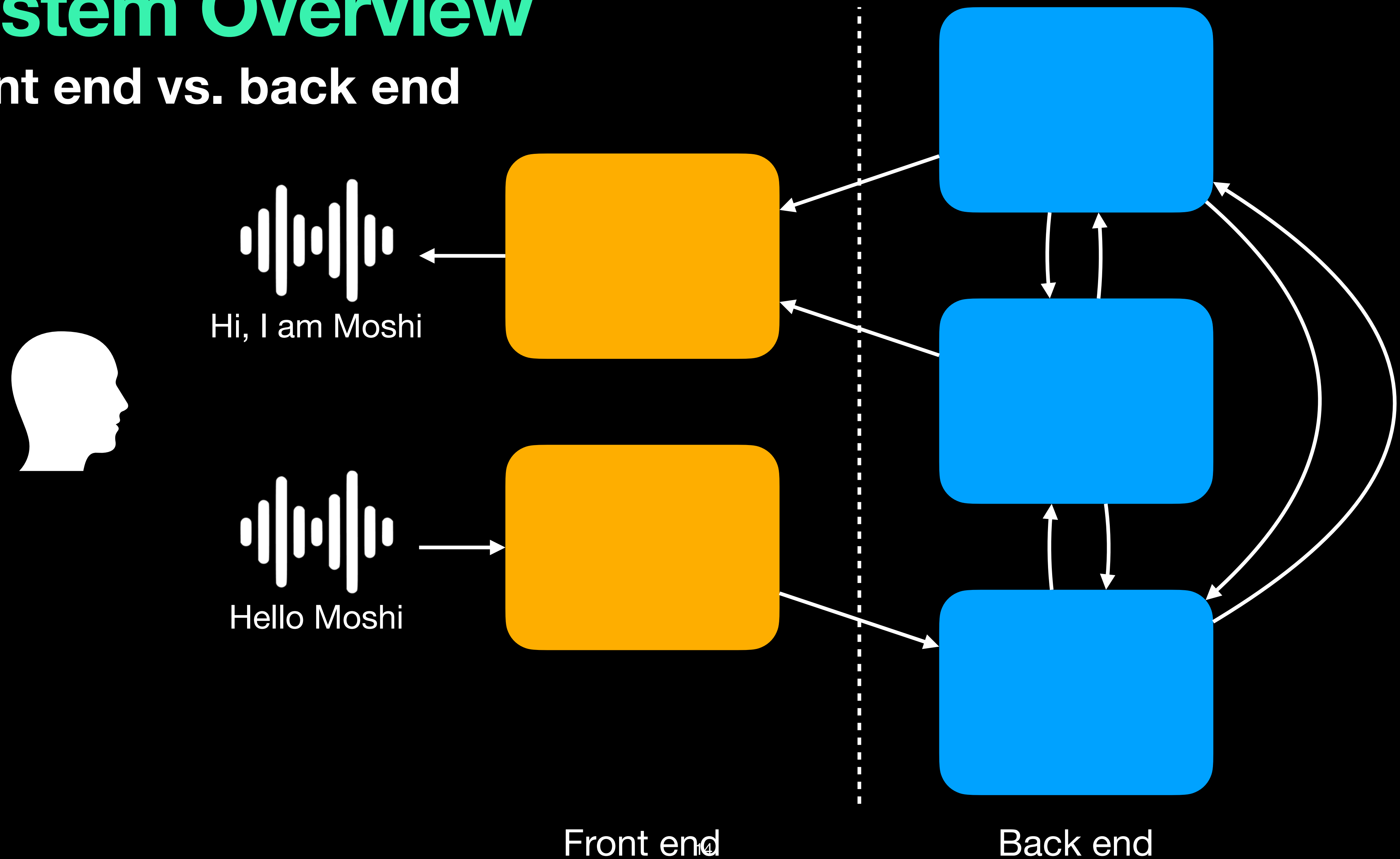
- Make Moshi be aware of its limitation: **accuracy** & consistency
- Teach Moshi to improve its responses with retrieved information obtained from **asynchronous RAG**

Goal

- Make Moshi be aware of its limitation: **accuracy** & consistency
- Teach Moshi to improve its responses with retrieved information obtained from **asynchronous RAG**
- Keep the **interactivity** of original moshi
→ timing is a critical concern

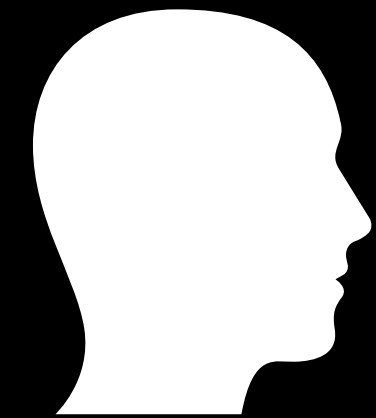
System Overview

Front end vs. back end



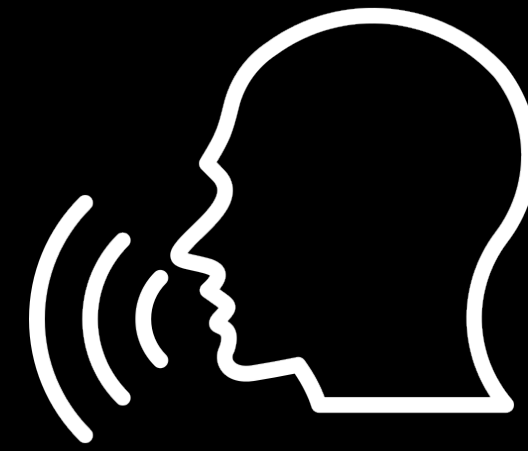
System Overview

Front end vs. back end

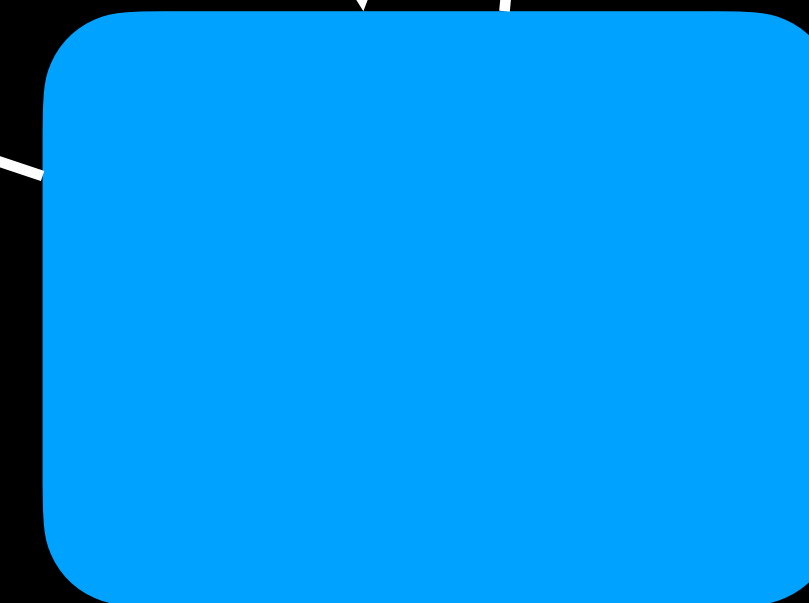



Hi, I am Moshi


Hello Moshi



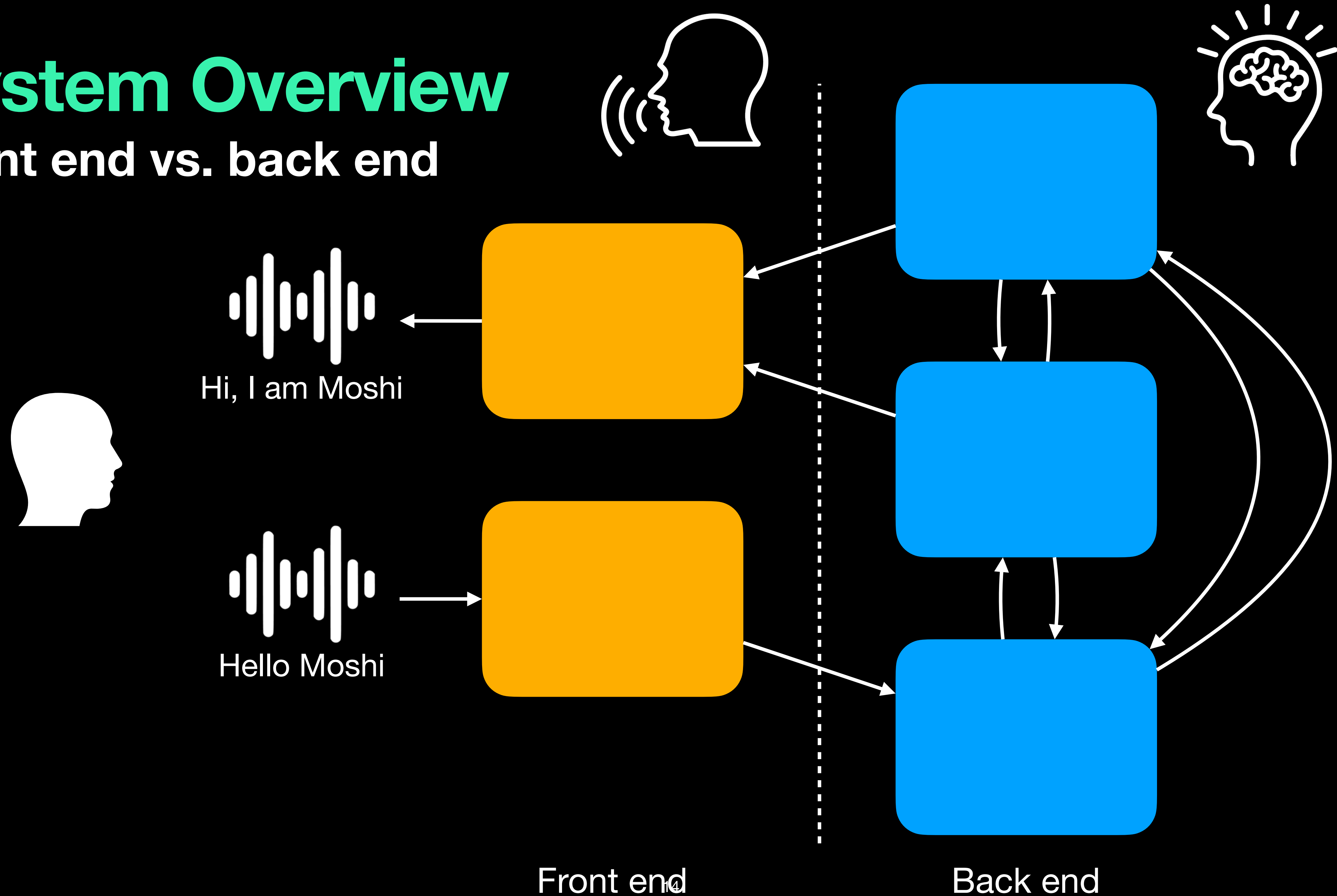
Front end



Back end

System Overview

Front end vs. back end



System Design

Timing control between front end and back end

System Design

Timing control between front end and back end

- Front-end moshi: directly interacts with user

System Design

Timing control between front end and back end

- Front-end moshi: directly interacts with user
 - Real-time

System Design

Timing control between front end and back end

- Front-end moshi: directly interacts with user
 - Real-time
 - Smoothness

System Design

Timing control between front end and back end

- Front-end moshi: directly interacts with user
 - Real-time
 - Smoothness
- Back-end retrieval

System Design

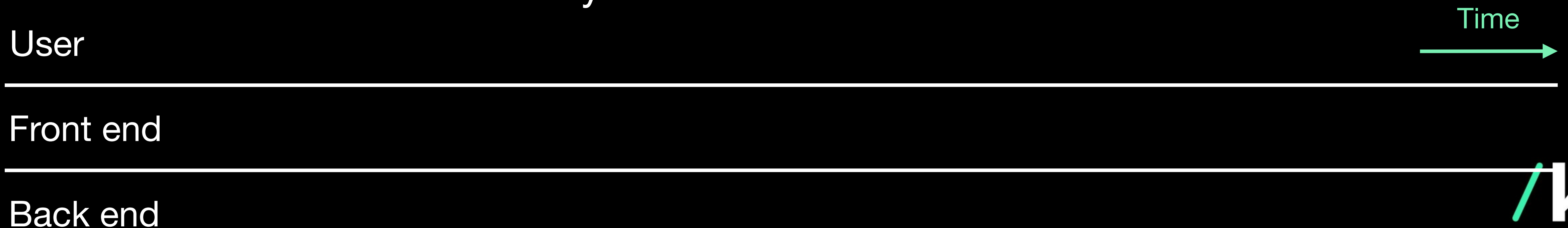
Timing control between front end and back end

- Front-end moshi: directly interacts with user
 - Real-time
 - Smoothness
- Back-end retrieval
 - Correctness and accuracy

System Design

Timing control between front end and back end

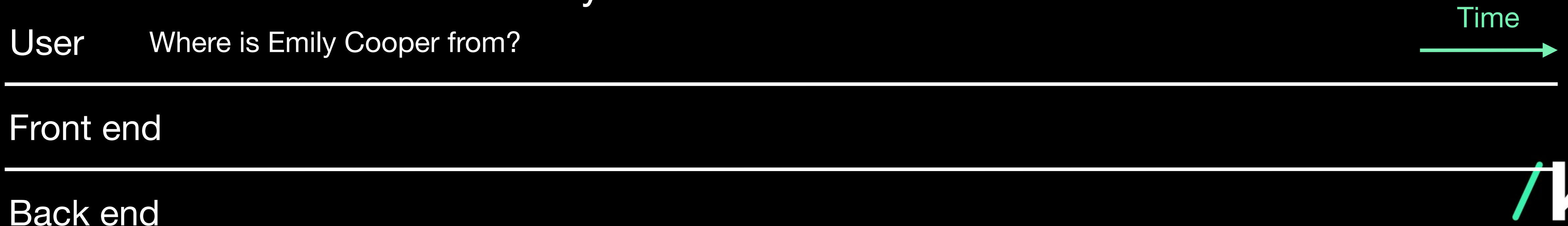
- Front-end moshi: directly interacts with user
 - Real-time
 - Smoothness
- Back-end retrieval
 - Correctness and accuracy



System Design

Timing control between front end and back end

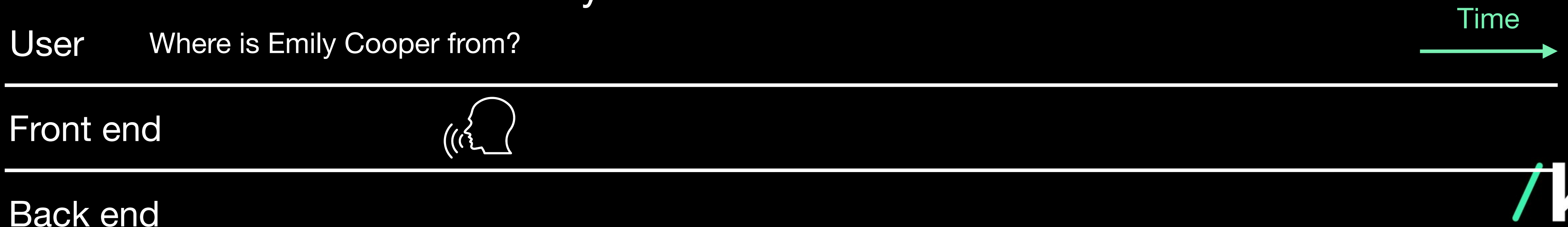
- Front-end moshi: directly interacts with user
 - Real-time
 - Smoothness
- Back-end retrieval
 - Correctness and accuracy



System Design

Timing control between front end and back end

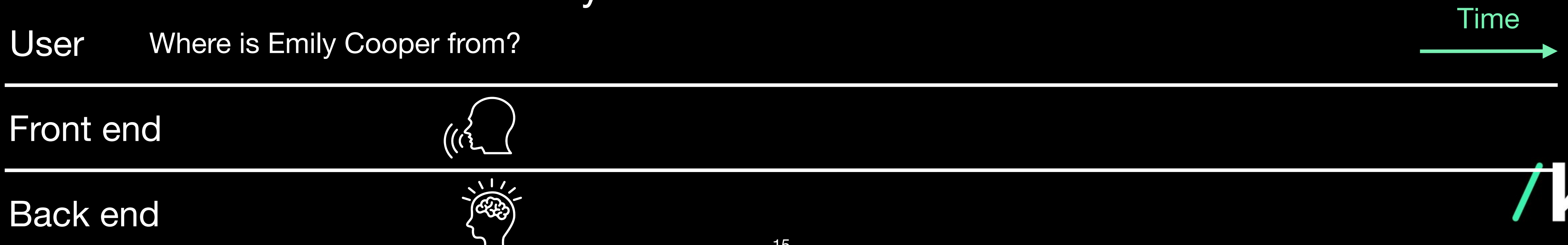
- Front-end moshi: directly interacts with user
 - Real-time
 - Smoothness
- Back-end retrieval
 - Correctness and accuracy



System Design

Timing control between front end and back end

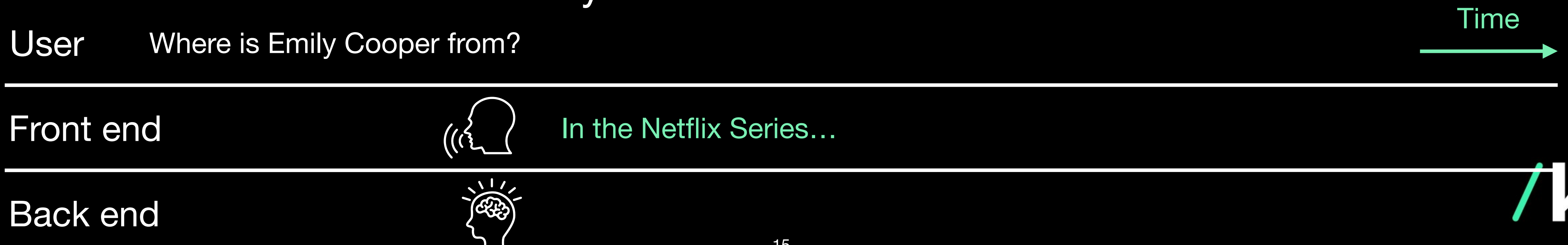
- Front-end moshi: directly interacts with user
 - Real-time
 - Smoothness
- Back-end retrieval
 - Correctness and accuracy



System Design

Timing control between front end and back end

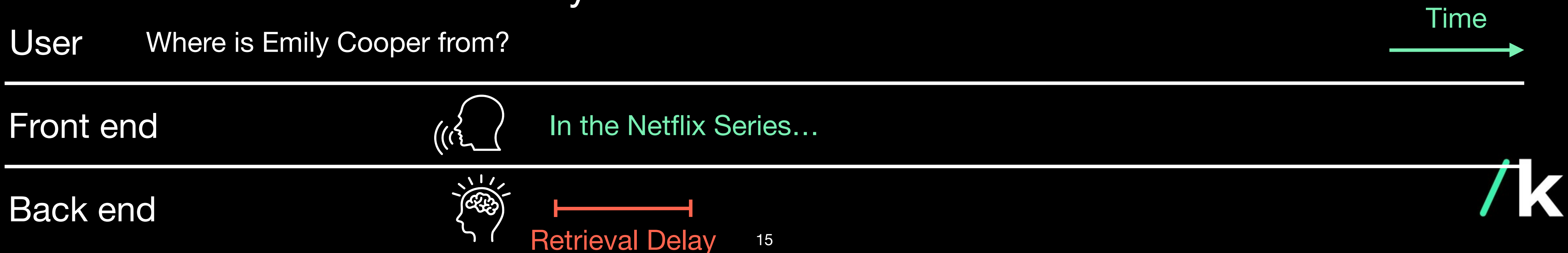
- Front-end moshi: directly interacts with user
 - Real-time
 - Smoothness
- Back-end retrieval
 - Correctness and accuracy



System Design

Timing control between front end and back end

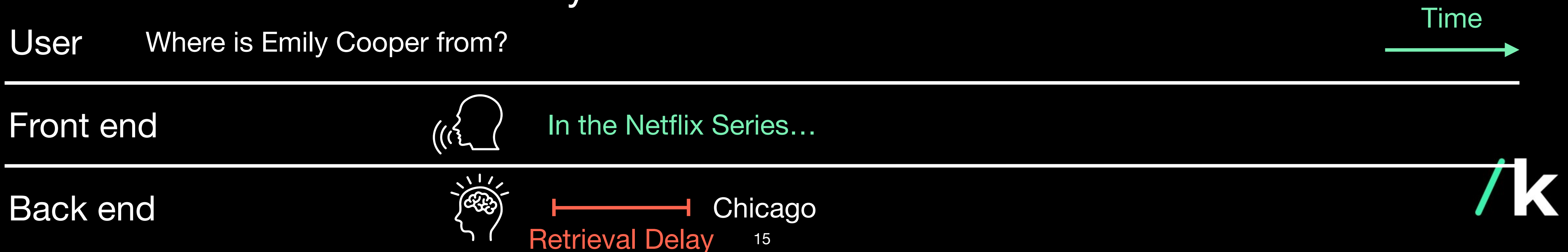
- Front-end moshi: directly interacts with user
 - Real-time
 - Smoothness
- Back-end retrieval
 - Correctness and accuracy



System Design

Timing control between front end and back end

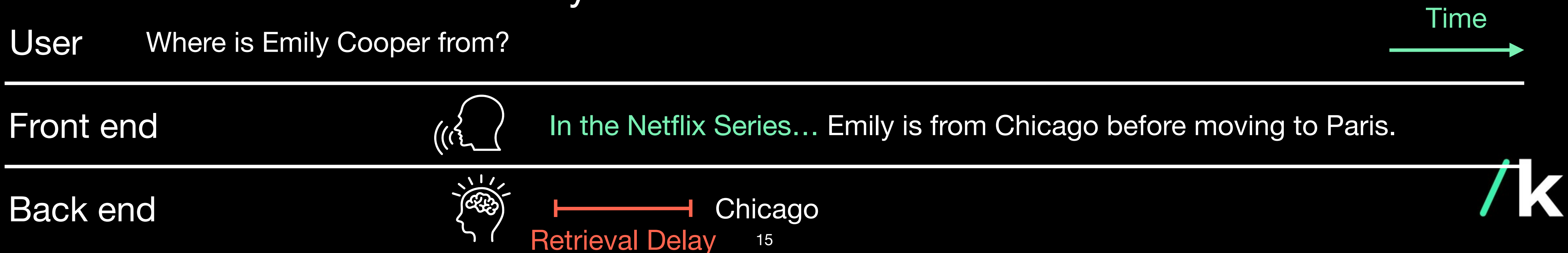
- Front-end moshi: directly interacts with user
 - Real-time
 - Smoothness
- Back-end retrieval
 - Correctness and accuracy



System Design

Timing control between front end and back end

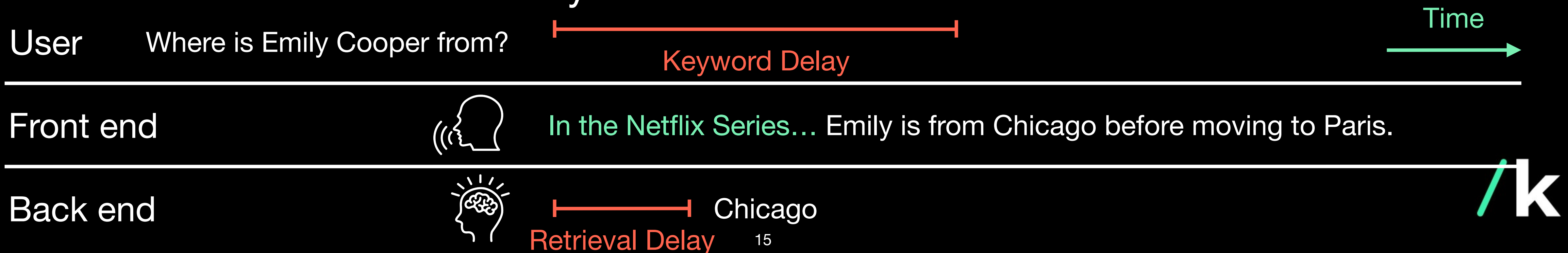
- Front-end moshi: directly interacts with user
 - Real-time
 - Smoothness
- Back-end retrieval
 - Correctness and accuracy



System Design

Timing control between front end and back end

- Front-end moshi: directly interacts with user
 - Real-time
 - Smoothness
- Back-end retrieval
 - Correctness and accuracy



System Design

Timing control between front end and back end

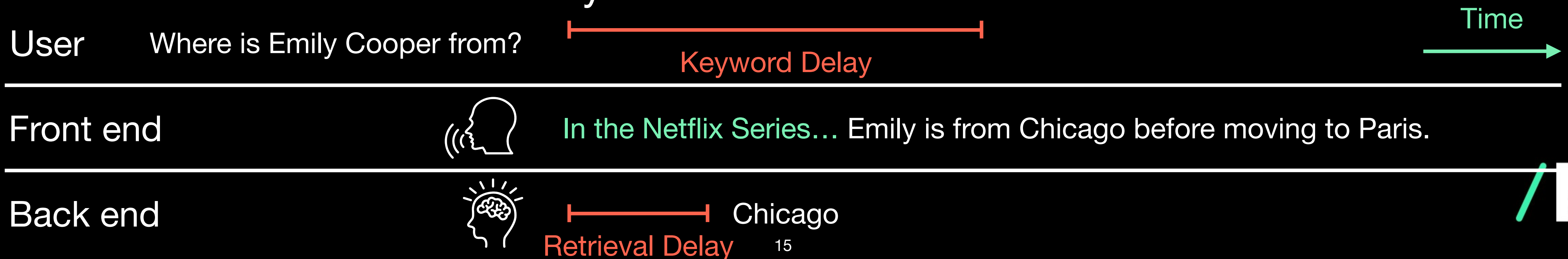
- Front-end moshi: directly interacts with user

- Real-time
- Smoothness

$$\text{Retrieval Delay} \leq \text{Keyword Delay}$$

- Back-end retrieval

- Correctness and accuracy



Streaming Front-End

Streaming Front-End

Front-end

Moshi

Streaming Front-End

Front-end

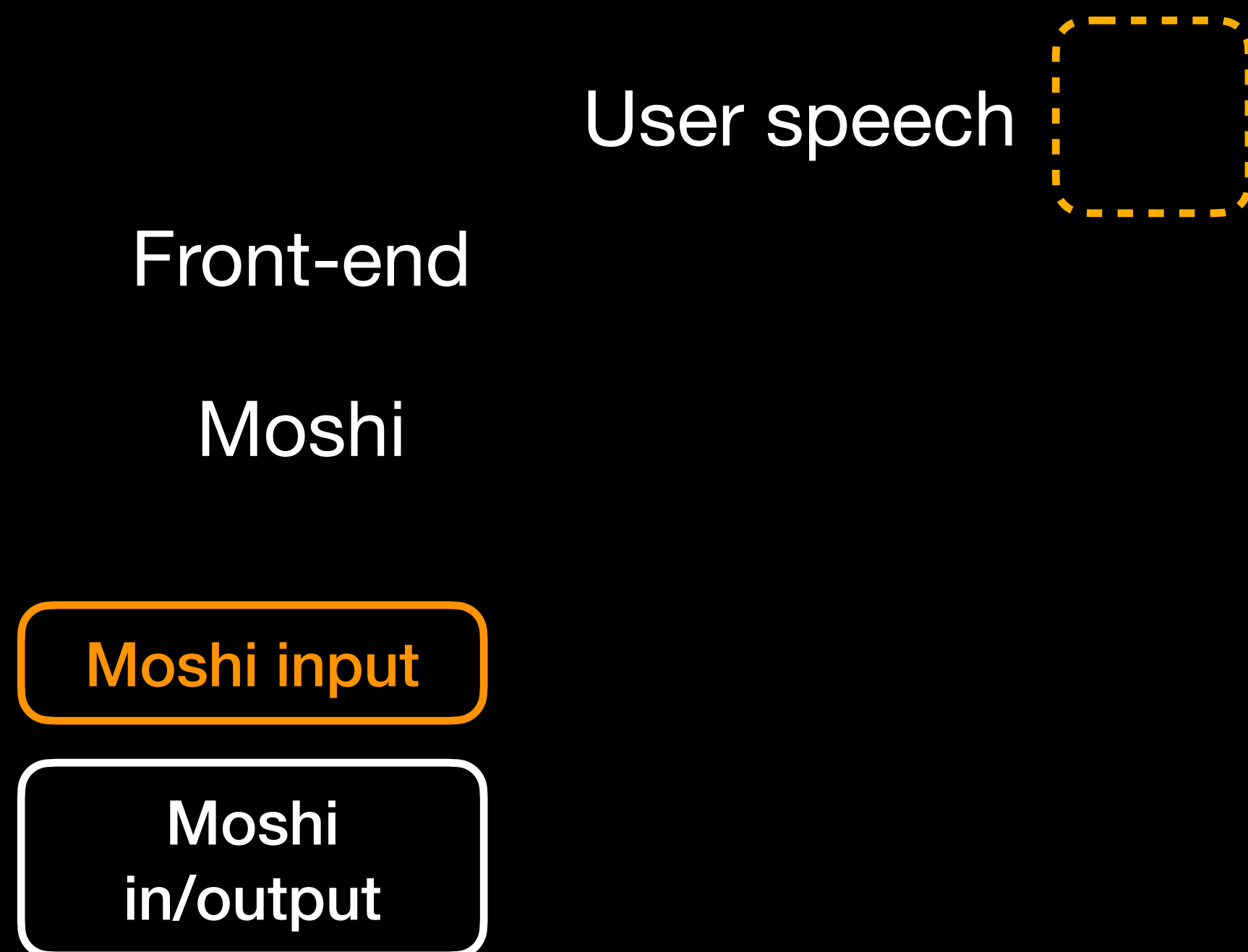
Moshi

Moshi input

Moshi
in/output

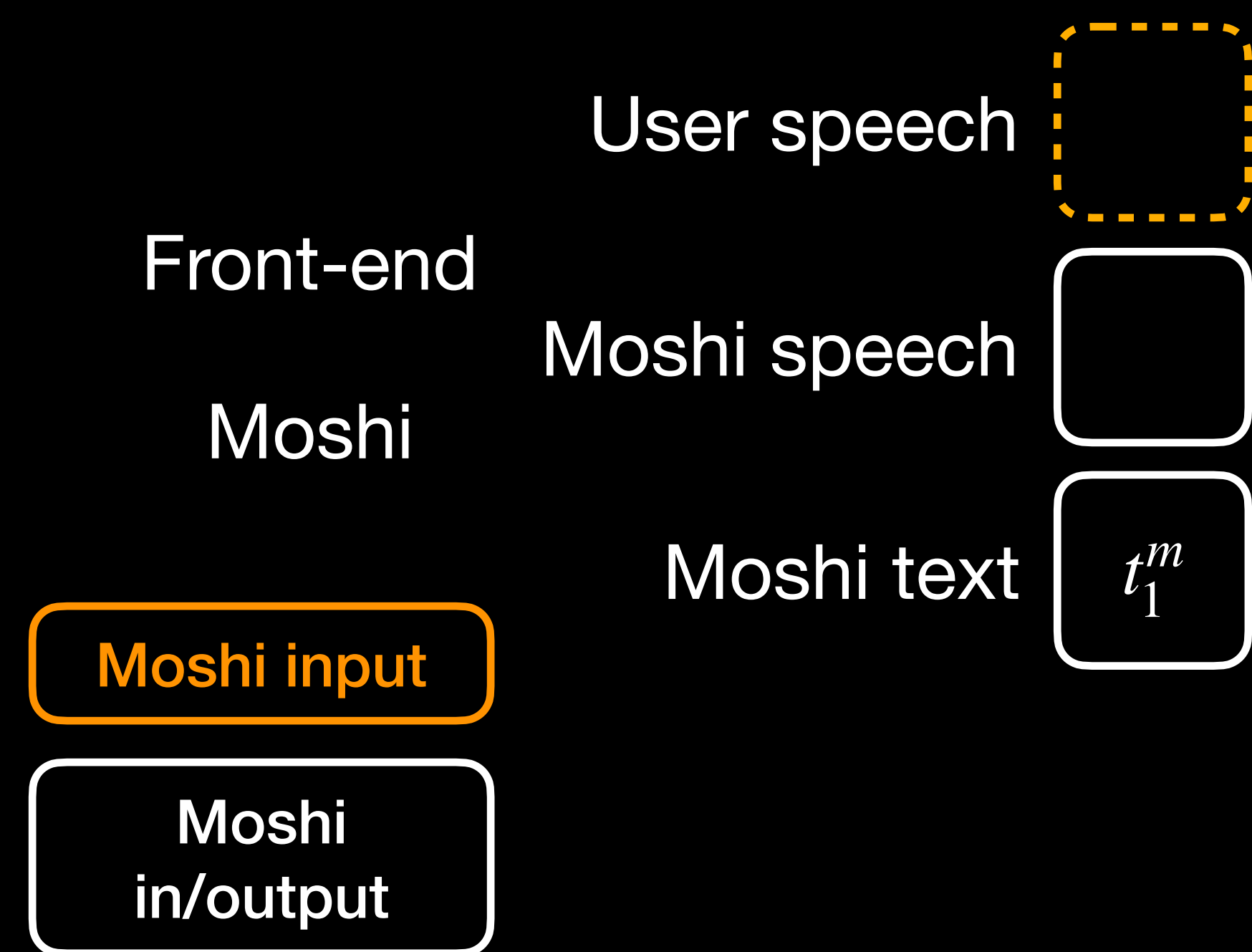
- Streaming operations
- Asynchronous operations

Streaming Front-End



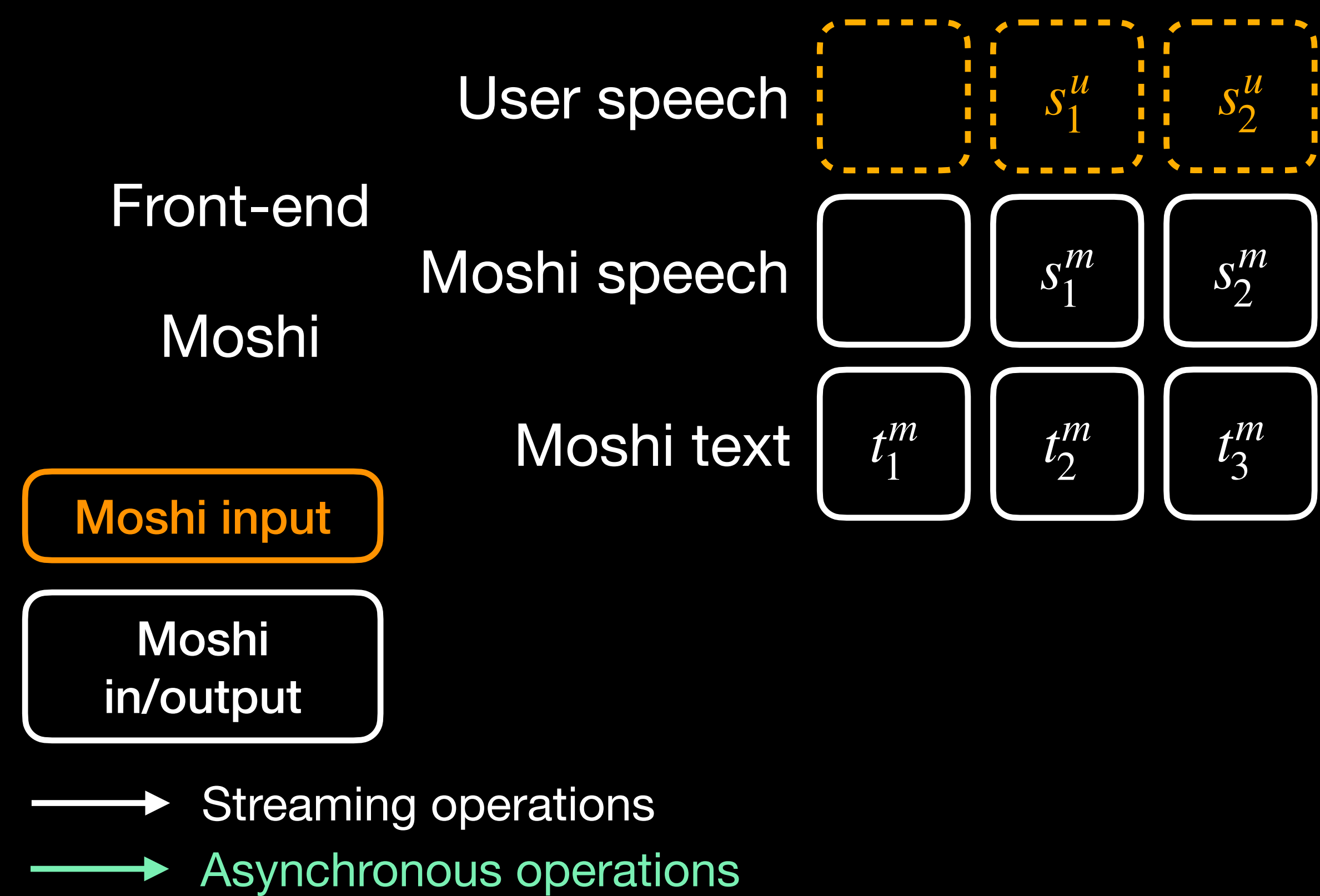
- Streaming operations
- Asynchronous operations

Streaming Front-End

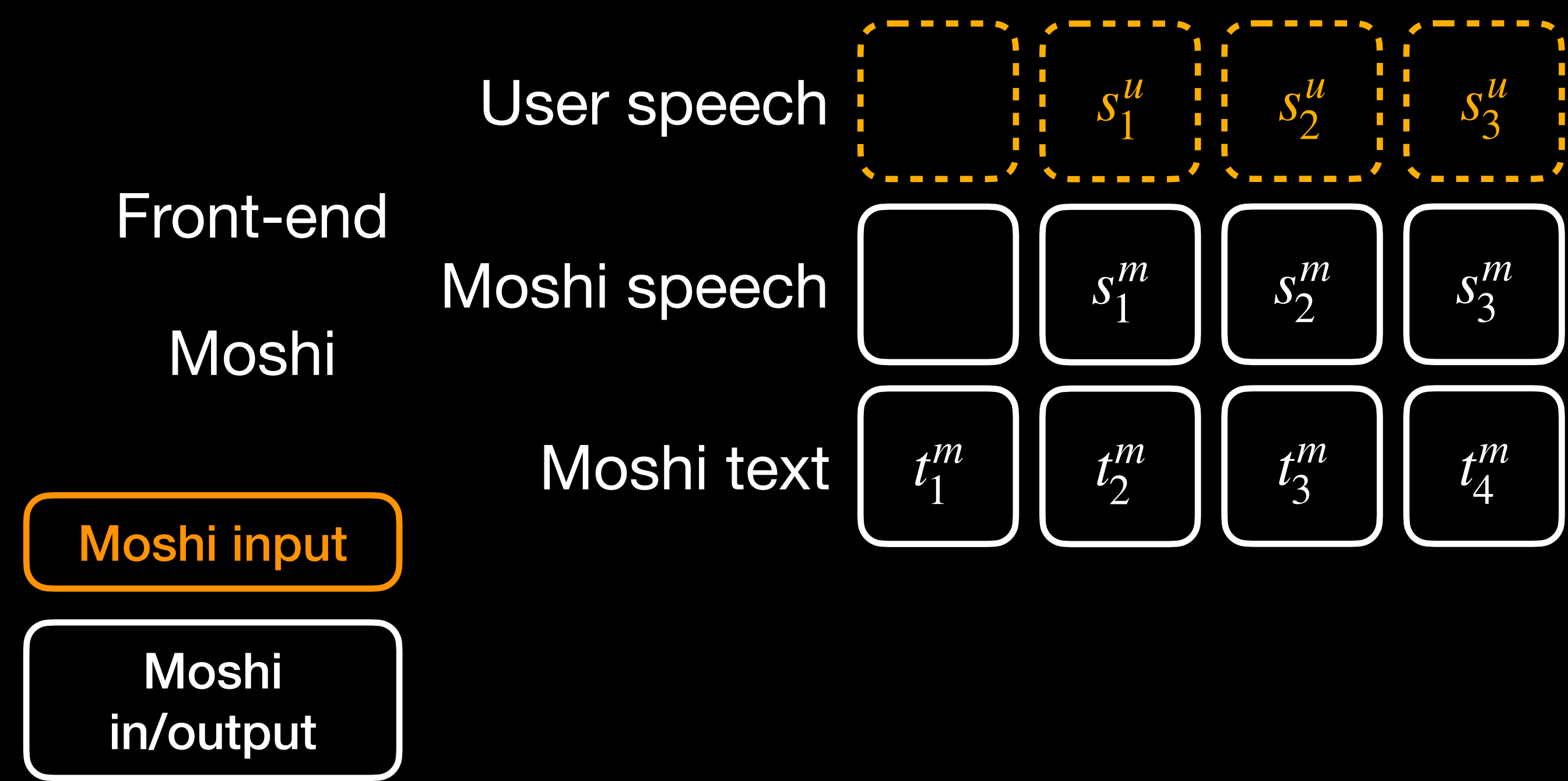


- Streaming operations
- Asynchronous operations

Streaming Front-End



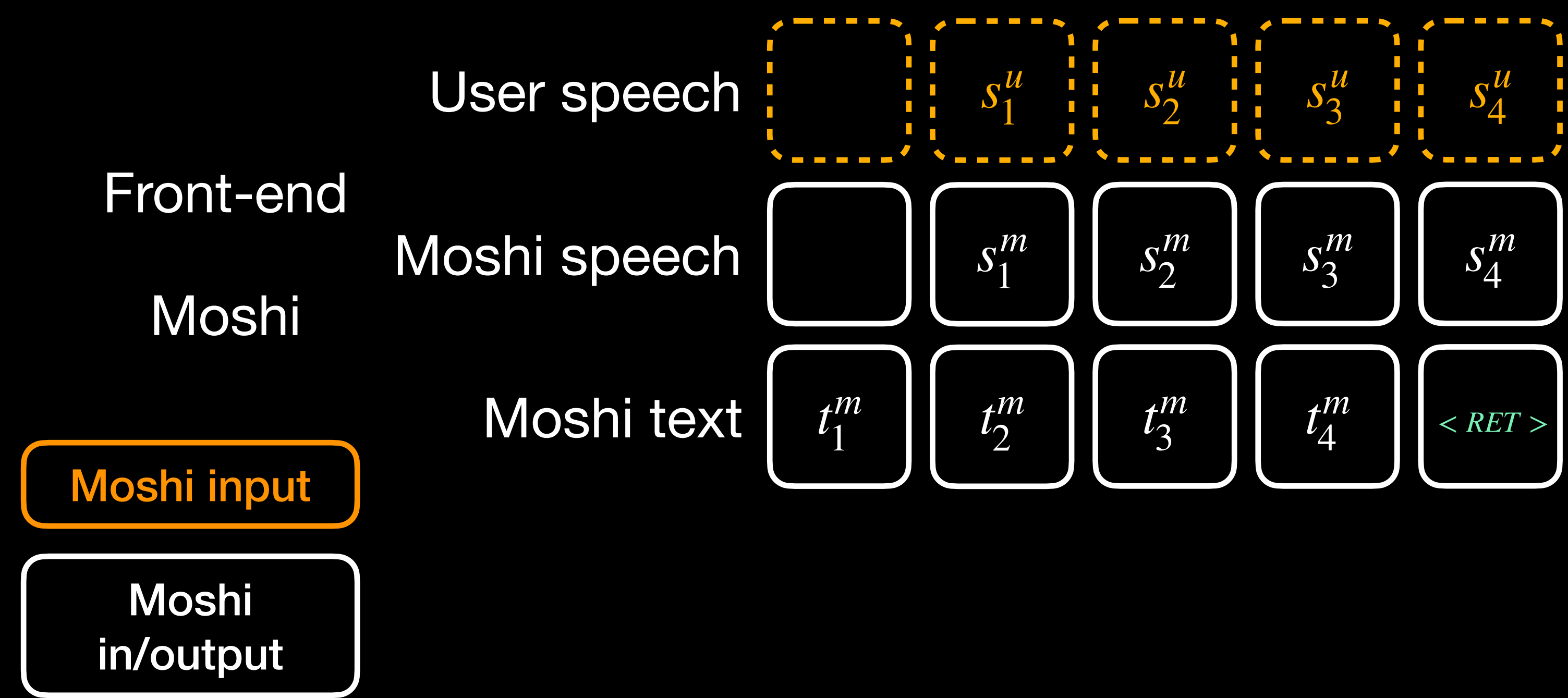
Streaming Front-End



→ Streaming operations

→ Asynchronous operations

Streaming Front-End



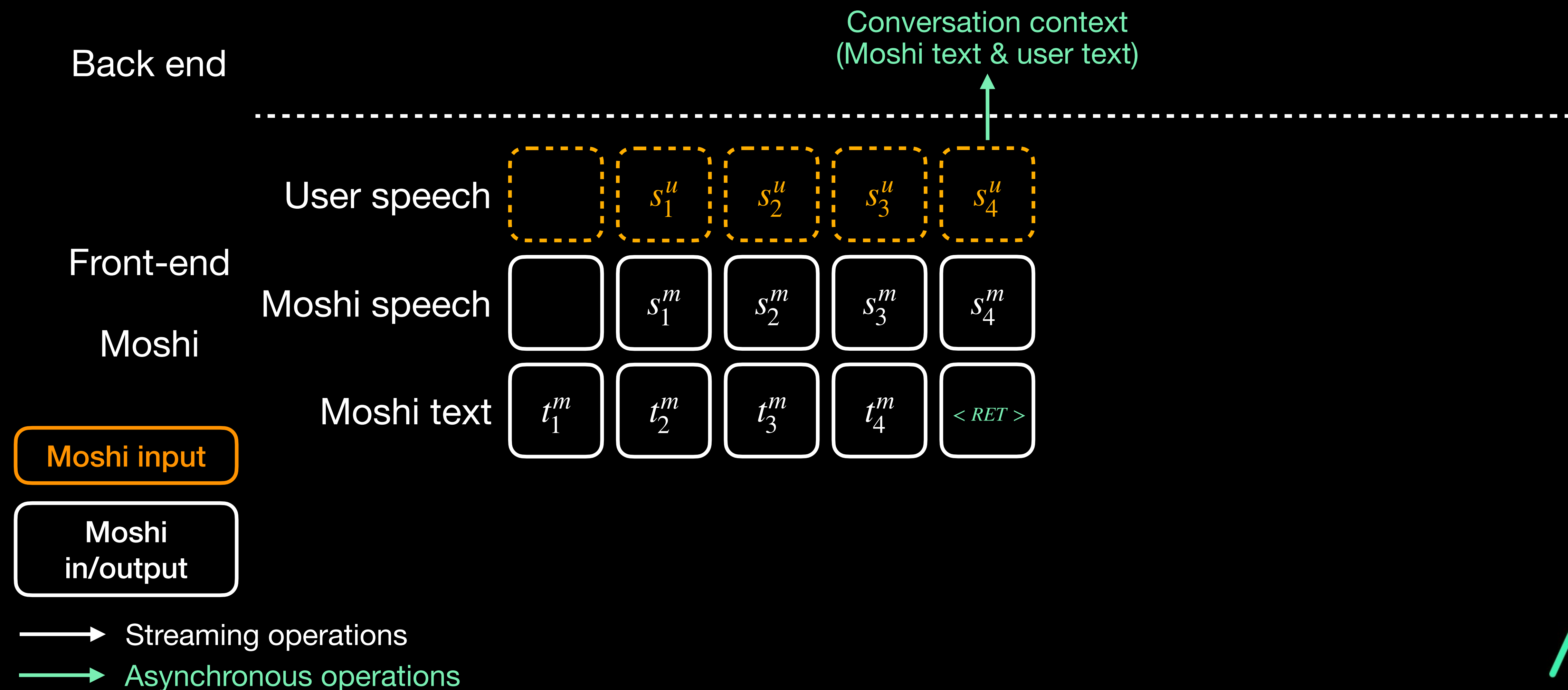
→ Streaming operations

→ Asynchronous operations

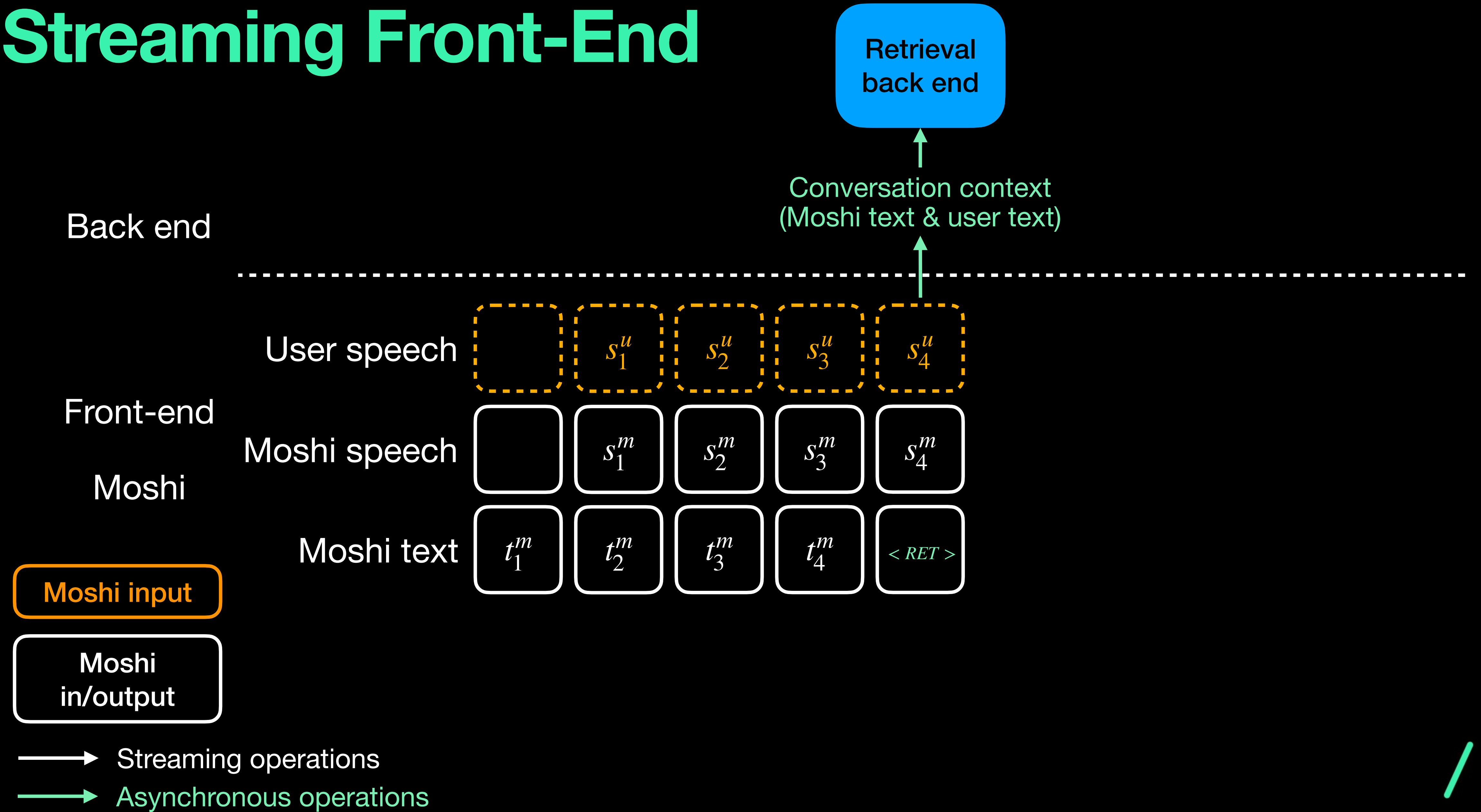
Streaming Front-End



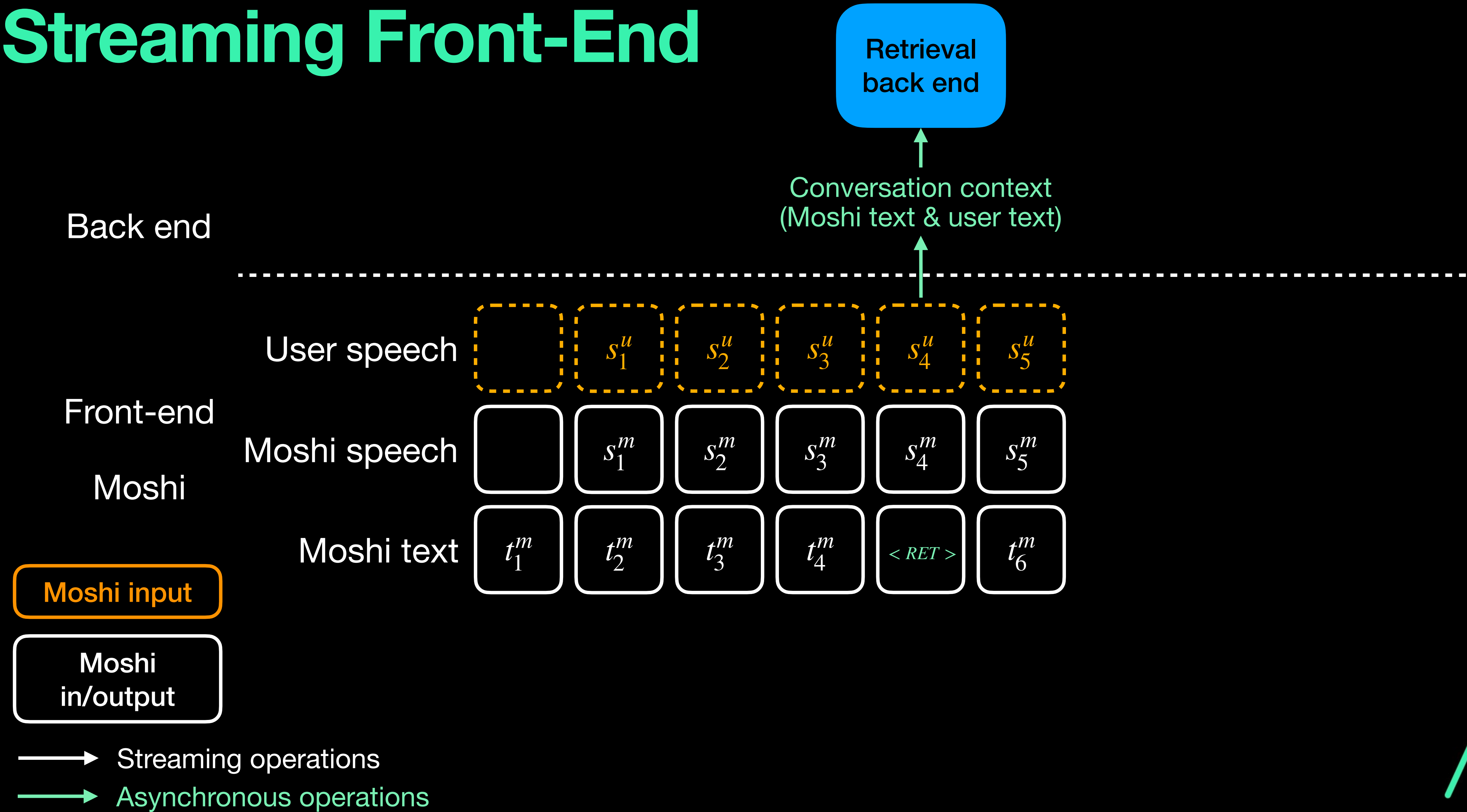
Streaming Front-End



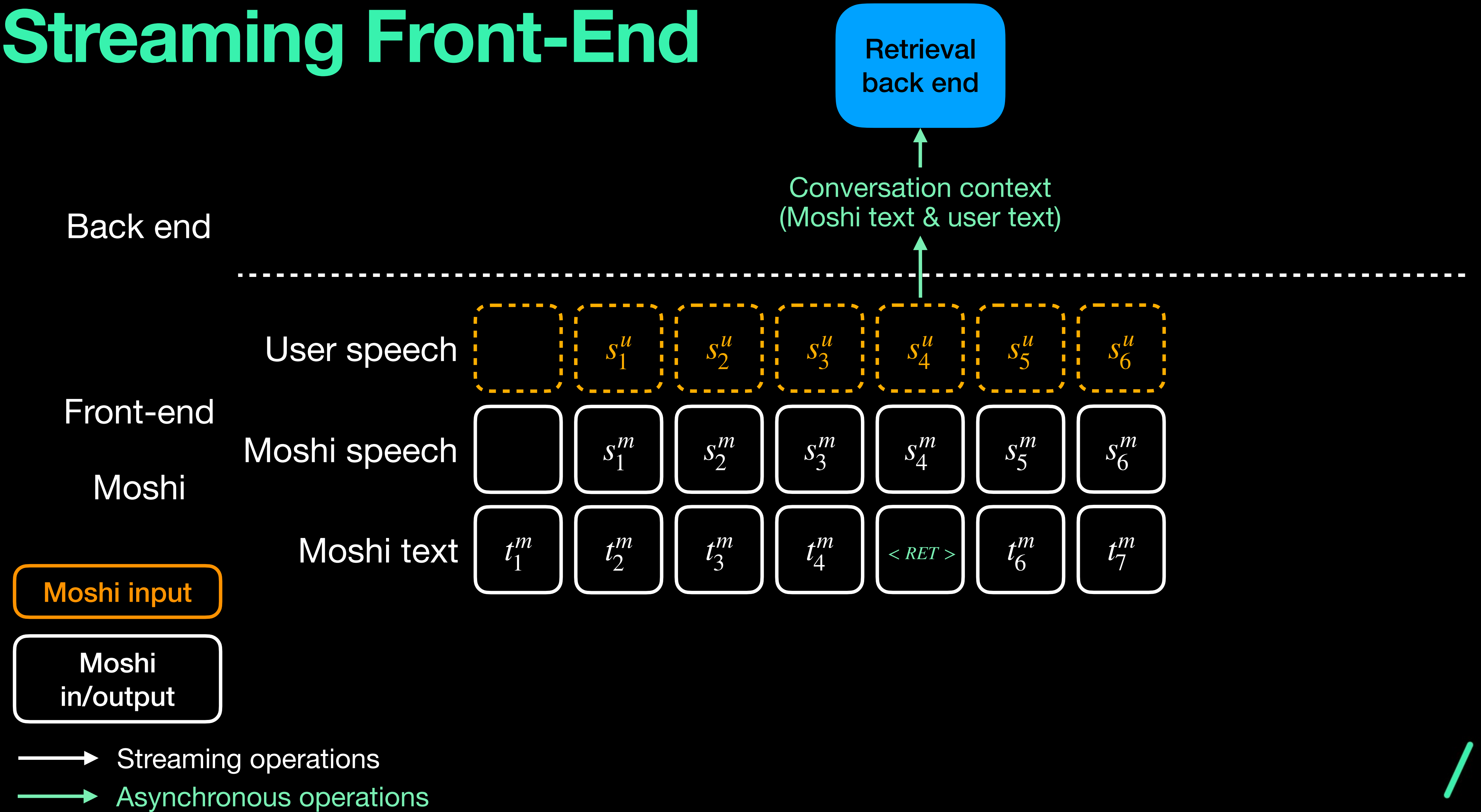
Streaming Front-End



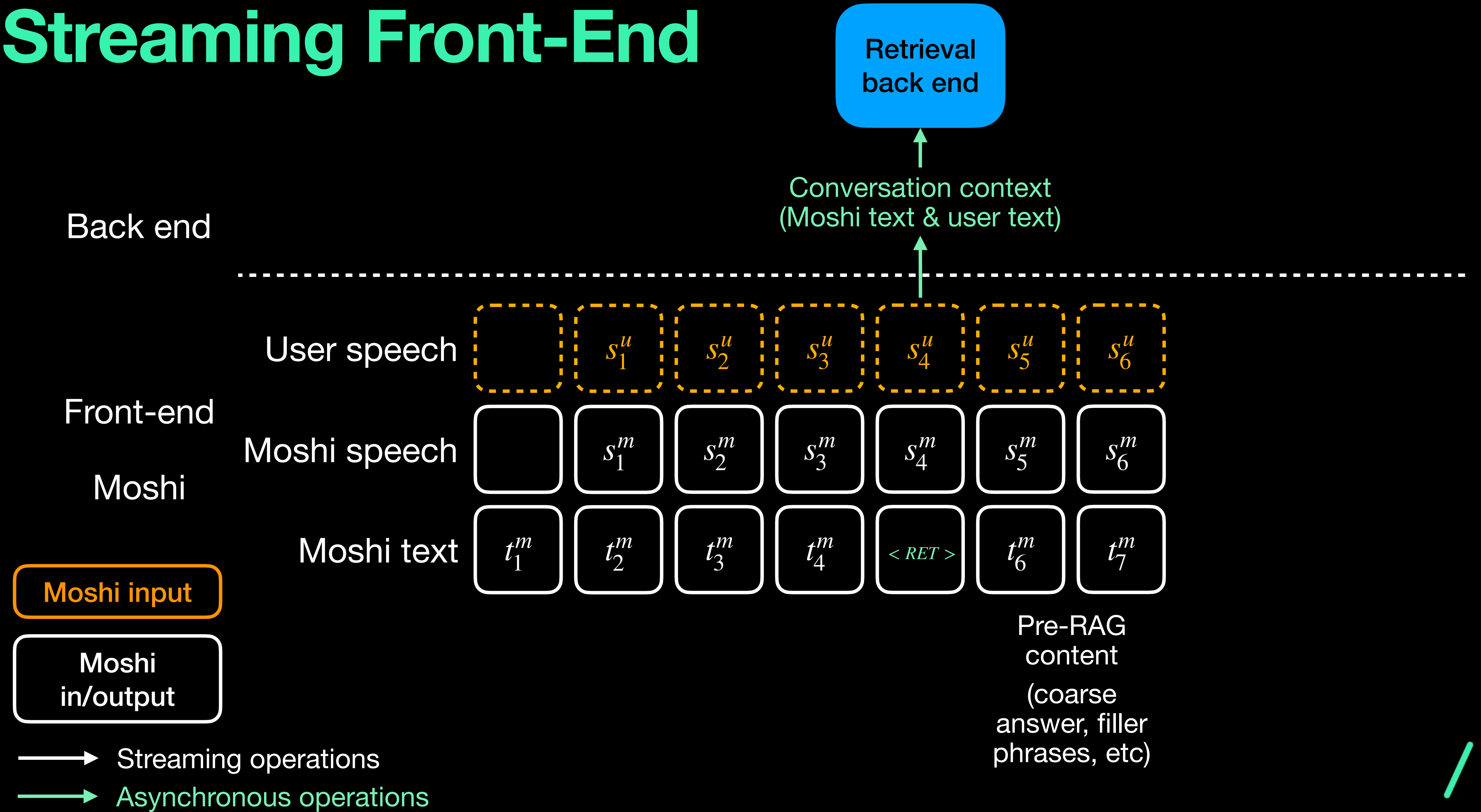
Streaming Front-End



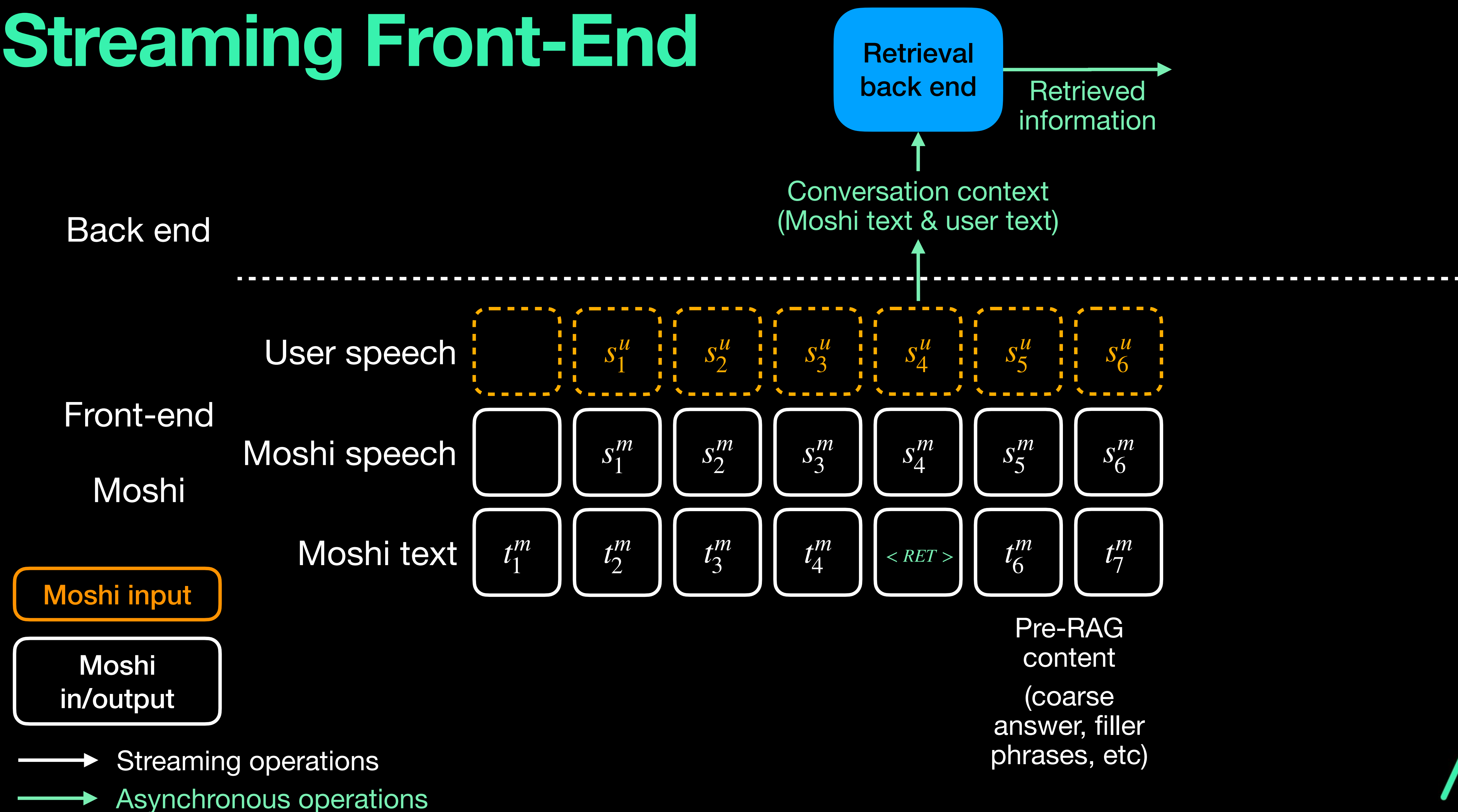
Streaming Front-End



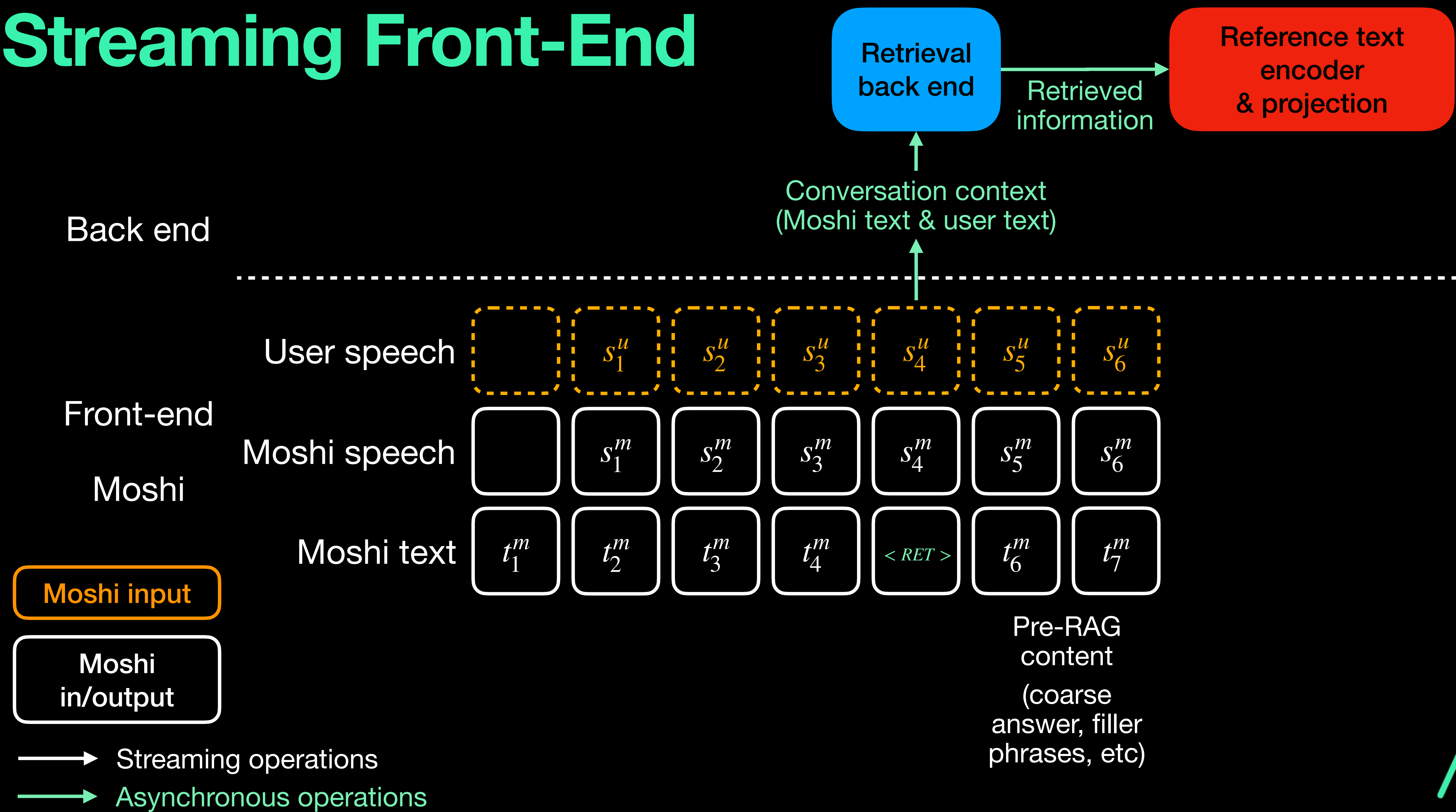
Streaming Front-End



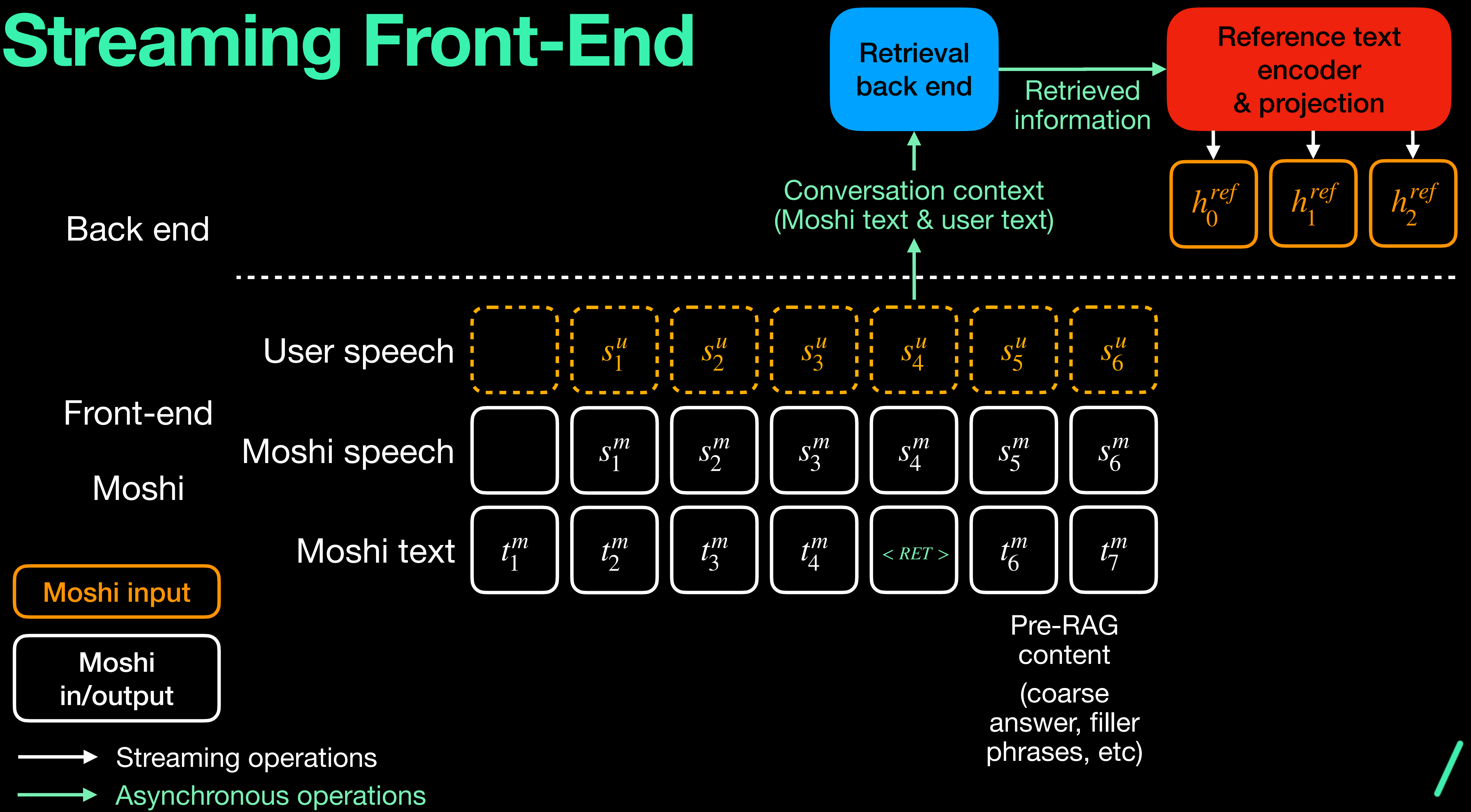
Streaming Front-End



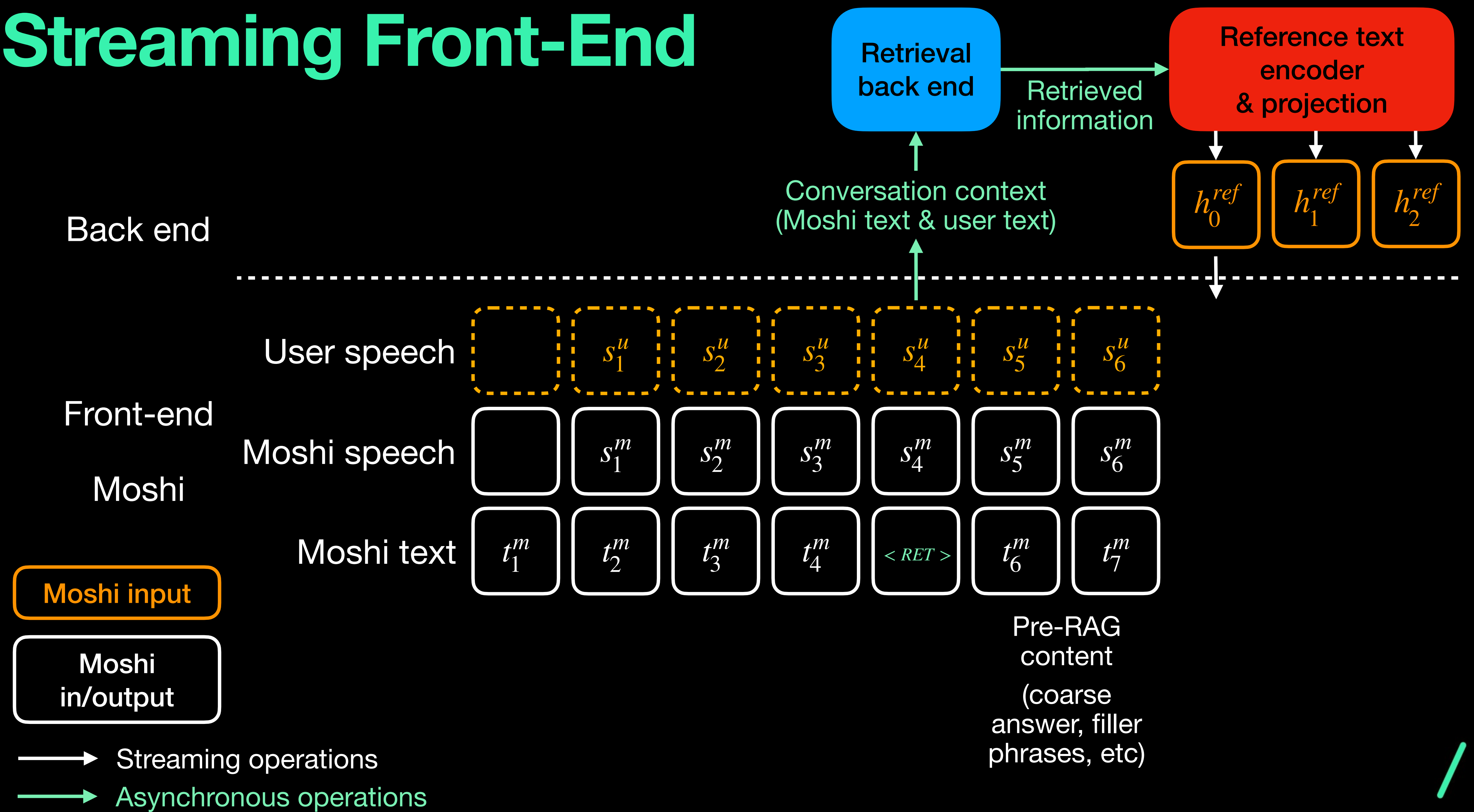
Streaming Front-End



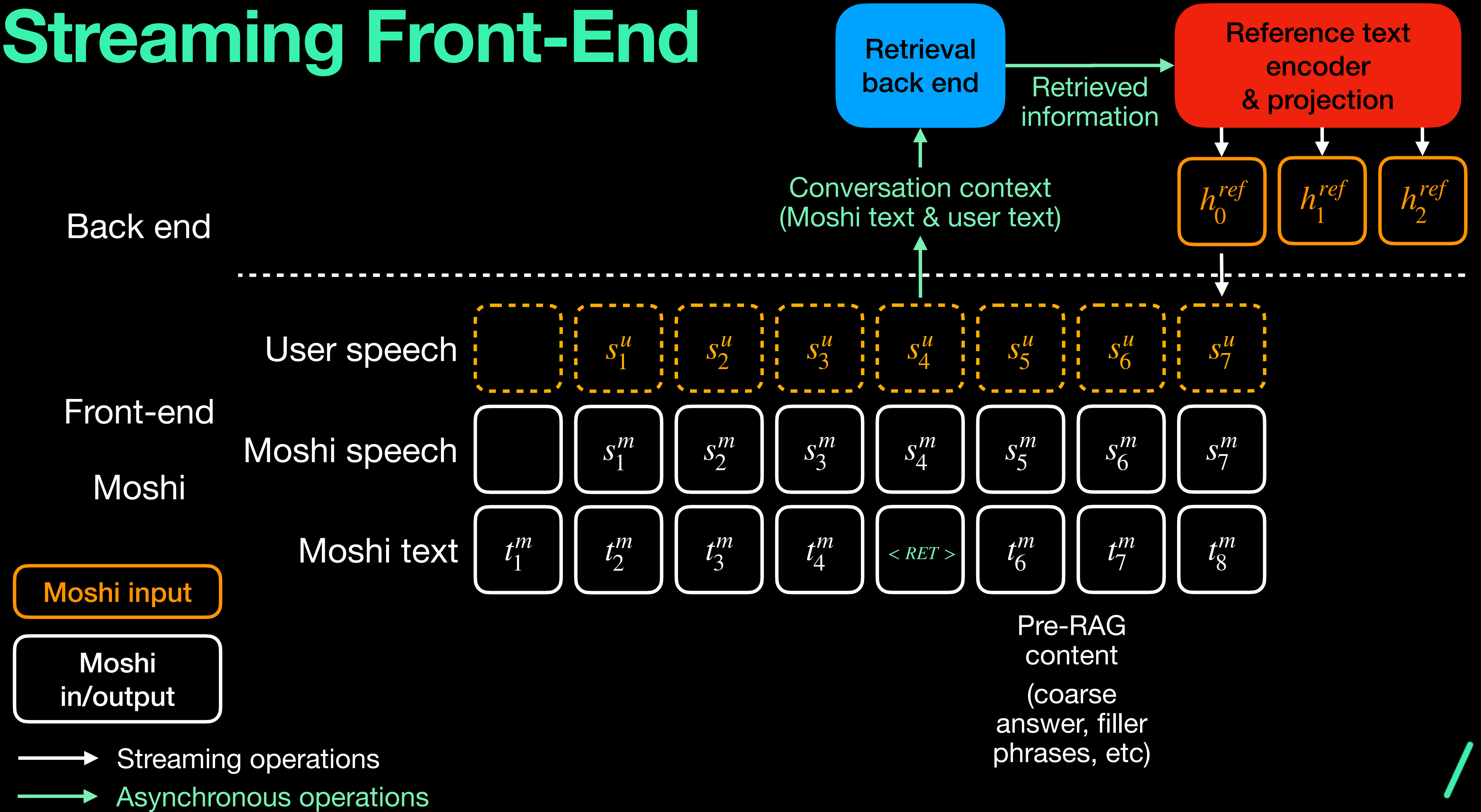
Streaming Front-End



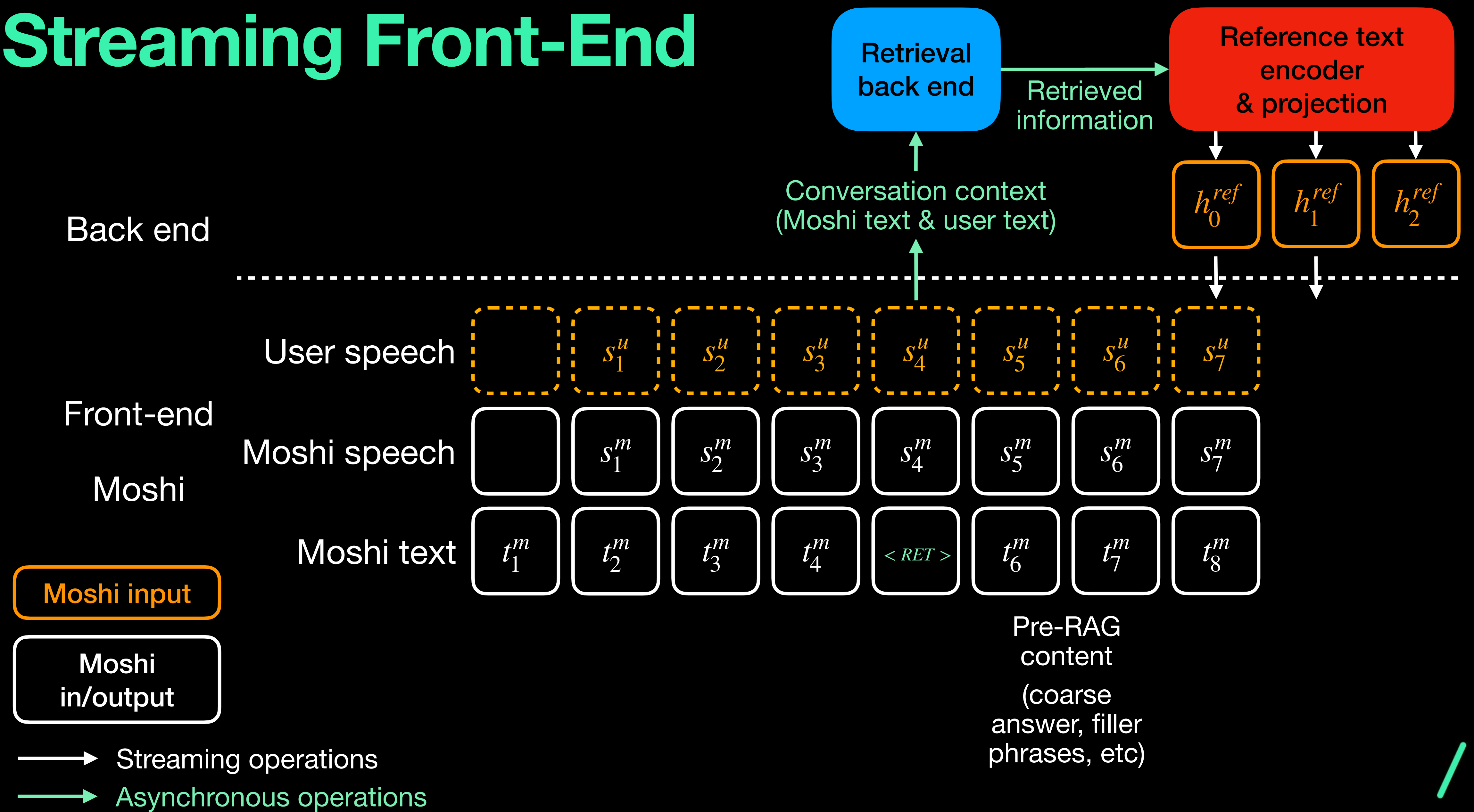
Streaming Front-End



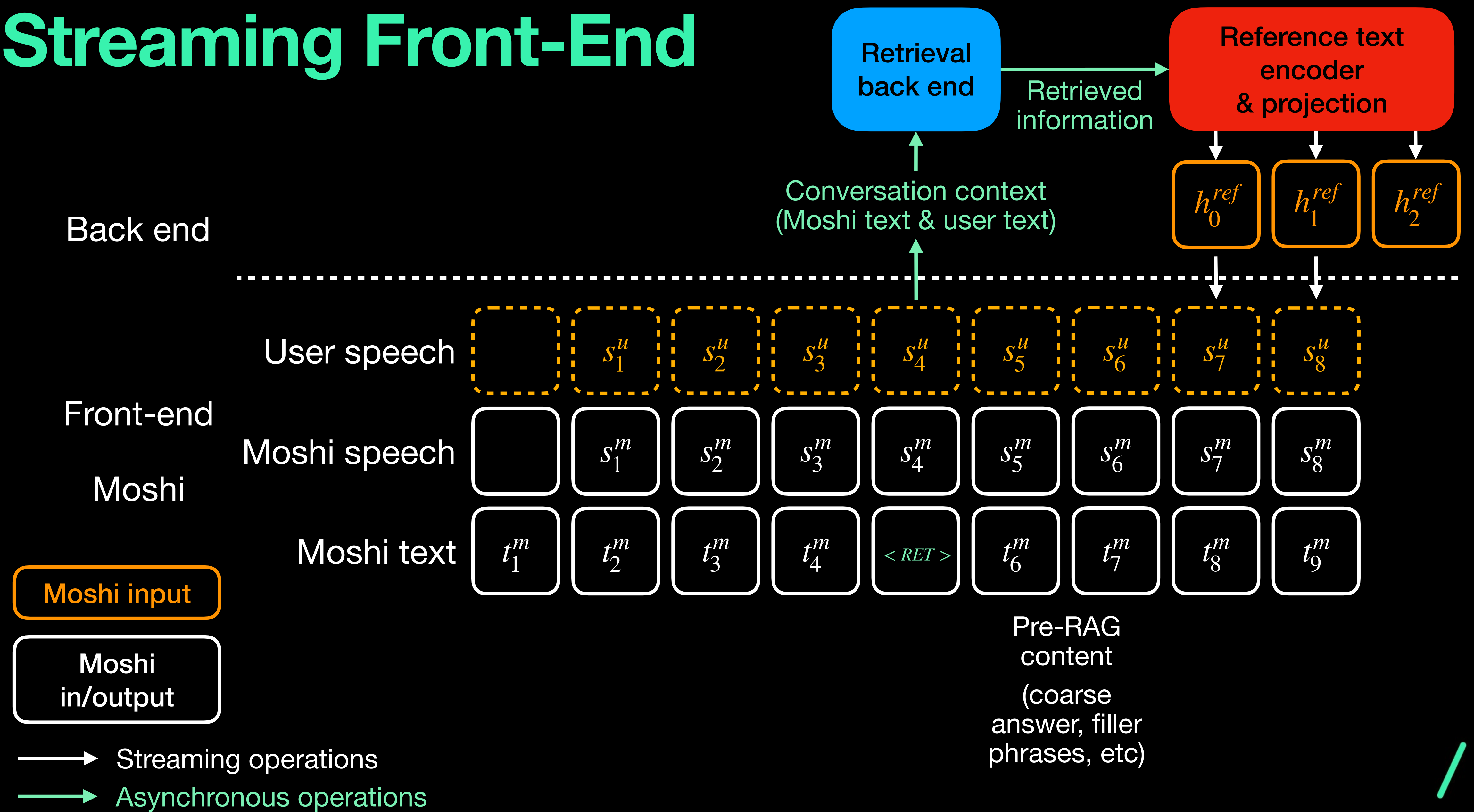
Streaming Front-End



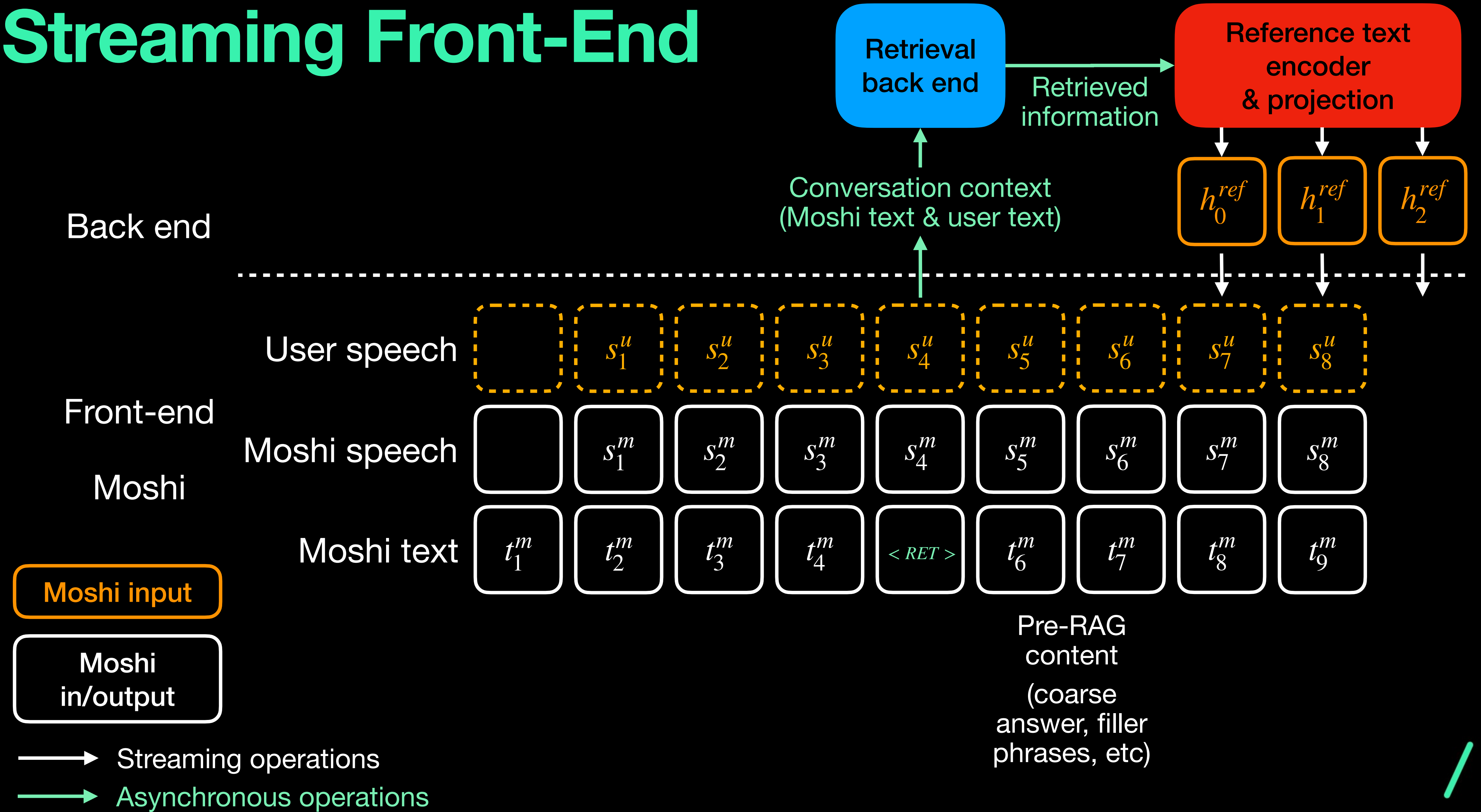
Streaming Front-End



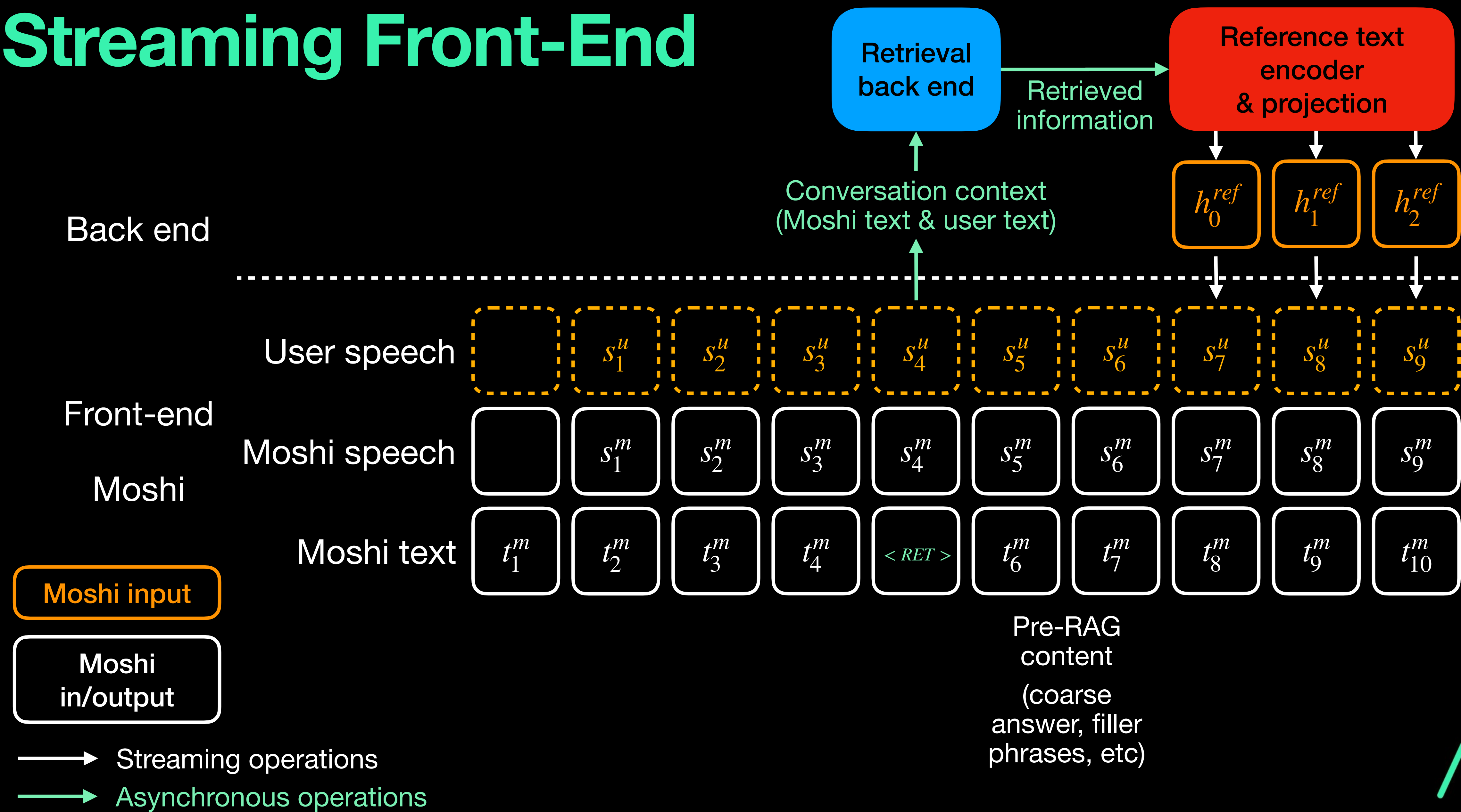
Streaming Front-End



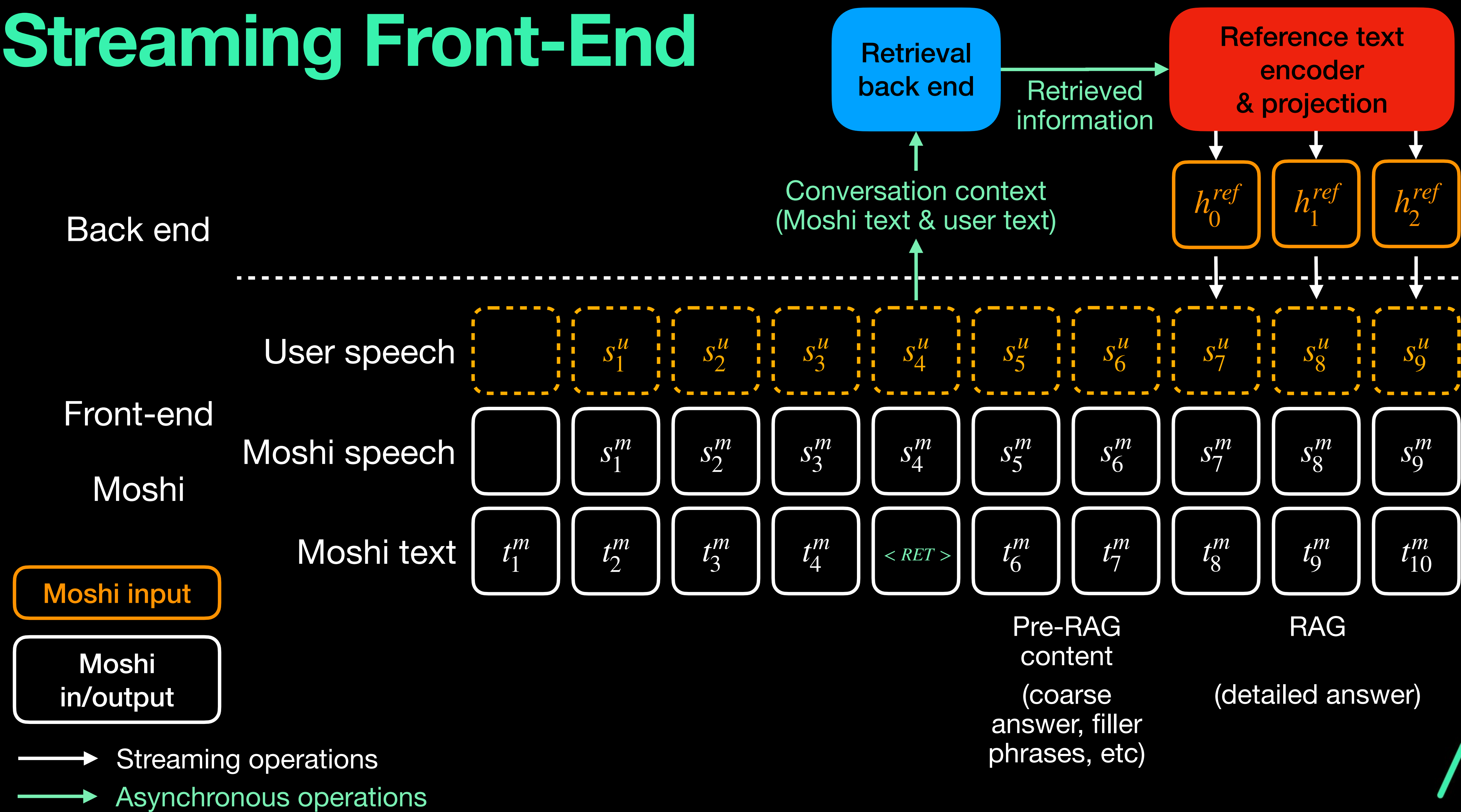
Streaming Front-End



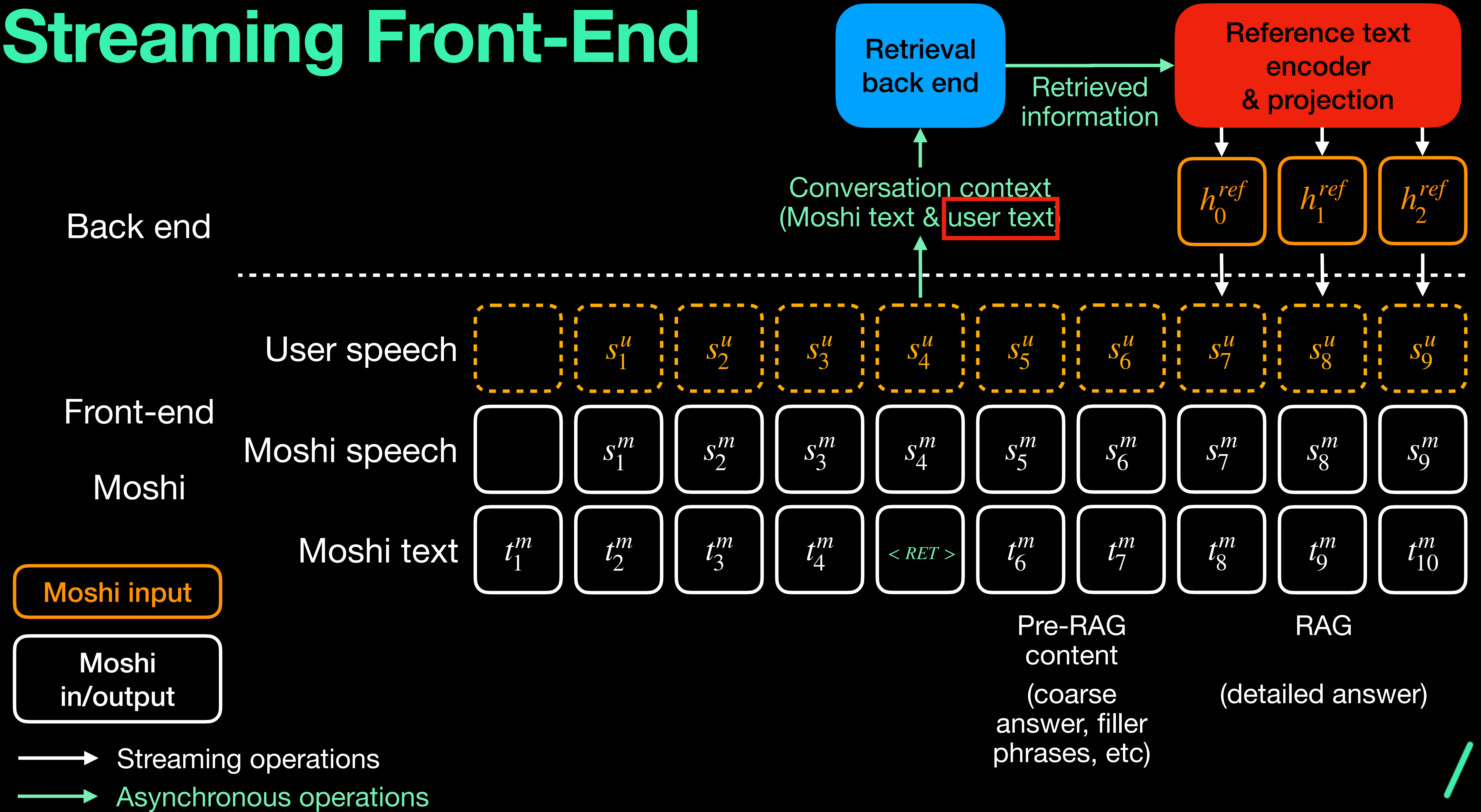
Streaming Front-End



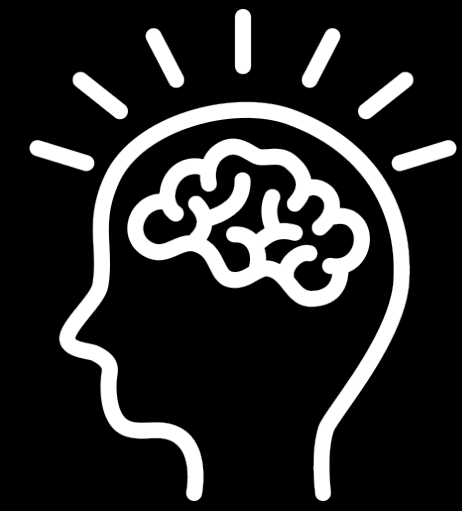
Streaming Front-End



Streaming Front-End



System Design



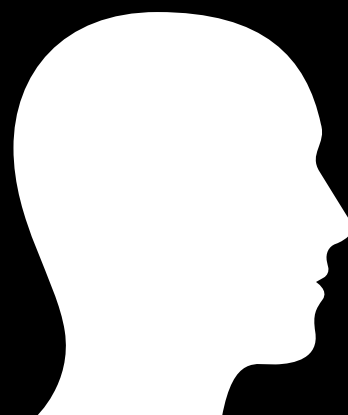
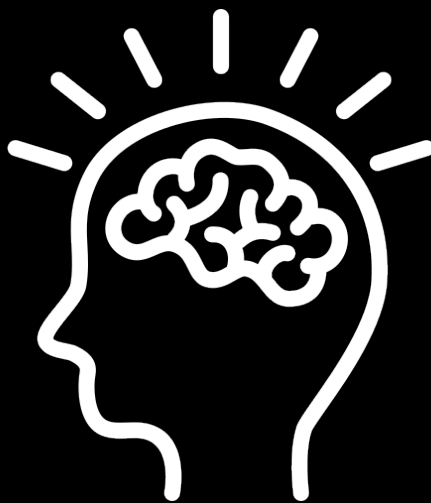
- Streaming operations
- Asynchronous operations

Front end

Back end



System Design



Streaming operations



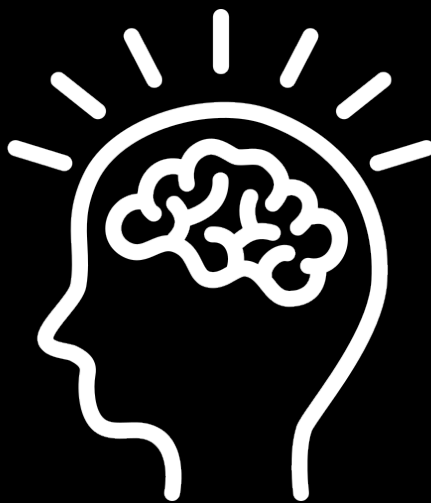
Asynchronous operations

Front end

Back end



System Design



user speech

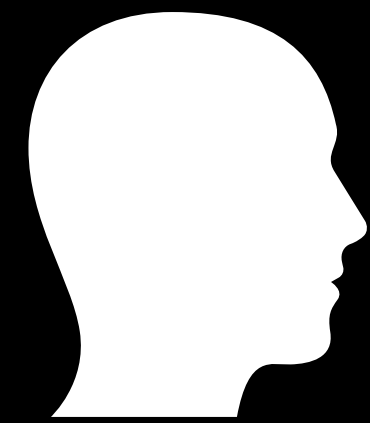
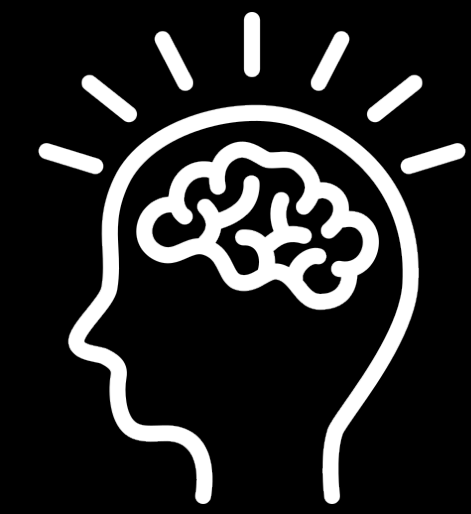
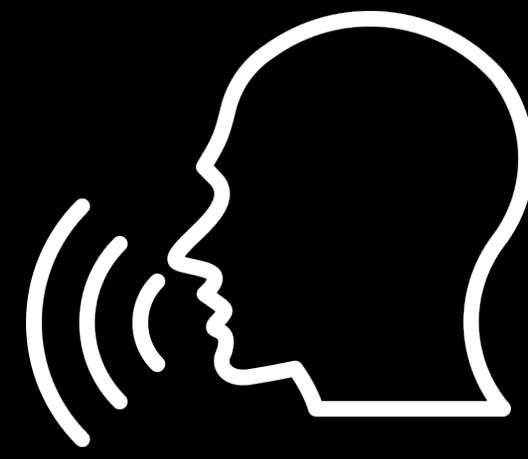
- Streaming operations
- Asynchronous operations

Front end

Back end



System Design



user speech



Moshi



Streaming operations

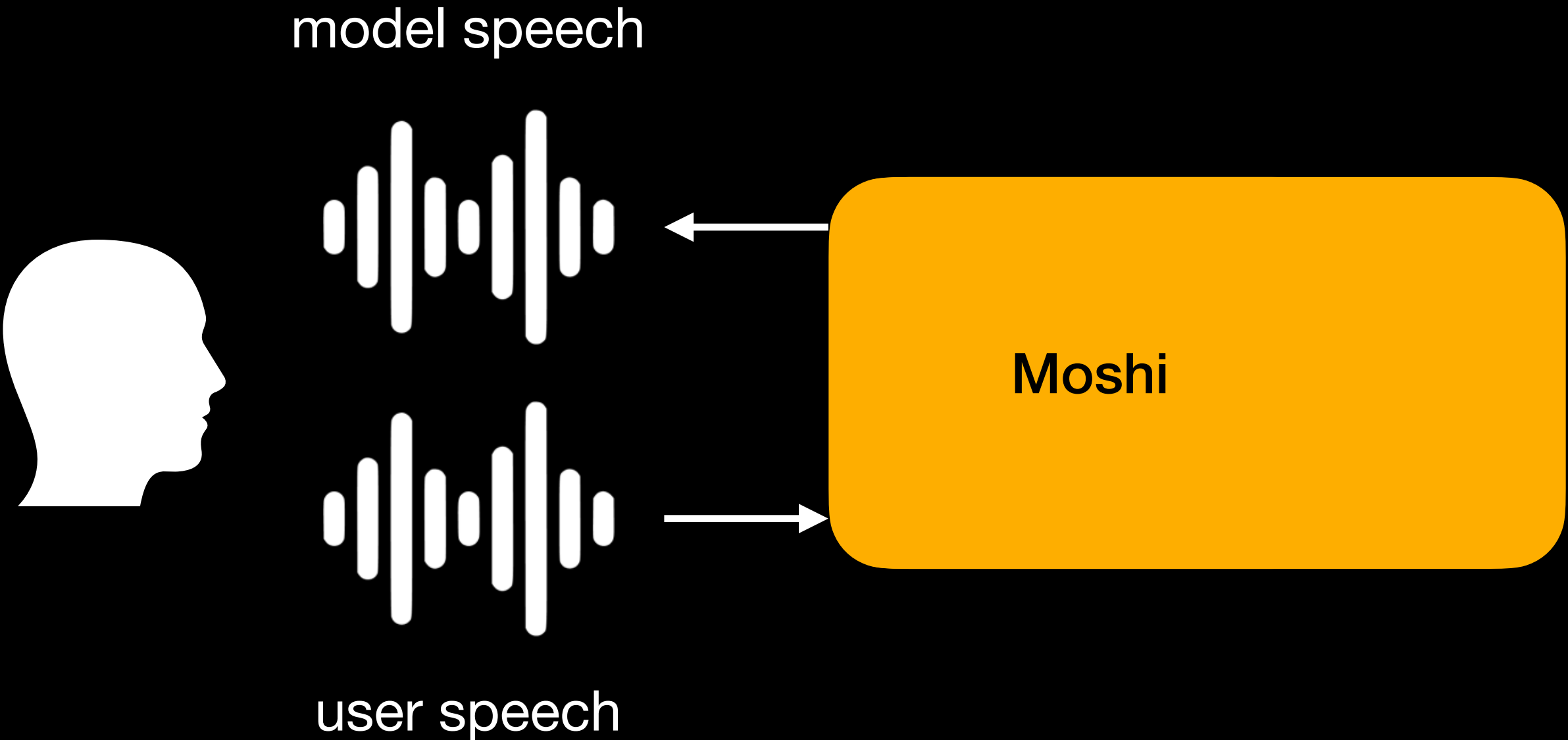
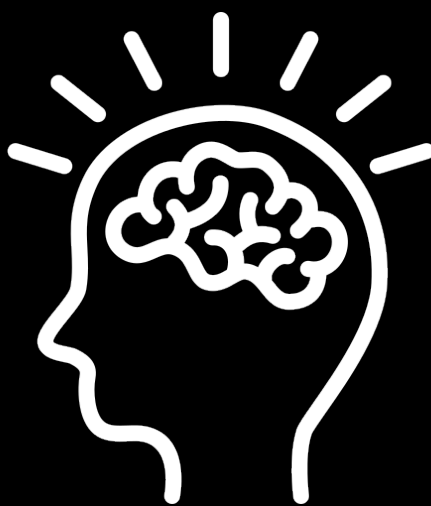


Asynchronous operations

Front end

Back end

System Design



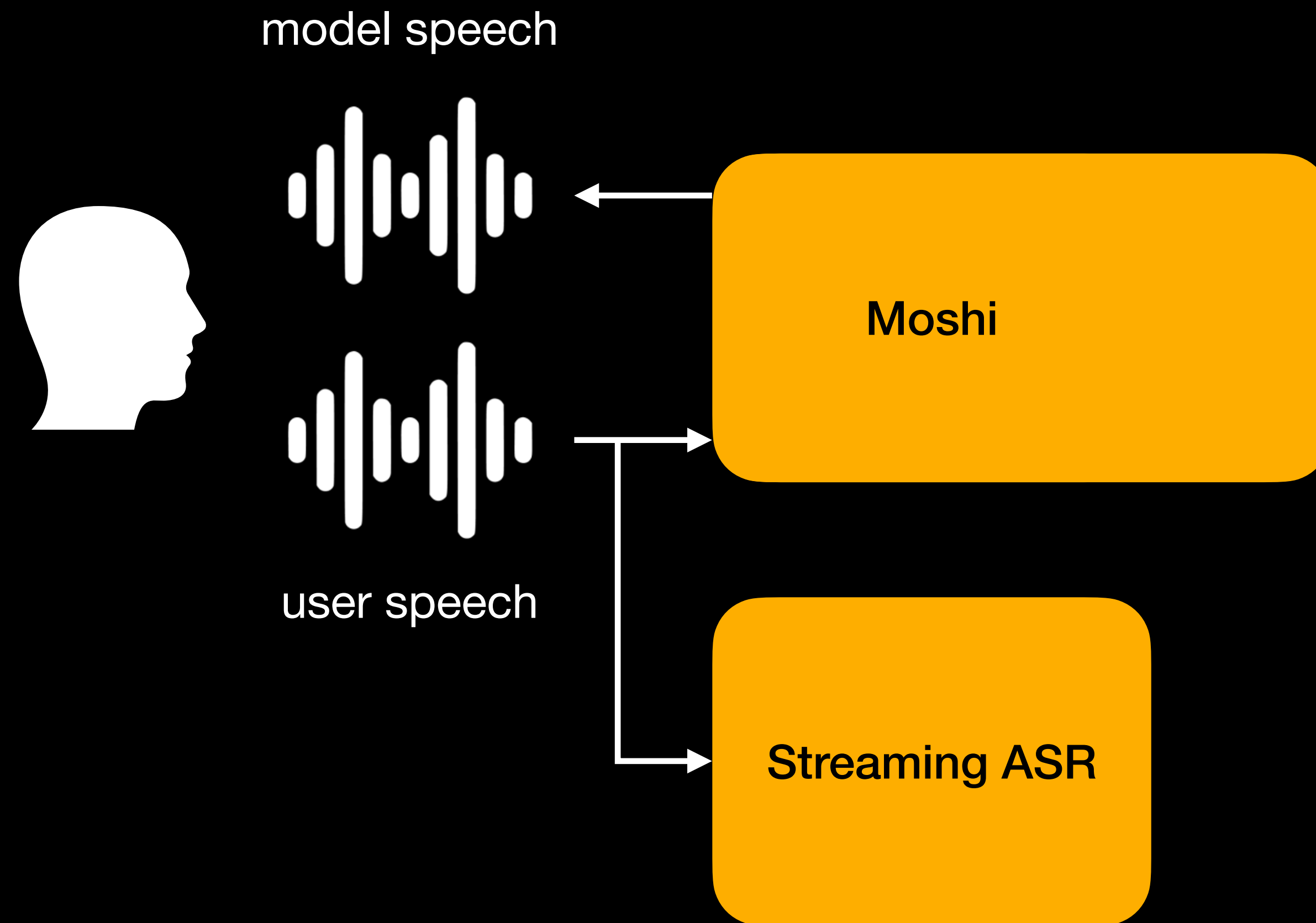
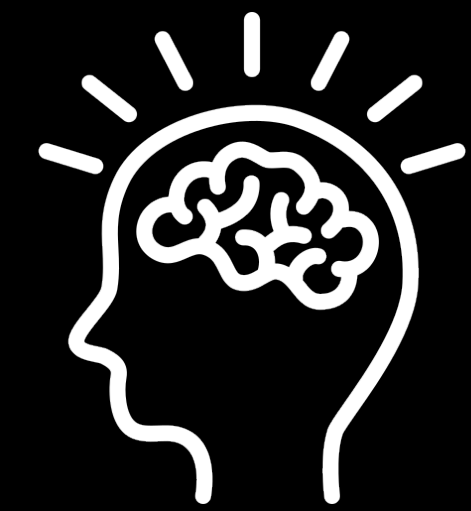
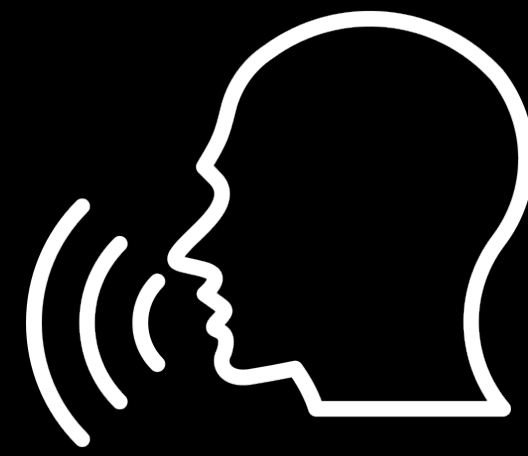
→ Streaming operations
→ Asynchronous operations

Front end

Back end



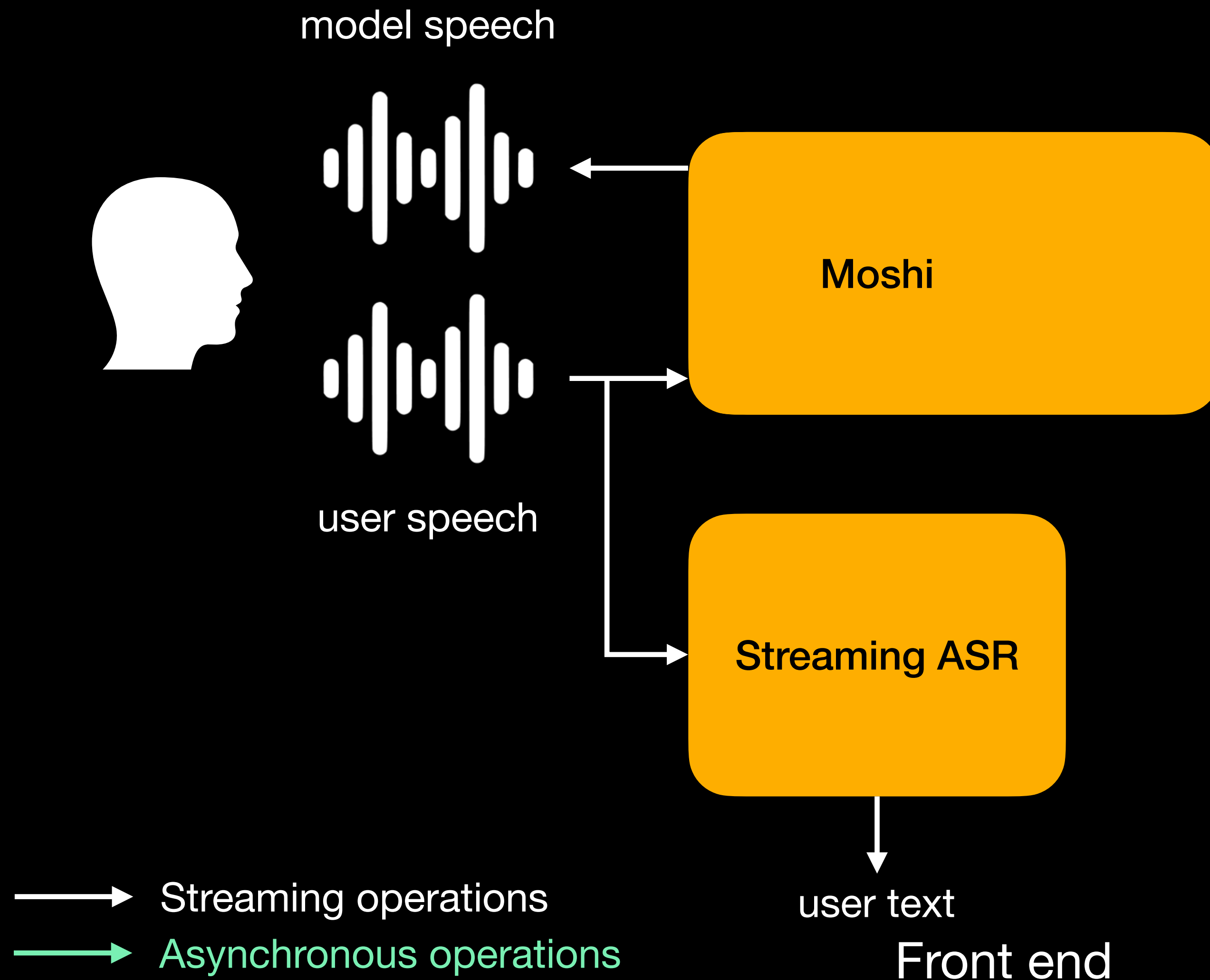
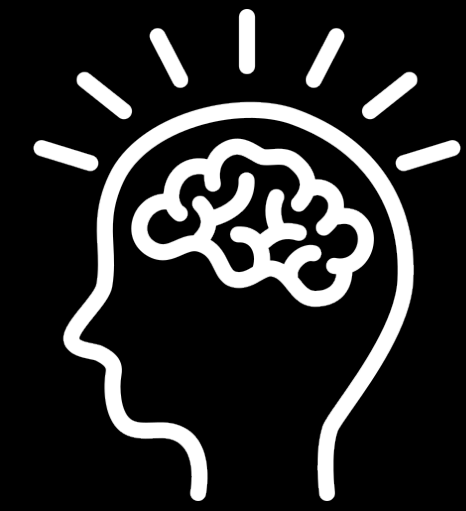
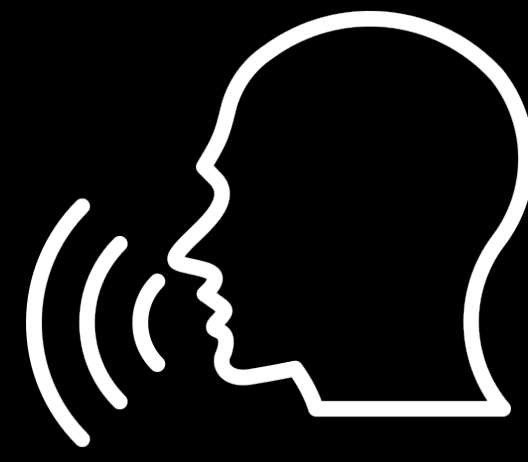
System Design



Front end

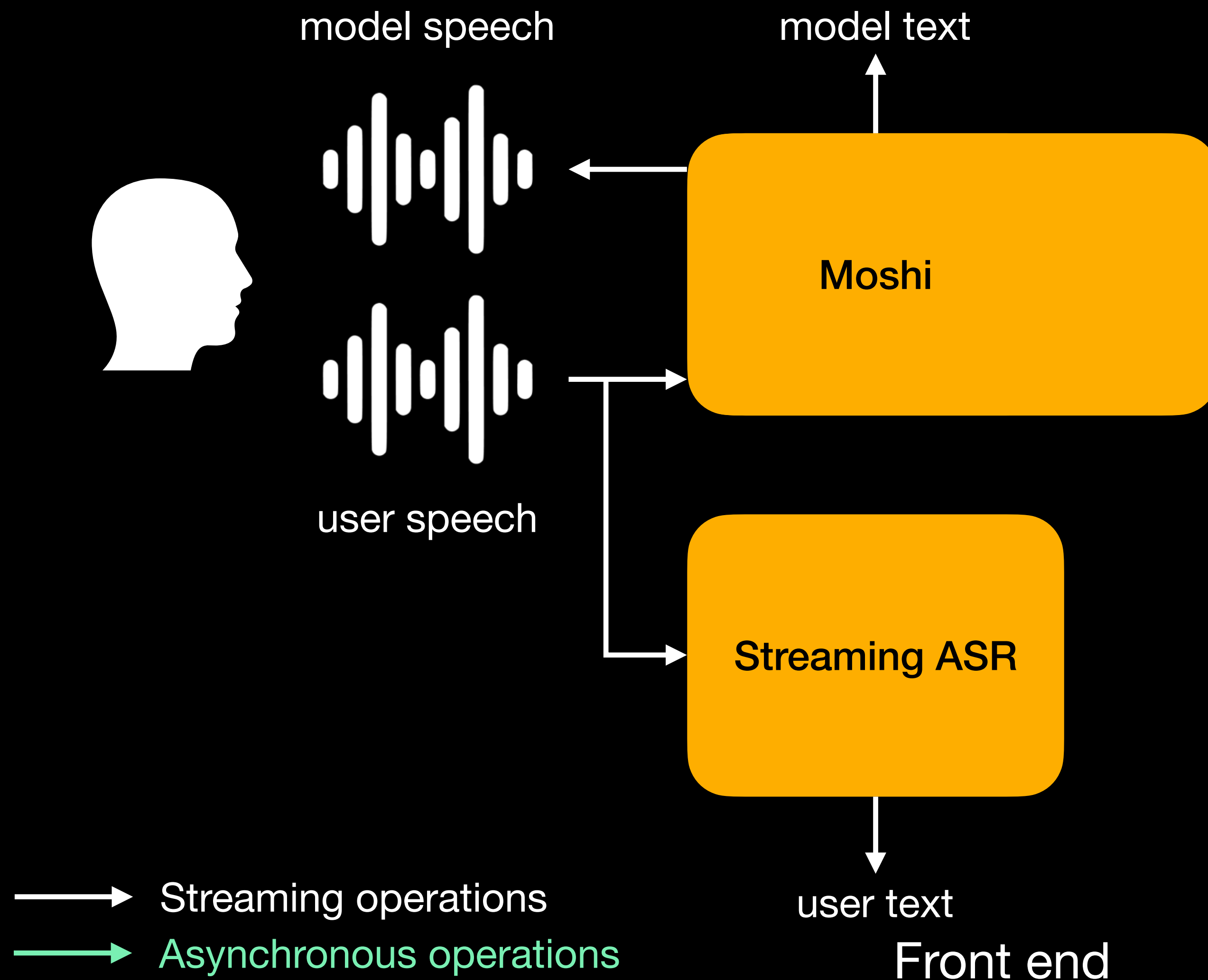
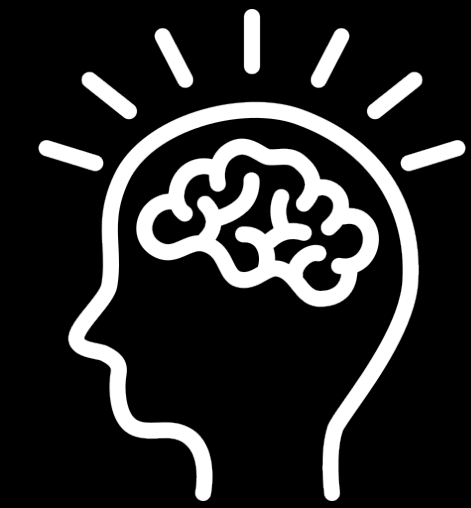
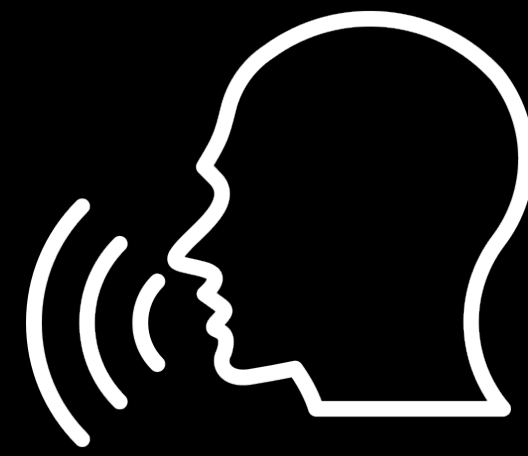
Back end

System Design

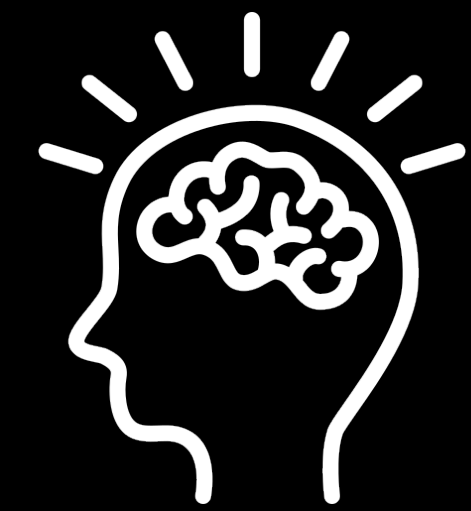
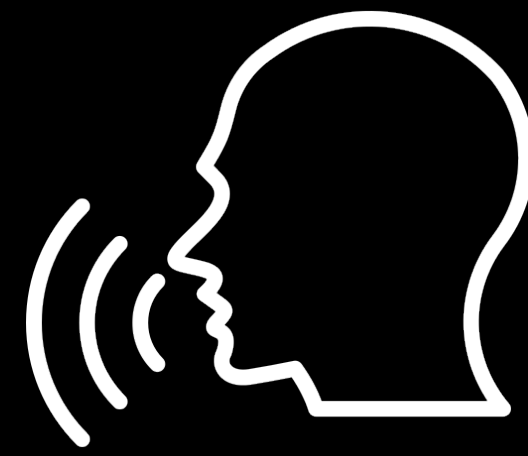


Back end

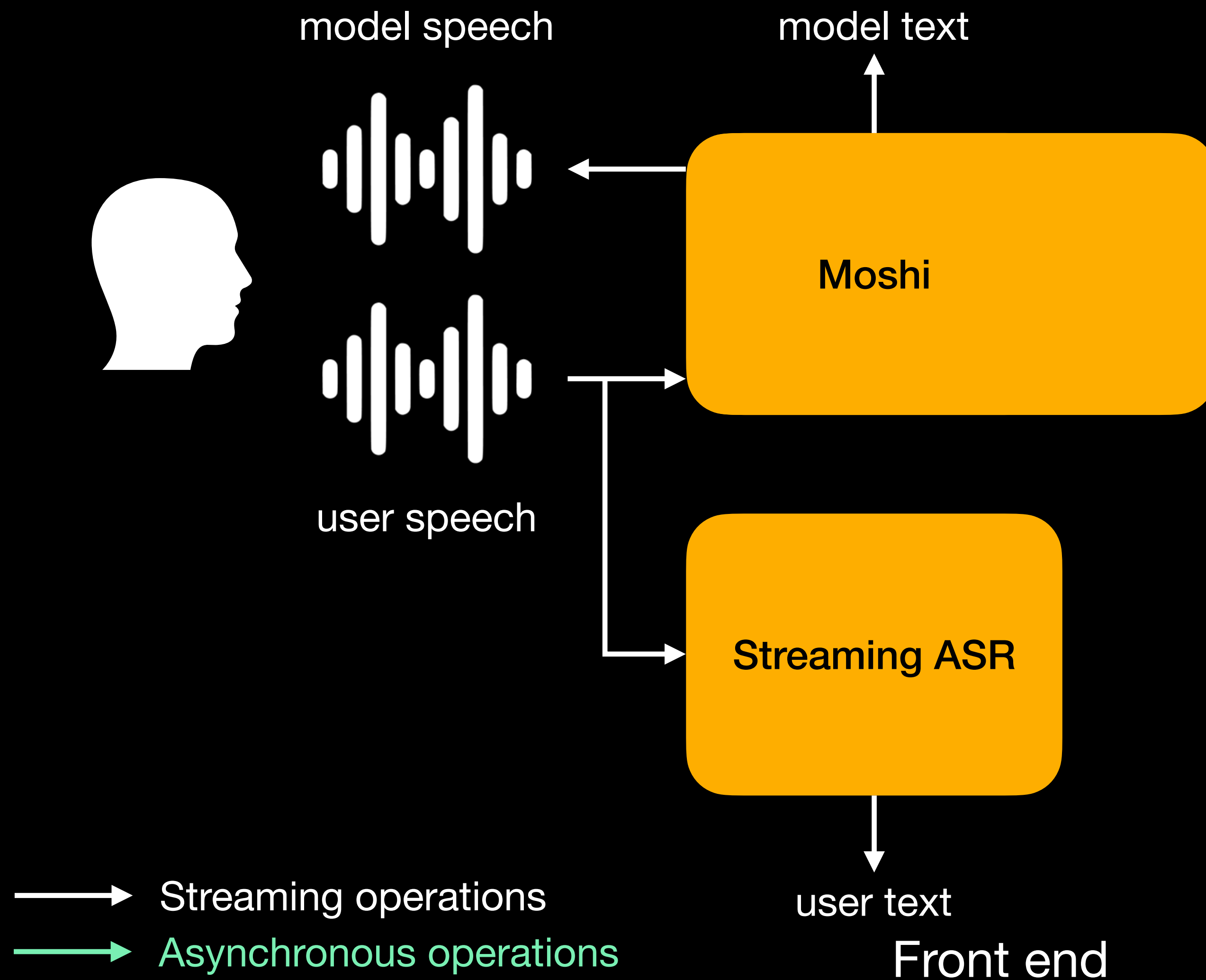
System Design



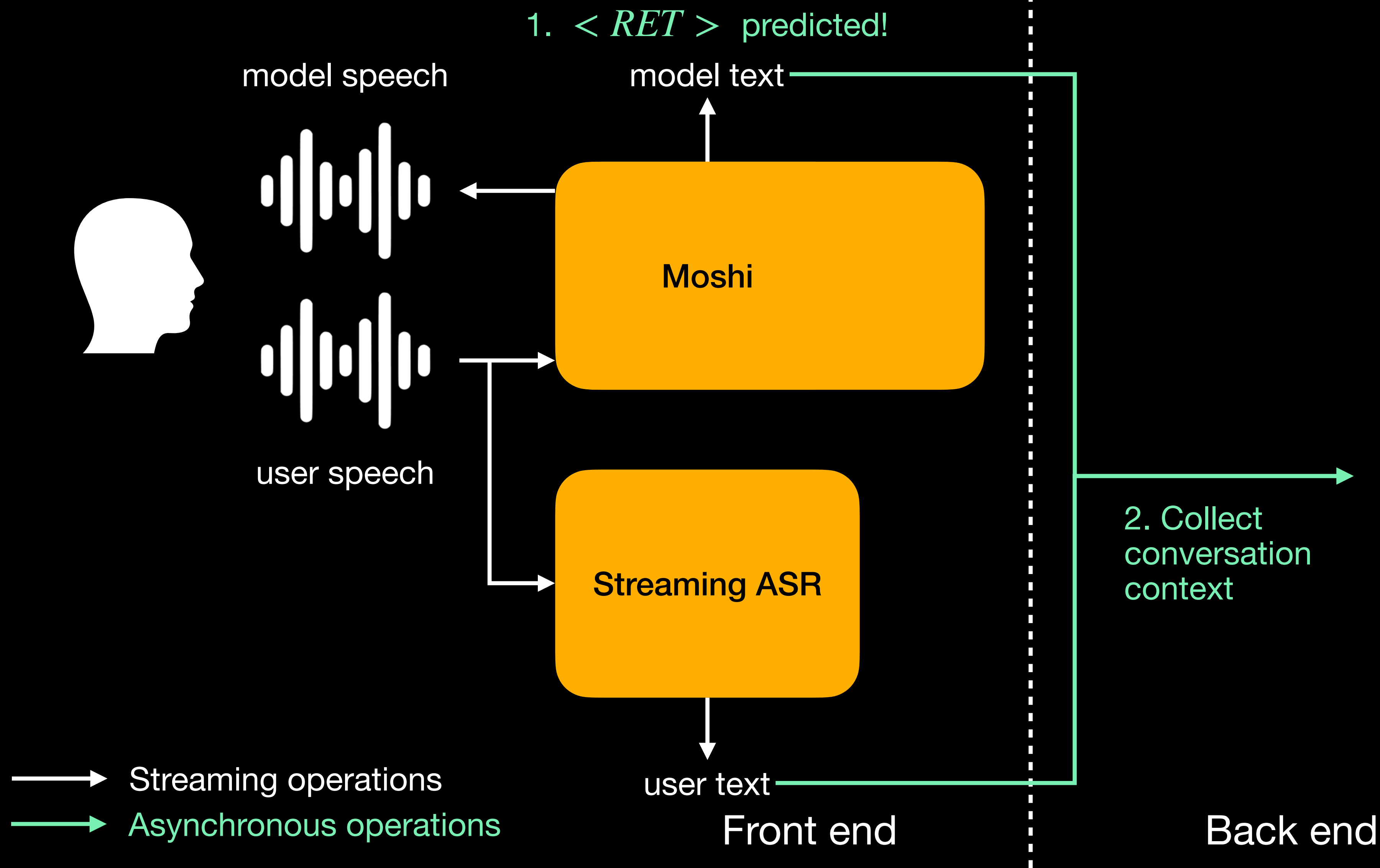
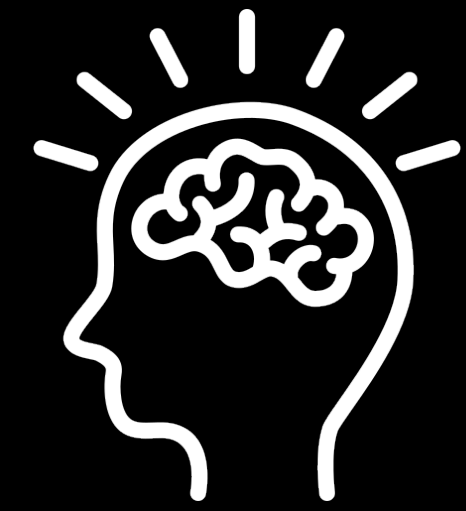
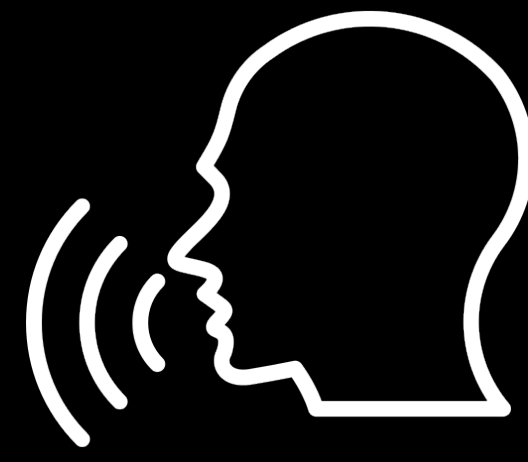
System Design



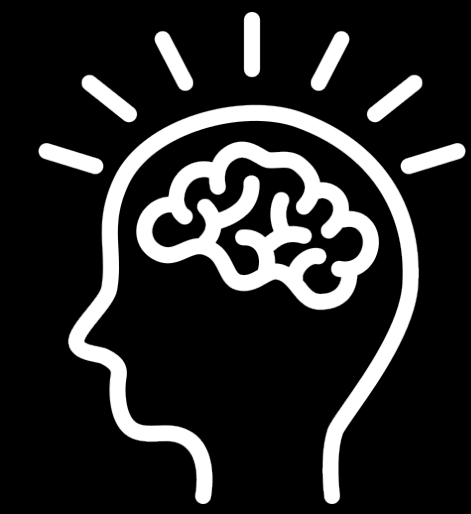
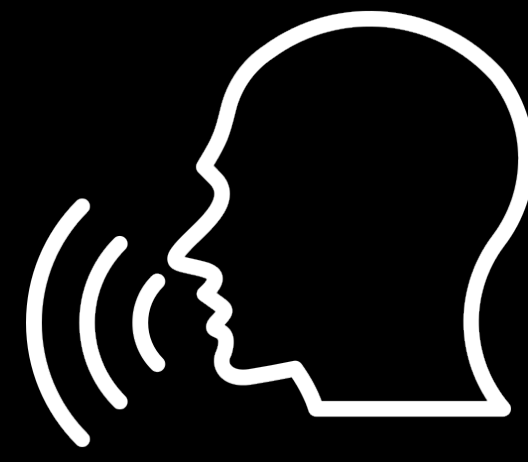
1. *<RET>* predicted!



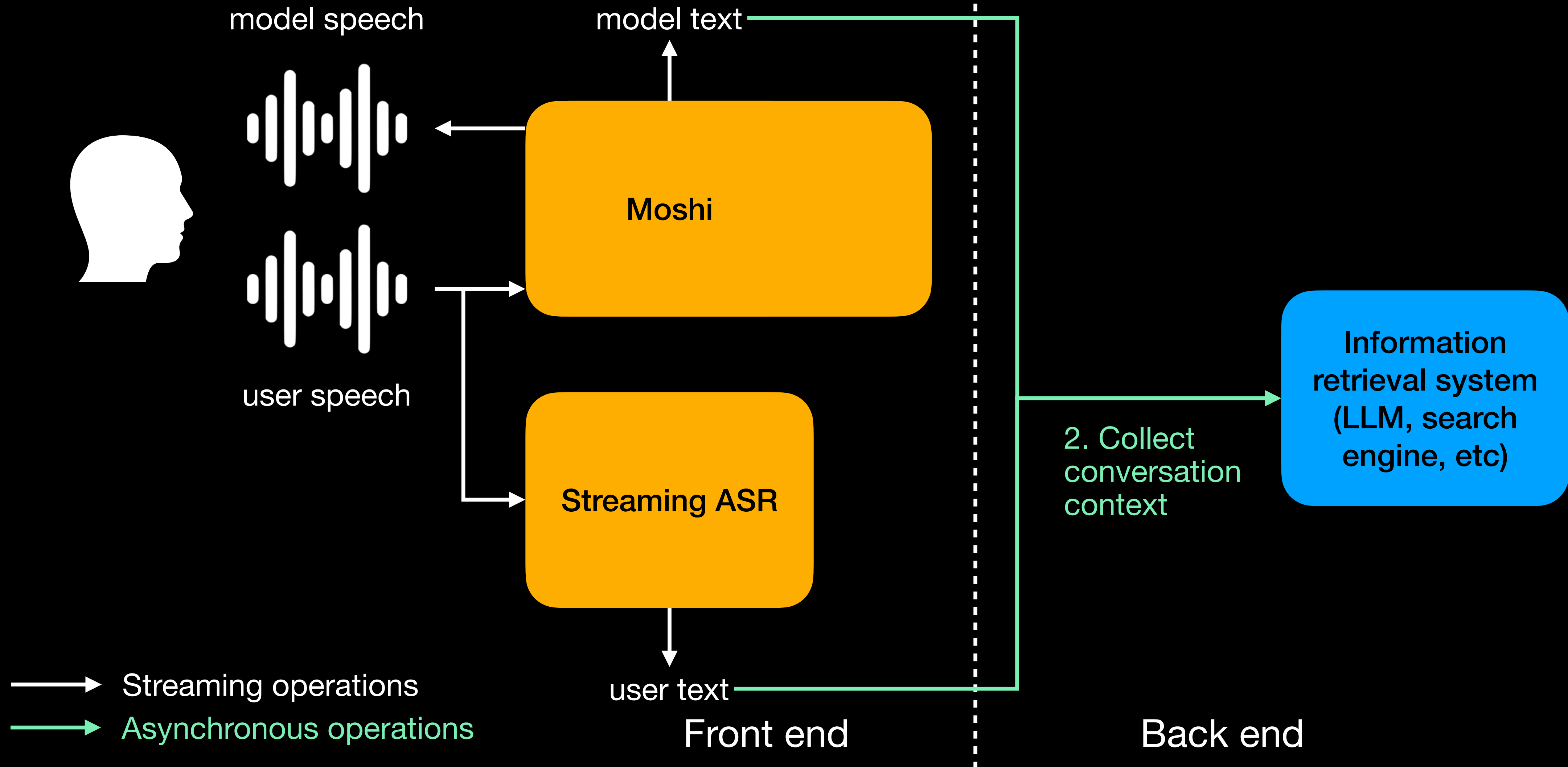
System Design



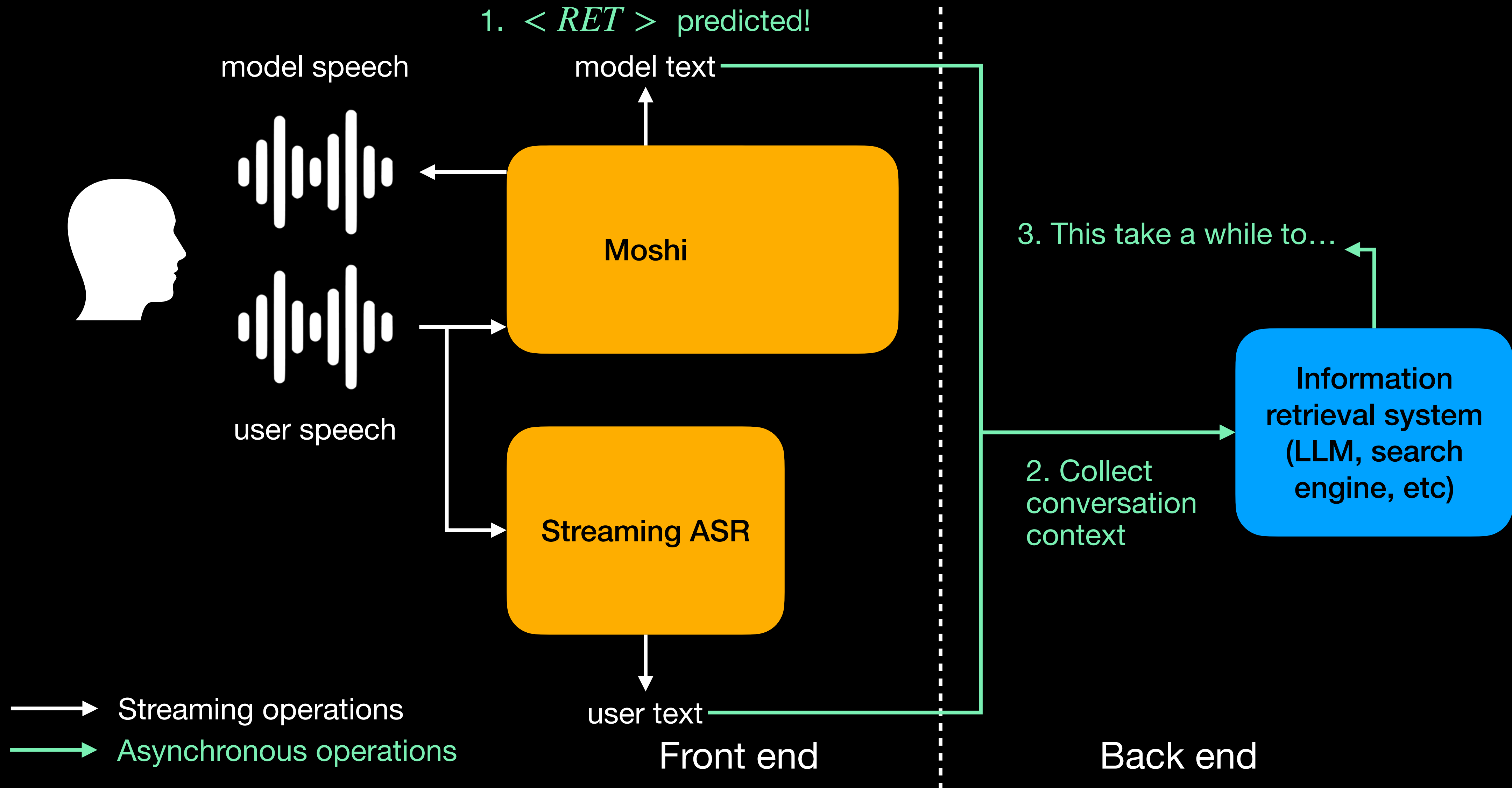
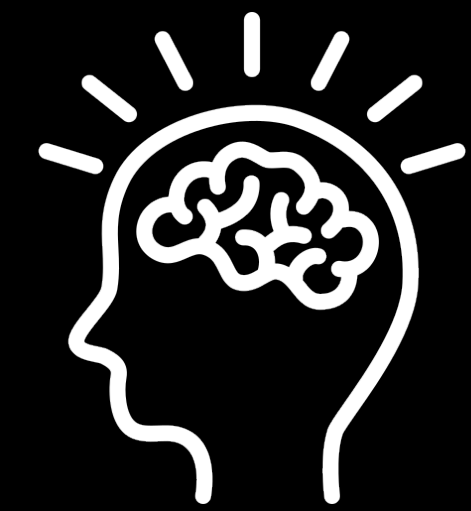
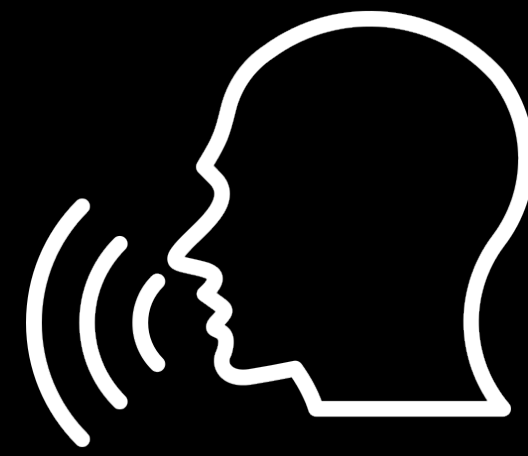
System Design



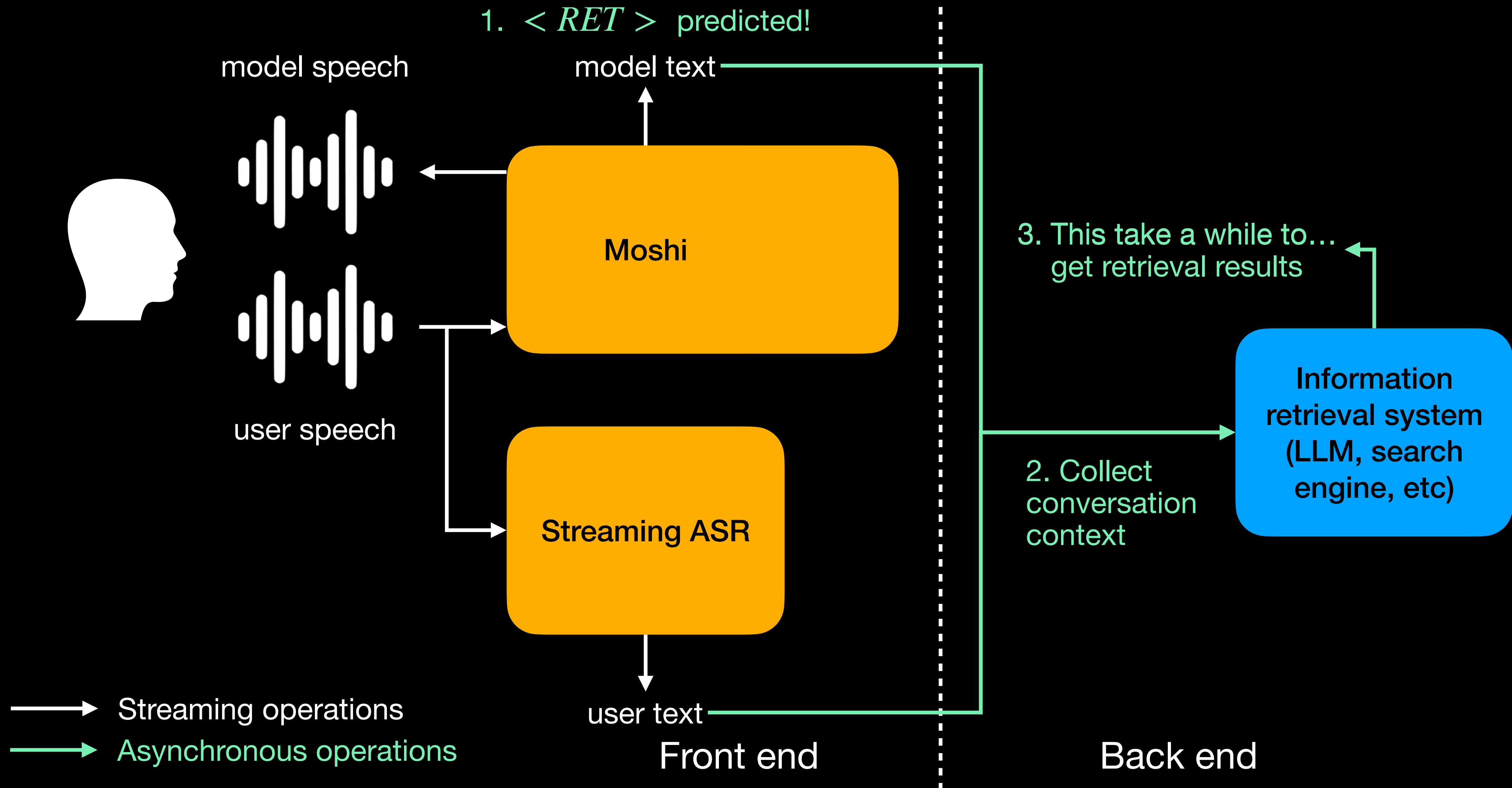
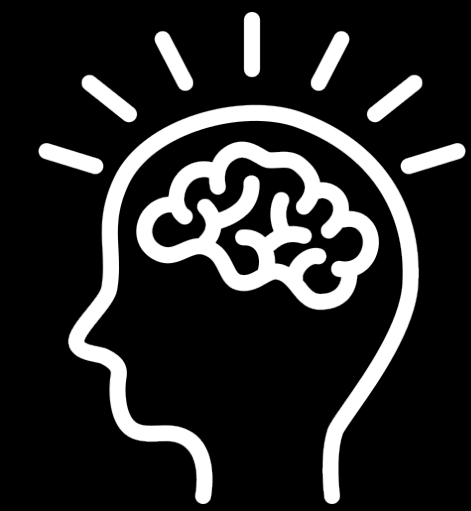
1. *< RET >* predicted!



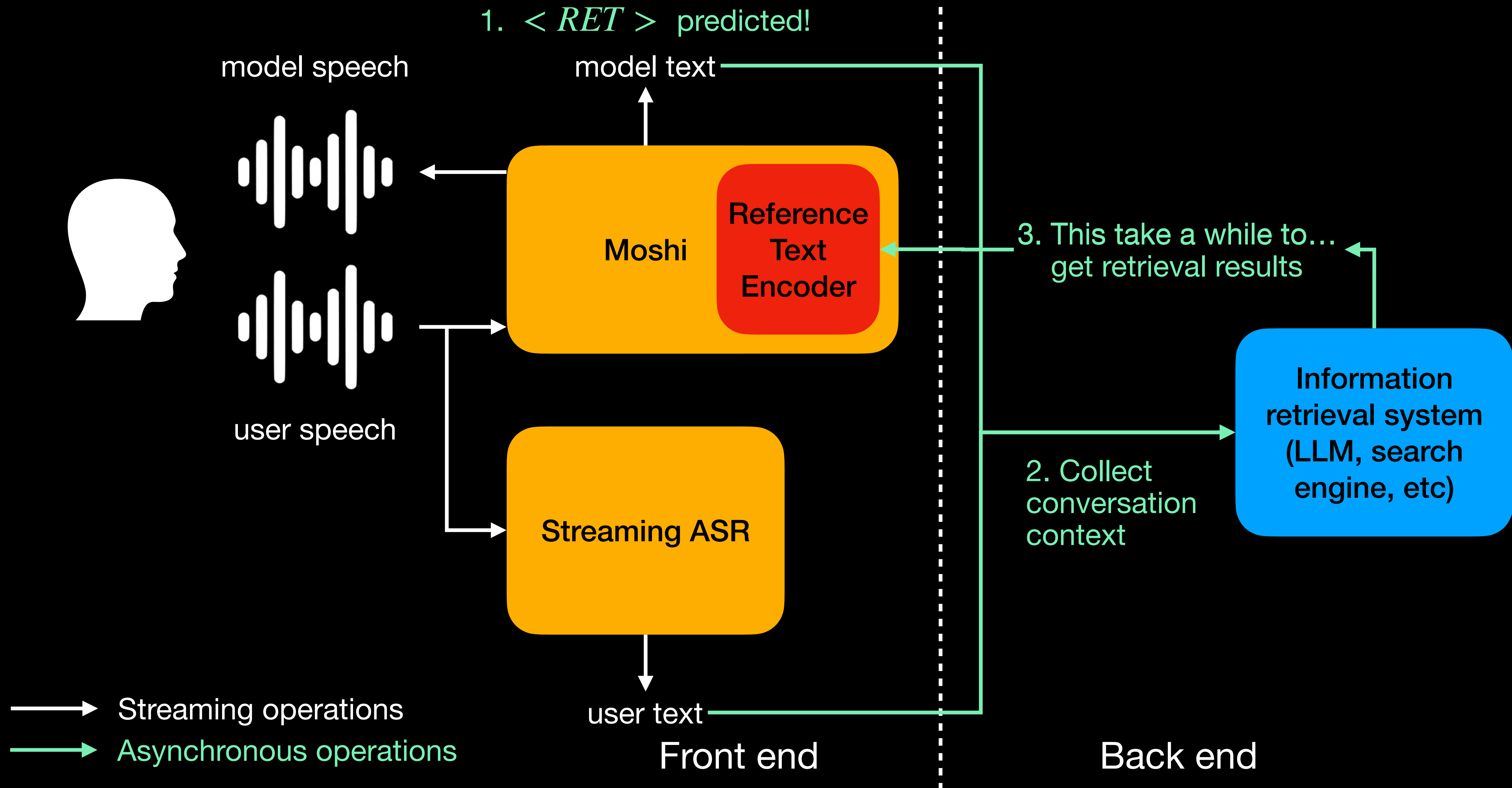
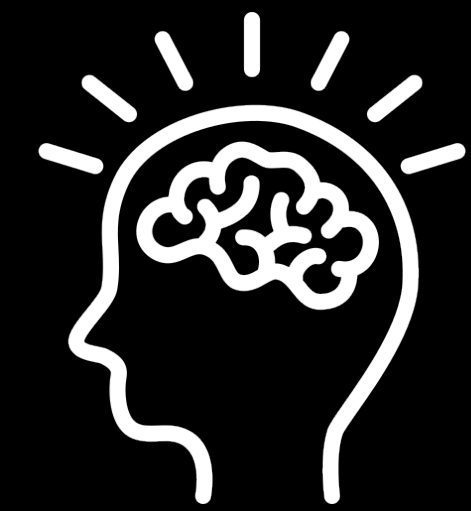
System Design



System Design



System Design





Training

Data Format

Data Format

User: Who is the current president of France?

Data Format

User: Who is the current president of France?

Reference: Emmanuel Macron has been served as the president of France since 2017 until now...

Data Format

User: Who is the current president of France?

Moshi: *< RET >* Based on what I know, the president of France is now Emmanuel Macron...

Reference: Emmanuel Macron has been served as the president of France since 2017 until now...

Data Format

User: Who is the current president of France?

Moshi: *< RET >* Based on what I know, the president of France is now Emmanuel Macron...

User: When was his second term election?

Reference: Emmanuel Macron has been served as the president of France since 2017 until now...

Data Format

User: Who is the current president of France?

Moshi: *< RET >* Based on what I know, the president of France is now Emmanuel Macron...

User: When was his second term election?

Reference: Emmanuel Macron has been served as the president of France since 2017 until now...

Reference: Macron was re-elected in 2022, becoming the first French president in 20 years to win a second term...

Data Format

User: Who is the current president of France?

Moshi: *< RET >* Based on what I know, the president of France is now Emmanuel Macron...

User: When was his second term election?

Moshi: *< RET >* That's a good question ... Macron's second term ...

Reference: Emmanuel Macron has been served as the president of France since 2017 until now...

Reference: Macron was re-elected in 2022, becoming the first French president in 20 years to win a second term...

Data Format

User: Who is the current president of France?

Moshi: *< RET >* Based on what I know, the president of France is now Emmanuel Macron...

User: When was his second term election?

Moshi: *< RET >* That's a good question ... Macron's second term ...

User: Is Macron a woman?

Reference: Emmanuel Macron has been served as the president of France since 2017 until now...

Reference: Macron was re-elected in 2022, becoming the first French president in 20 years to win a second term...

Data Format

User: Who is the current president of France?

Moshi: *< RET >* Based on what I know, the president of France is now Emmanuel Macron...

User: When was his second term election?

Moshi: *< RET >* That's a good question ... Macron's second term ...

User: Is Macron a woman?

Reference: Emmanuel Macron has been served as the president of France since 2017 until now...

Reference: Macron was re-elected in 2022, becoming the first French president in 20 years to win a second term...

No reference needed

Data Format

User: Who is the current president of France?

Moshi: *< RET >* Based on what I know, the president of France is now Emmanuel Macron...

User: When was his second term election?

Moshi: *< RET >* That's a good question ... Macron's second term ...

User: Is Macron a woman?

Moshi: No, he is a man!

Reference: Emmanuel Macron has been served as the president of France since 2017 until now...

Reference: Macron was re-elected in 2022, becoming the first French president in 20 years to win a second term...

No reference needed

Training Setup

Supervised fine-tuning

Training Setup

Supervised fine-tuning

- Initialized from Moshi checkpoint

Training Setup

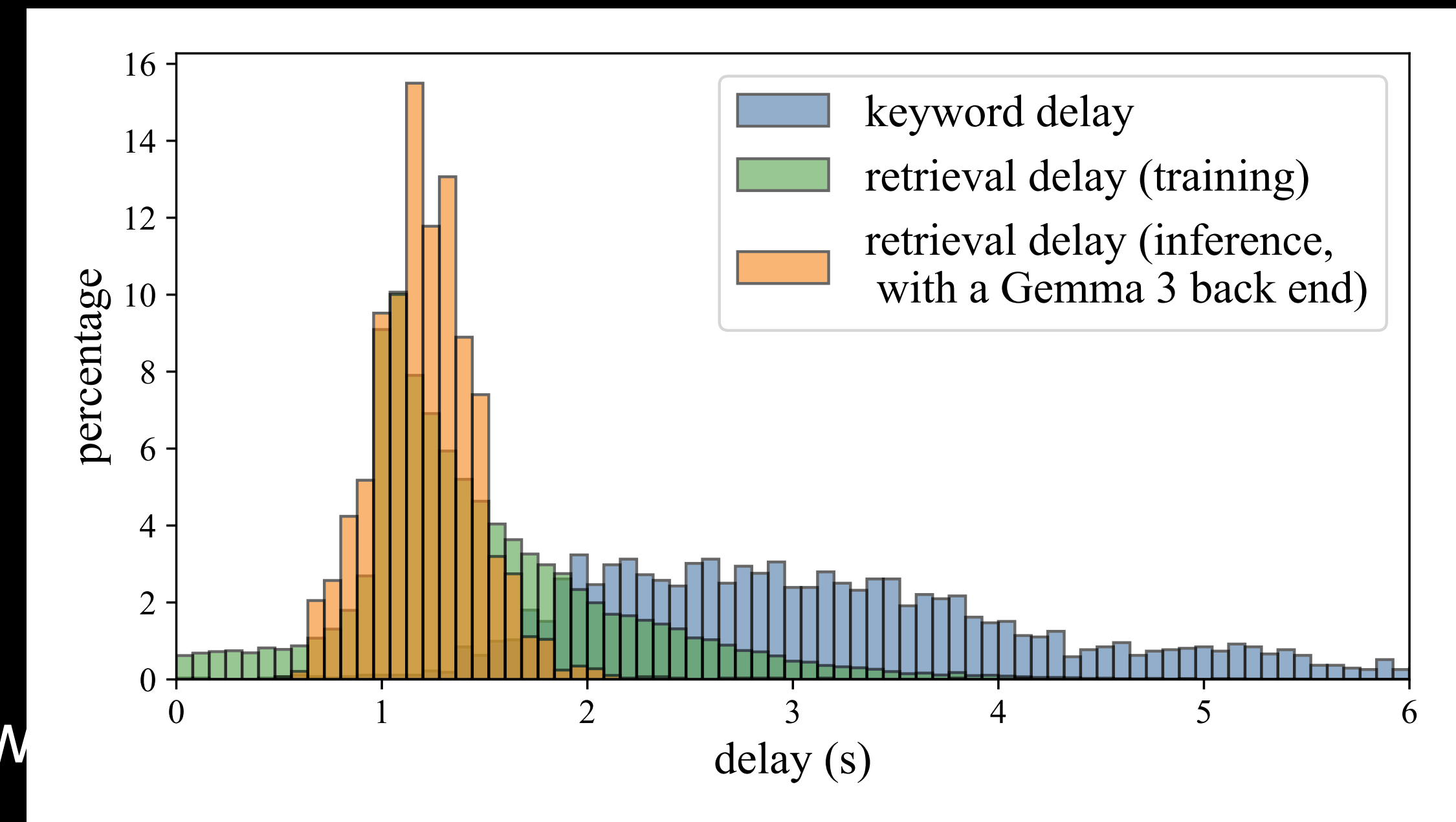
Supervised fine-tuning

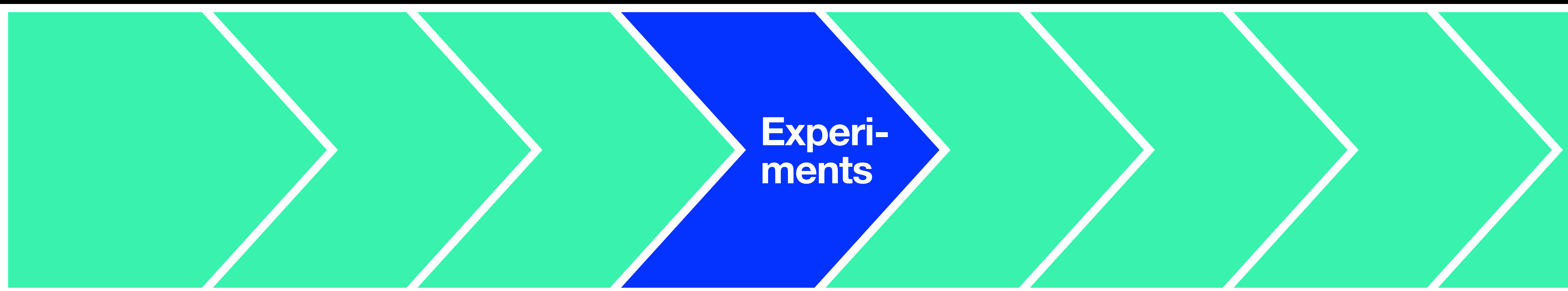
- Initialized from Moshi checkpoint
- Fine-tuned with synthetic spoken conversation with textual reference
 - 50k hours (2 minutes per conversation on average)
 - 100k steps (2 days) with 32 GPUs
- Moshi learns to extract useful information from text reference regardless of retrieval methods

Training Setup

Supervised fine-tuning

- Initialized from Moshi checkpoint
- Fine-tuned with synthetic spoken conversation w
 - 50k hours (2 minutes per conversation on average)
 - 100k steps (2 days) with 32 GPUs
 - Moshi learns to extract useful information from text reference regardless of retrieval methods
- Simulated retrieval delay
 - Shorter than keyword delays
 - Match the distribution of inference-time retrieval delay





**Experi-
ments**

Evaluation

Metrics

Evaluation

Metrics

- Factuality
 - Llama Questions, Web Questions, Trivia QA, HaluEval

Evaluation

Metrics

- Factuality
 - Llama Questions, Web Questions, Trivia QA, HaluEval
- Response delay
 - End-to-end keyword delay

Evaluation

Metrics

- Factuality
 - Llama Questions, Web Questions, Trivia QA, HaluEval
- Response delay
 - End-to-end keyword delay
- Interactiveness
 - Full-Duplex-Bench

Experiments

Performance on QA datasets

Experiments

Performance on QA datasets

Moshi
Qwen3-Omni
Kimi-Audio
MoshiRAG
GPT-4o

Experiments

Performance on QA datasets

Non-duplex	
	Moshi
	Qwen3-Omni
	Kimi-Audio
	MoshiRAG
	GPT-4o

Experiments

Performance on QA datasets

Non-duplex

	LlamaQ (%) ↑	WebQ (%) ↑	TriviaQA (%) ↑	HaluEval (%) ↑
Moshi	62.3	26.6	22.8	10.5
Qwen3-Omni	84.7	68.8	73.6	38.9
Kimi-Audio	79.3	70.2	62.1	43.2
MoshiRAG	80.3	67.2	78.2	51.3
GPT-4o	88.4	81.0	90.6	68.7



Experiments

Performance on QA datasets

Non-duplex

	LlamaQ (%) ↑	WebQ (%) ↑	TriviaQA (%) ↑	HaluEval (%) ↑	End-to-end keyword delay (sec.) ↓
Moshi	62.3	26.6	22.8	10.5	2.1
Qwen3-Omni	84.7	68.8	73.6	38.9	5.7
Kimi-Audio	79.3	70.2	62.1	43.2	3.5
MoshiRAG	80.3	67.2	78.2	51.3	3.1
GPT-4o	88.4	81.0	90.6	68.7	—



Experiments

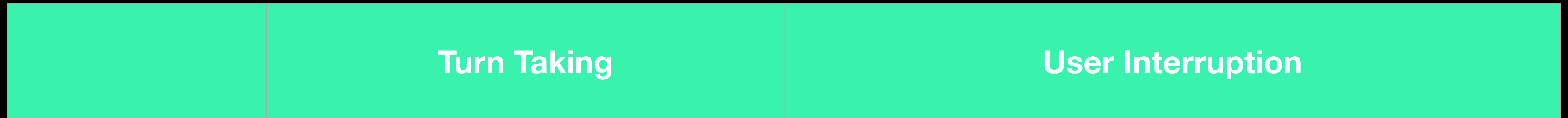
Interactiveness evaluation

Experiments

Interactiveness evaluation

Experiments

Interactiveness evaluation



Experiments

Interactiveness evaluation

	Turn Taking		User Interruption		
	Takeover % ↑	Latency (s) ↓	Takeover % ↑	GPT Score ↑	Latency (s) ↓

Experiments

Interactiveness evaluation

	Turn Taking		User Interruption		
	Takeover % ↑	Latency (s) ↓	Takeover % ↑	GPT Score ↑	Latency (s) ↓
Moshi	94.1	0.27	100.0	0.77	0.26

Experiments

Interactiveness evaluation

	Turn Taking		User Interruption		
	Takeover % ↑	Latency (s) ↓	Takeover % ↑	GPT Score ↑	Latency (s) ↓
Moshi	94.1	0.27	100.0	0.77	0.26
Moshi-RAG	83.2	0.18	85.0	3.75	1.02

Experiments

Interactiveness evaluation

	Turn Taking		User Interruption		
	Takeover % ↑	Latency (s) ↓	Takeover % ↑	GPT Score ↑	Latency (s) ↓
Moshi	94.1	0.27	100.0	0.77	0.26
Moshi-RAG	83.2	0.18	85.0	3.75	1.02
Gemini	65.5	1.30	89.1	3.38	1.18



Experiments

Interactiveness evaluation

Less aggressive turn-taking but improved response relevance

	Turn Taking		User Interruption		
	Takeover % ↑	Latency (s) ↓	Takeover % ↑	GPT Score ↑	Latency (s) ↓
Moshi	94.1	0.27	100.0	0.77	0.26
Moshi-RAG	83.2	0.18	85.0	3.75	1.02
Gemini	65.5	1.30	89.1	3.38	1.18



Experiments

Performance on QA datasets

Experiments

Performance on QA datasets

- Factuality
 - Comparable to SoTA non-duplex models

Experiments

Performance on QA datasets

- Factuality
 - Comparable to SoTA non-duplex models
- Latency
 - Better than all non-duplex models
 - Increased keyword delay than the original Moshi

Experiments

Performance on QA datasets

- Factuality
 - Comparable to SoTA non-duplex models
- Latency
 - Better than all non-duplex models
 - Increased keyword delay than the original Moshi
- Interactiveness
 - Less aggressive turn-taking but improved response relevance

More Interesting Results

Flexibility to switch to different retrieval back ends

More Interesting Results

Flexibility to switch to different retrieval back ends

Retrieval backend	LlamaQ (%) ↑	WebQ (%) ↑	TriviaQA (%) ↑	HaluEval (%) ↑
Gemma 27B	80.3	67.2	69.6	36.3
GPT 4.1	80.6	75.4	82.9	51.3
Tavily (search engine)	79.2	76.2	81.6	47.0



More Interesting Results

Generalization to unseen mathematical reasoning tasks

More Interesting Results

Generalization to unseen mathematical reasoning tasks

	AddSub (%)	MultiArith (%)	SinglEq (%)	SVAMP (%)	GSM8K (%)

More Interesting Results

Generalization to unseen mathematical reasoning tasks

	AddSub (%)	MultiArith (%)	SinglEq (%)	SVAMP (%)	GSM8K (%)
MoshiRAG	61.7	69.0	68.2	55.0	33.8
GLM-4-Voice	59.4	62.0	71.0	44.0	29.0

More Interesting Results

Generalization to unseen mathematical reasoning tasks

	AddSub (%)	MultiArith (%)	SinglEq (%)	SVAMP (%)	GSM8K (%)
MoshiRAG	61.7	69.0	68.2	55.0	33.8
GLM-4-Voice	59.4	62.0	71.0	44.0	29.0
STITCH-S (fine-tuned on math)	81.7	87.9	91.7	72.3	56.7



Ablation Studies

Sensitivity to ASR errors

Ablation Studies

Sensitivity to ASR errors

Ablation Studies

Sensitivity to ASR errors

(%)	Llama Questions	Web Questinos	Trivia QA	Halu Eval
-----	-----------------	---------------	-----------	-----------

Ablation Studies

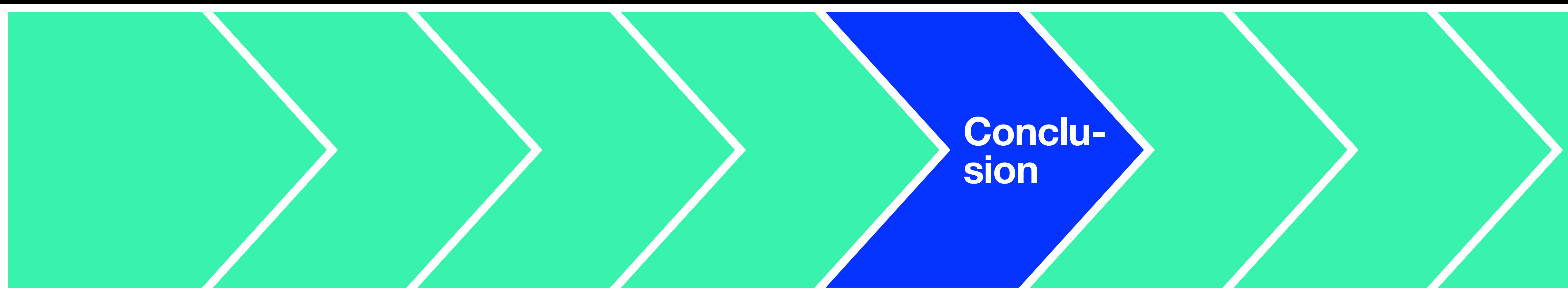
Sensitivity to ASR errors

(%)	Llama Questions	Web Questinos	Trivia QA	Halu Eval
MoshiRAG	80.3	74.7	73.2	36.3

Ablation Studies

Sensitivity to ASR errors

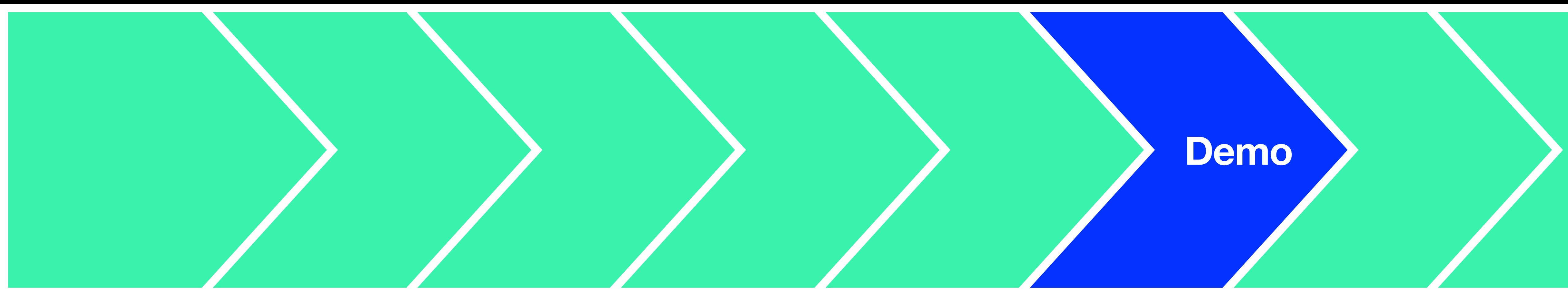
(%)	Llama Questions	Web Questinos	Trivia QA	Halu Eval
MoshiRAG	80.3	74.7	73.2	36.3
MoshiRAG (ground-truth ASR transcript)	80.2	78.3	82.5	50.8



**Conclu-
sion**

Conclusion

- First fully-duplex model with RAG capability
- Asynchronous retrieval without interrupting the conversation
- **Interactivity and latency** comparable to SoTA SLMs
- **Factuality** (QA benchmarks) comparable to non-duplex models
- **Plug-and-play** support for various retrieval sources — boosting performance without re-training the model
- **Generalizability** to unseen domain by using LLM as a tool (math reasoning benchmarks)



Demo



Hide Retrieval

RETRIEVAL BACK END



System Idle

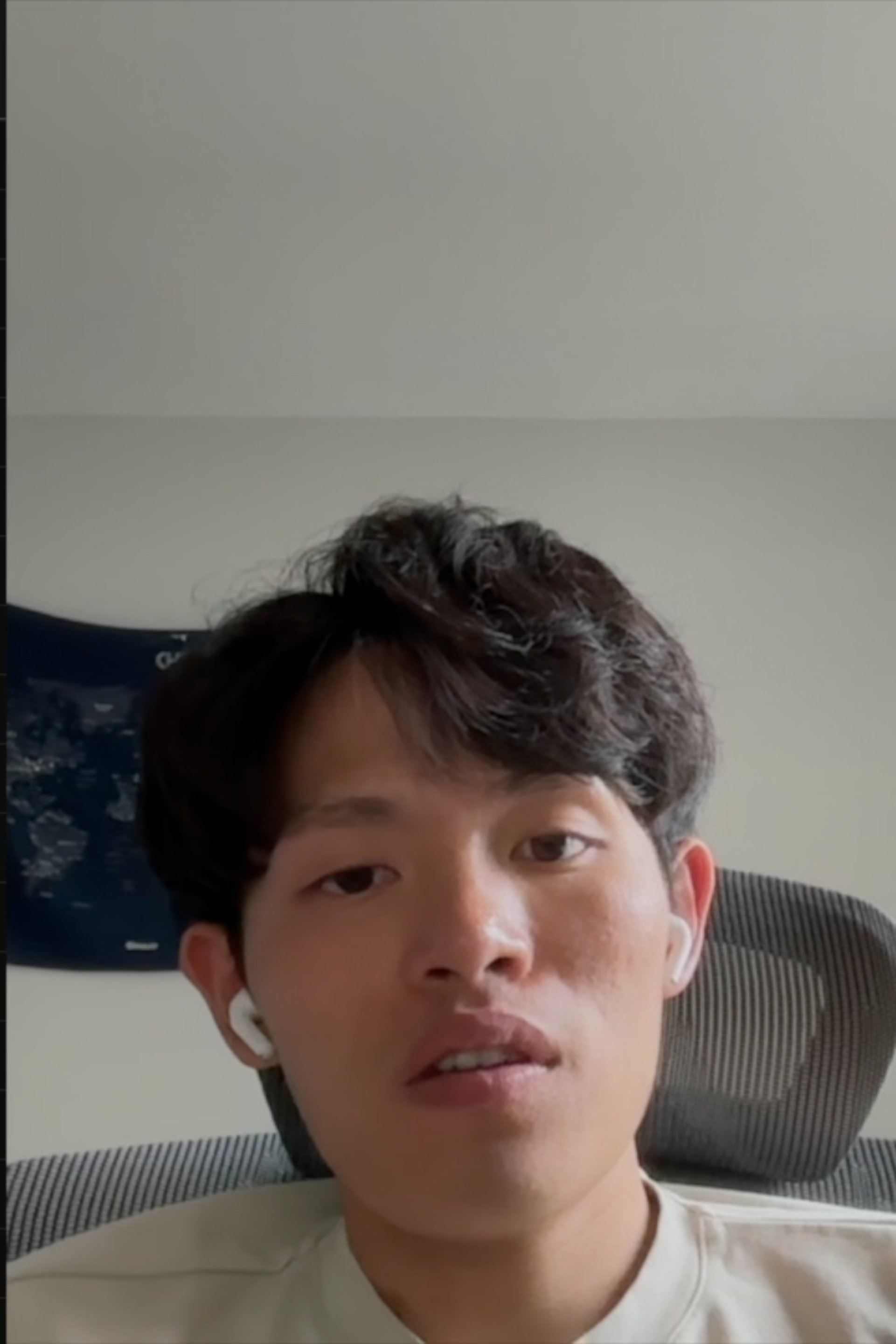
LIVE TRANSCRIPT

NOTICE

MOSHI

This is moshi. Ready when you are!

YOU





Hide Retrieval

RETRIEVAL BACK END



System Idle

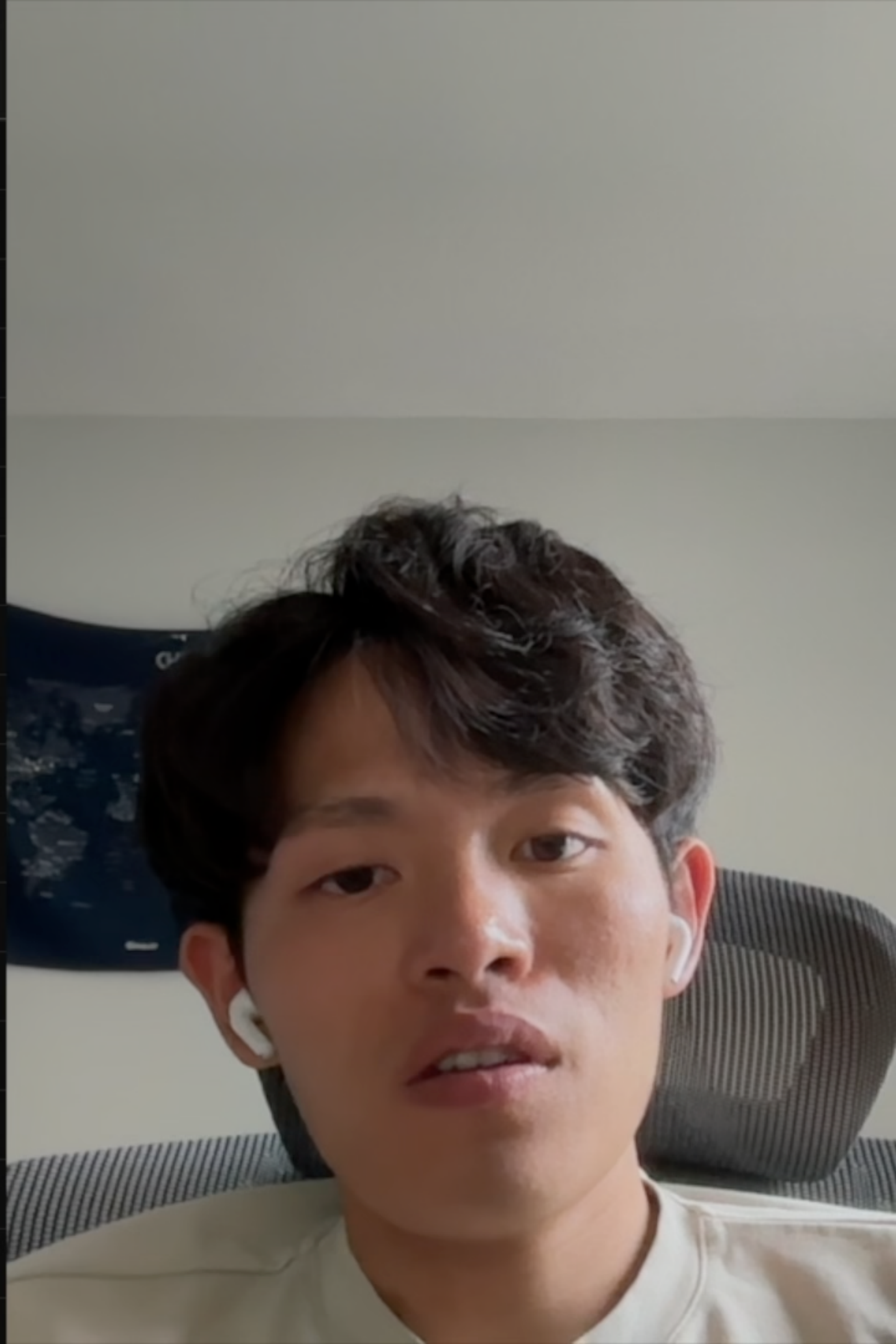
LIVE TRANSCRIPT

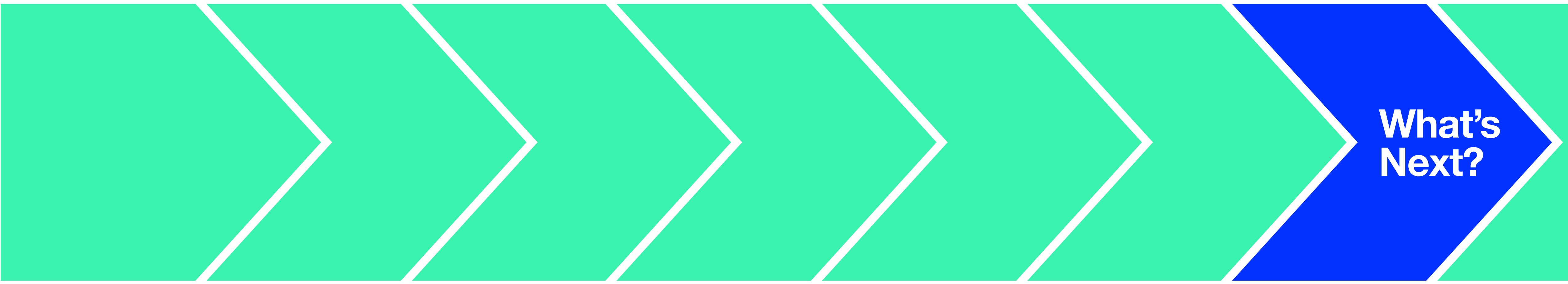
NOTICE

MOSHI

This is moshi. Ready when you are!

YOU





**What's
Next?**

What We Could Have Done Better?

What We Could Have Done Better?

- Training data is not general enough (too biased towards QA)

What We Could Have Done Better?

- Training data is not general enough (too biased towards QA)
- Robustness against
 - ASR error
 - Backend delay

What We Could Have Done Better?

- Training data is not general enough (too biased towards QA)
- Robustness against
 - ASR error
 - Backend delay
- Rethinking the timing of retrieval
 - Filler-words are not elegant — try to explain what's happening to the user, etc
 - Retrieval earlier before the end of the user input

What We Could Have Done Better?

What We Could Have Done Better?

- More general tool-use aside from retrieval
 - Different tool calls take different amount of time
 - Multiple communications between front-end and back-end are necessary for long-horizon tool calls or agentic workflows

What We Could Have Done Better?

- More general tool-use aside from retrieval
 - Different tool calls take different amount of time
 - Multiple communications between front-end and back-end are necessary for long-horizon tool calls or agentic workflows
- Smarter tool call
 - “To know what you know, and to know what you don't know” is hard
 - Use RL to ensure that the retrieval decision is associated with answer correctness
 - Use a separate model for the tool call decision?

What We Could Have Done Better?

What We Could Have Done Better?

- How to integrate retrieved information into Moshi?
 - Directly inserting text can probably improve latency and information accuracy, with the price of longer sequence and specific handling to enable batching

What We Could Have Done Better?

- How to integrate retrieved information into Moshi?
 - Directly inserting text can probably improve latency and information accuracy, with the price of longer sequence and specific handling to enable batching
- Channel fusion or interleaving for speech & text tokens?
 - Personally I think interleaving in the MiniCPM-o 4.5 style is a good practice — separating text and speech could make our life easier when debugging and probably results in better knowledge transfer

Questions?