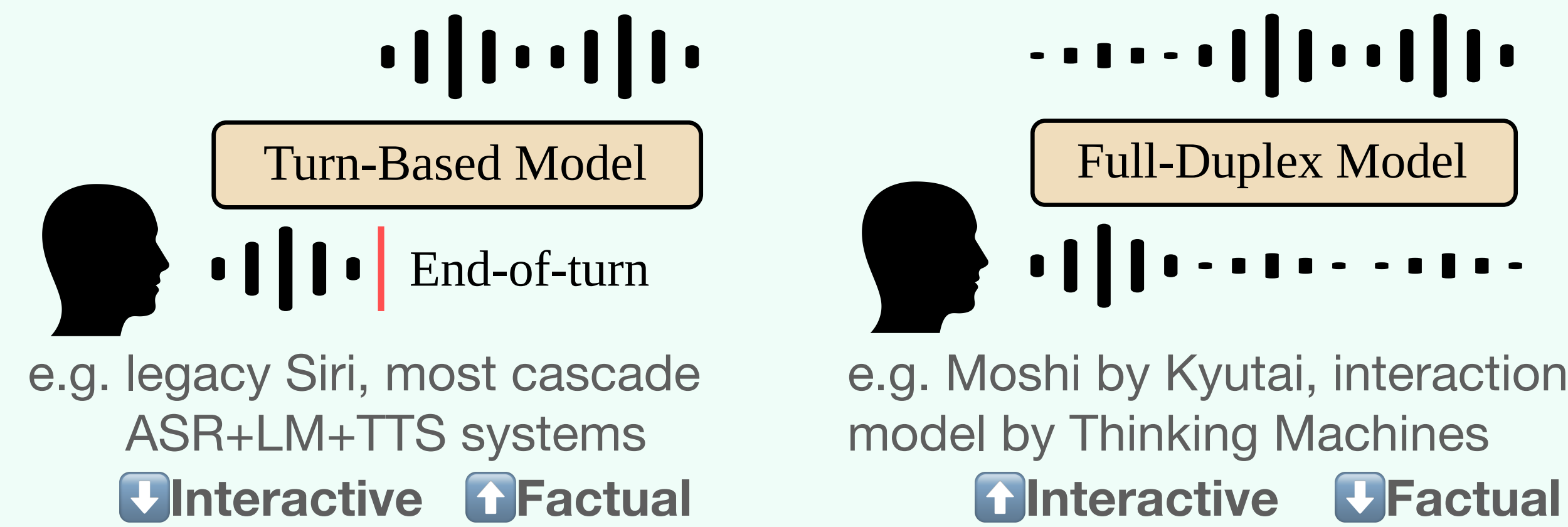
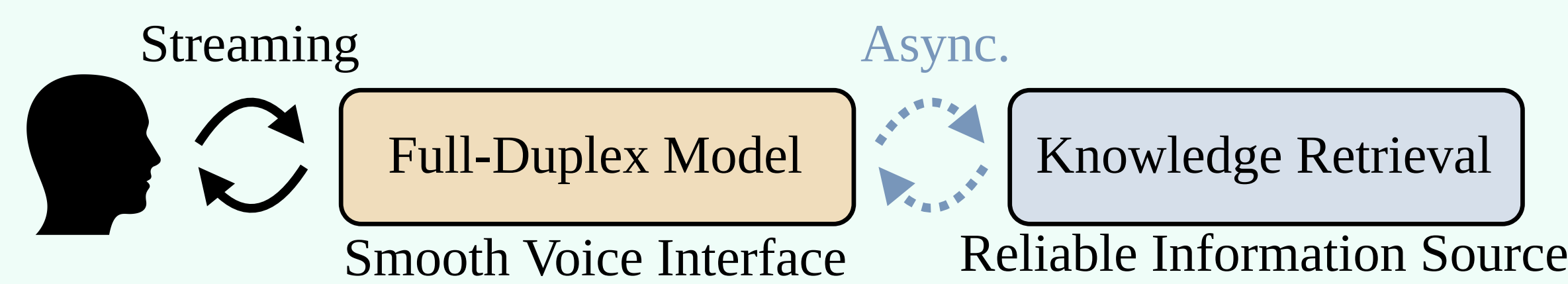


Highlights

- Turn-based vs. full-duplex models



- This work: combines a **full-duplex** speech language model with **asynchronous knowledge retrieval**



- Interactivity & latency** comparable to SoTA speech language models
- Factuality** comparable to non-duplex models
- Model, code, and dataset open-sourced

Training

- Initialized from Moshi checkpoint
- Fine-tuned with 50k hours of synthetic spoken conversation with textual reference
 - Moshi learns to extract useful information from text reference regardless of retrieval methods

...

User: Where is Emily Cooper from?

Moshi (lead): In the Netflix series "Emily in Paris,"

Reference: Emily Cooper, the protagonist of the Netflix series Emily in Paris, is from the Chicago area in the United States. She is a marketing executive who moves to Paris for a job opportunity, as described in the Netflix article.

Moshi (body): Emily Cooper was originally from Chicago and then relocated to Paris for a job opportunity.

...

Example LLM-generated conversation script

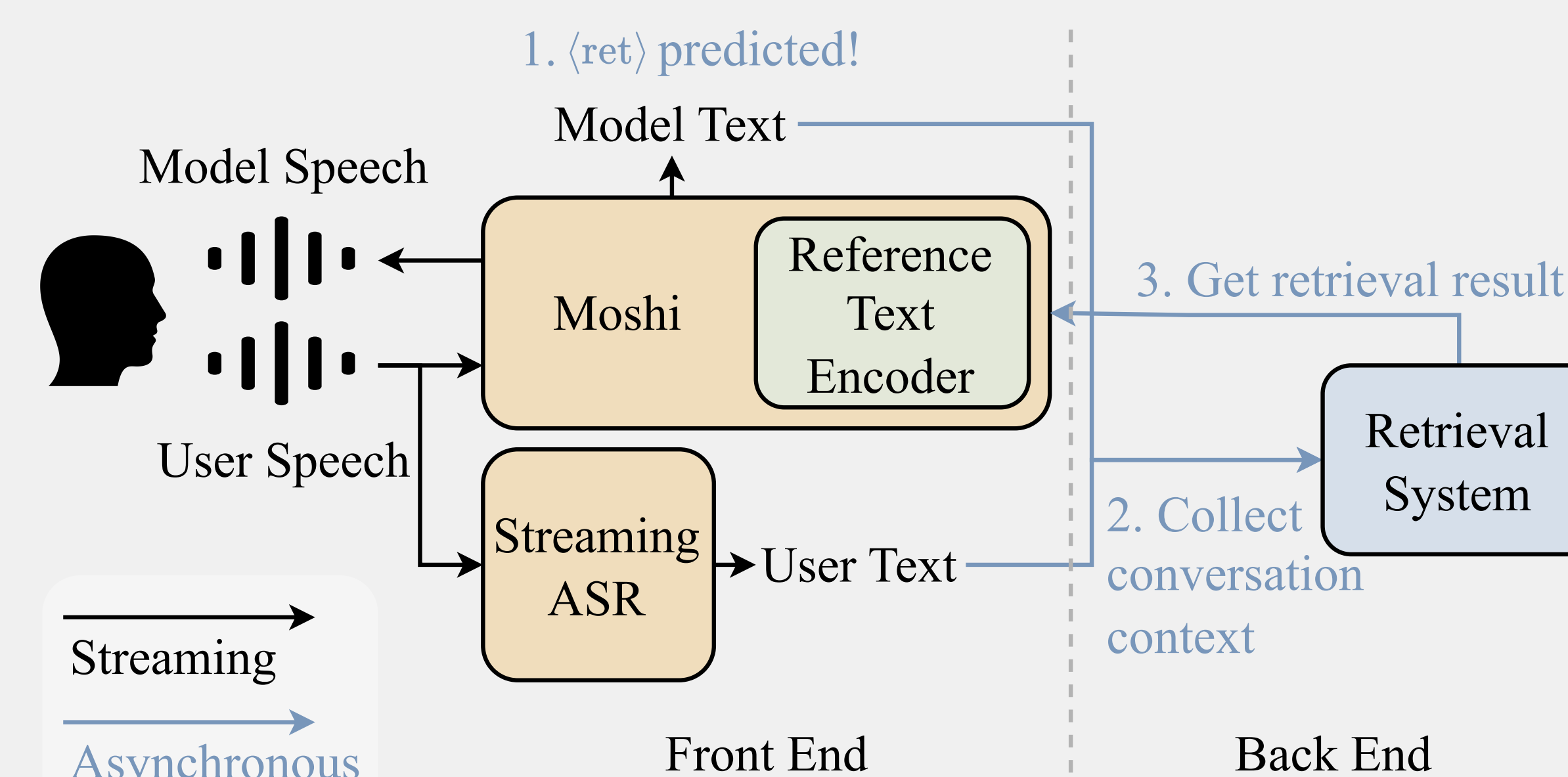
System Design



Observation: the core answer rarely appears in the first few words in a natural conversation

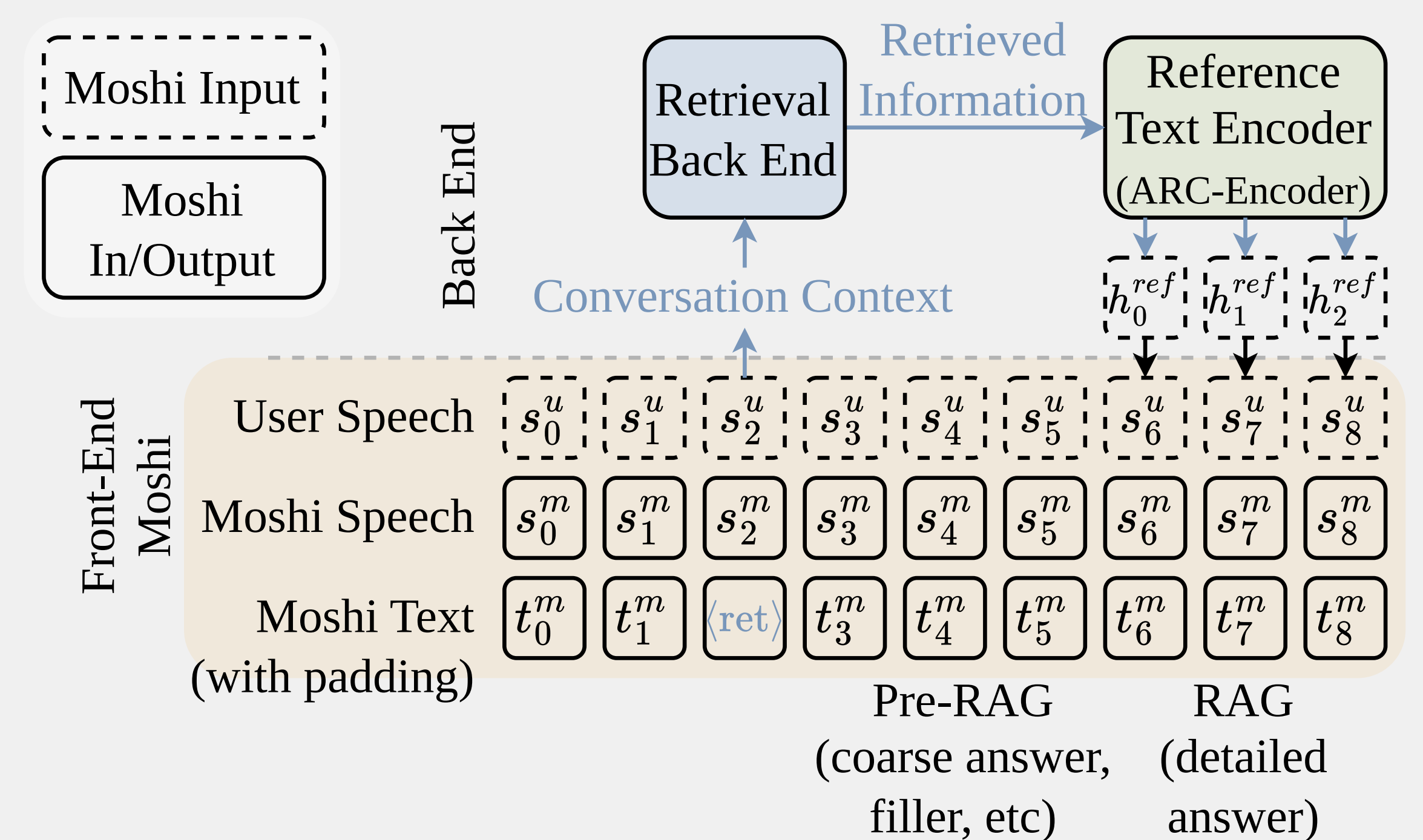
→ Use of the gap to perform retrieval as long as

$$\text{Retrieval Delay} \leq \text{E2EKD}$$



Streaming Inputs & Outputs

- Real-time duplex audio streaming at 12.5Hz
- Reference embedding added to input embedding once ready



Results

Factuality QA Benchmarks

Model	LlamaQ (%)↑	WebQ (%)↑	TriviaQA (%)↑	HaluEval (%)↑	E2EKD (sec.)↓
Vanilla Moshi	62.3	26.6	22.8	10.5	2.1
GLM-4-Voice	64.7	32.2	39.1	21.2	4.4
Kimi-Audio	79.3	70.2	62.1	43.2	3.5
Qwen 3 Omni	84.7	68.8	73.6	38.9	5.7
MoshiRAG (GPT-4o Back End)	<u>80.6</u>	<u>68.9</u>	78.2	51.3	<u>3.1</u>
MoshiRAG (Search Engine)	78.2	66.1	<u>77.5</u>	<u>47.0</u>	

- Factuality & latency:** comparable to SoTA non-duplex models
- Plug-and-play** support for various retrieval sources
- Generalizability** to unseen mathematical reasoning tasks

Mathematical Reasoning Benchmarks

Model	AddSub (%)↑	MultiArit h (%)↑	SingleEq (%)↑	SVAMP (%)↑	GSM8K (%)↑
Vanilla Moshi	8.3	9.8	18.4	9.7	2.1
GLM-4-Voice	59.4	62.0	71.0	4.0	29.0
STITCH (Fine-tuned on math)	81.7	87.9	91.7	72.2	56.7
MoshiRAG (GPT-4o back end)	<u>64.8</u>	<u>76.0</u>	<u>72.9</u>	<u>61.1</u>	<u>43.2</u>

Limitations & Next Steps

- Robustness against ASR error and information loss during injection
- Diverse retrieval sources & tools
- RL-based tool-use triggering



Paper



Blog Post