Chronological Thinking in Full-Duplex Spoken Dialogue Language Models

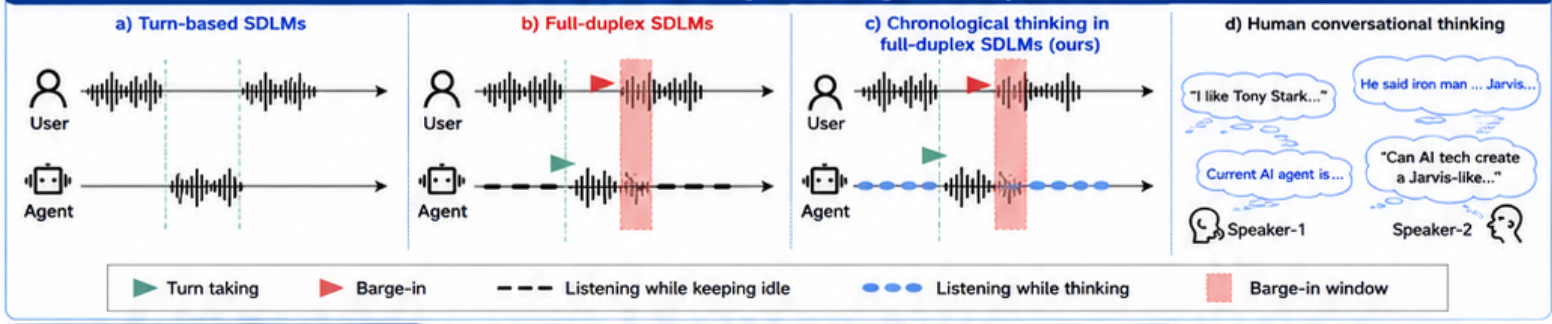
Donghang Wu[1,2,*], Haoyang Zhang[1,2,*], Chen Chen[1,†], Tianyu Zhang[3], Fei Tian[2], Xuerui Yang[2], Gang Yu[2], Hexin Liu[1], Nana Hou[1], Yuchen Hu[1], Eng Siong Chng[1,†]

[1] Nanyang Technological University, [2] StepFun, [3] Mila

[*] Equal contribution [†] Corresponding authors

arXiv:2510.05150

From Turn-Based to Chronological Thinking in Full-Duplex SDLMs



1. Motivation / Problem

- Full-duplex SDLMs continuously listen while speaking, enabling natural interaction and barge-in handling.
- Existing systems keep the agent idle during listening by repeatedly predicting silence tokens.
- This wastes the listening window and can bias autoregressive generation.
- Humans instead perform lightweight thinking while listening.

Research question: can a full-duplex SDLM think on the fly while listening, under strict streaming and latency constraints?

2. Key Idea / Contributions

1. Introduce chronological thinking for full-duplex SDLMs: a strictly causal, on-the-fly reasoning mechanism while listening.
2. Replace redundant silence tokens with compact structured thinking-chain tokens, enabling preemptible reasoning with no additional latency.
3. Show consistent gains in response quality and strong full-duplex interaction behavior.

4. Data Generation & Training

Synthetic dialogues generated with an LLM and converted to speech with multi-speaker TTS.

Speech synthesized with Step-Audio-TTS-3B.

Chronological thinking chains generated with Qwen2.5-72B-Instruct.

Datasets & Hours	
GenConv	10.5k hours
SpokenWOZ-G	2.0k hours
Llamaq-G	2.7k hours

- Model training with NeMo Toolkit.
- Backbone initialized from Qwen2.5-1.5B-Instruct.
- Trained on 8 L40s (48G) GPUs.

6. Example Chronological Thinking Chain

Help me order a restaurant to celebrate my birthday this weekend.

Thinking Chain (listening phase)

{Help me}{INTENT} Request assistance

{order a restaurant}{ACTION} Initiate booking process

{to celebrate my birthday}{LOGIC} Purpose: birthday celebration

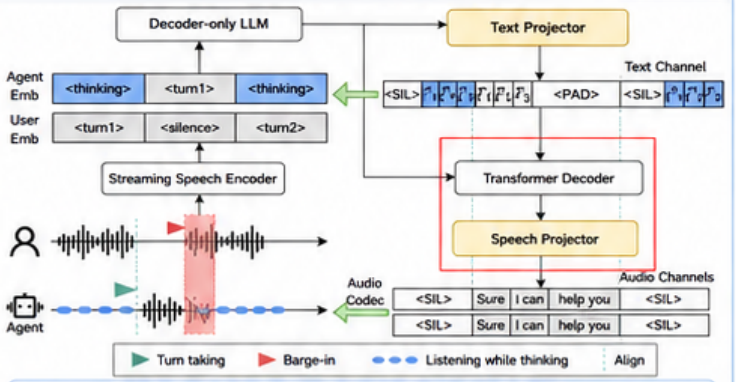
{this weekend}{ENTITY} Timeframe: weekend

{KNOWLEDGE} Birthday: decorations, discounts, or special perks

Agent Response:

Certainly! For a birthday celebration, I recommend a place that offers a special dining experience. Do you have a preferred cuisine or location in mind?

3. Method: CT-Duplex Architecture



- User speech is encoded by a streaming speech encoder at 12.5 Hz.
- A decoder-only LLM backbone predicts agent text tokens.
- An autoregressive Transformer decoder predicts audio tokens.
- Chronological thinking tokens are inserted during the listening phase instead of repeated <SIL> tokens, enabling **preemptible reasoning without extra latency.**
- Inspired by the Adaptive Control of Thought–Rational (ACT-R) cognitive architecture.

Node Types in Thinking Chain

Entity	Extracts entities from dialogue
Intent	Represents the user's goal
Action	Denotes the agent's executable operation
Knowledge	Retrieves factual or procedural knowledge
Logic	Captures rules or logical reasoning

Strictly causal

No additional latency

5. Results

A. Reasoning Quality

SpokenWOZ

Method	GPT score	BLEU	Sentence-BERT
GT-LM	2.48	7.60	0.55
SALM-Duplex	2.11	8.73	0.34
CT-Duplex w/o *thk*	2.40	12.92	0.52
CT-Duplex w/ *thk*	2.61	16.30	0.59

MtBenchEval

Method	GPT score	BLEU	Sentence-BERT
GT-LM	3.15	10.18	0.73
SALM-Duplex	2.25	5.31	0.47
CT-Duplex w/o *thk*	2.39	7.00	0.64
CT-Duplex w/ *thk*	2.44	7.34	0.67

+8.75% improvement on SpokenWOZ with chronological thinking.

B. Factual Knowledge Capability (Accuracy %)

Method	Llama Questions	Web Questions
TWIST	4.0	1.5
SpeechGPT	21.6	6.5
Spectron	21.9	6.1
Moshi	21.0	9.2
GLM-4-Voice	50.7	15.9
SALM-Duplex*	15.0	6.7
CT-Duplex w/o *thk*	30.4	13.2
CT-Duplex w/ *thk*	31.4	13.3

Chronological thinking mainly helps reasoning-intensive tasks; factual QA shows only marginal gain.

C. Full-Duplex Interaction Metrics (Impatient dataset)

Method	Turn-taking Latency (↓)	Barge-in Latency (↓)	Barge-in Success rate (↑)
Freeze-Omni	1.17	1.20	79.50%
dGSLM	0.57	0.86	85.00%
Moshi	n.a.	0.81	55.10%
ORISE	0.43	0.61	96.80%
SALM-Duplex*	0.92	0.69	87.50%
CT-Duplex w/o *thk*	0.45	0.53	88.63%
CT-Duplex w/ *thk*	0.68	0.54	94.05%

Chronological thinking does not harm turn-taking or barge-in performance.

D. Subjective A/B Test (CT-Duplex w/ *thk* vs. w/o thk)

Human evaluators prefer the model with chronological thinking in both audio fidelity and content quality.

7. Conclusion / Takeaways

- Chronological thinking gives full-duplex SDLMs a human-like thinking-while-listening ability.
- It is causal, compact, and preemptible.
- It improves response quality, especially on reasoning-heavy tasks, without adding latency.
- It robustly supports conversational dynamics such as turn-taking and user barge-in.