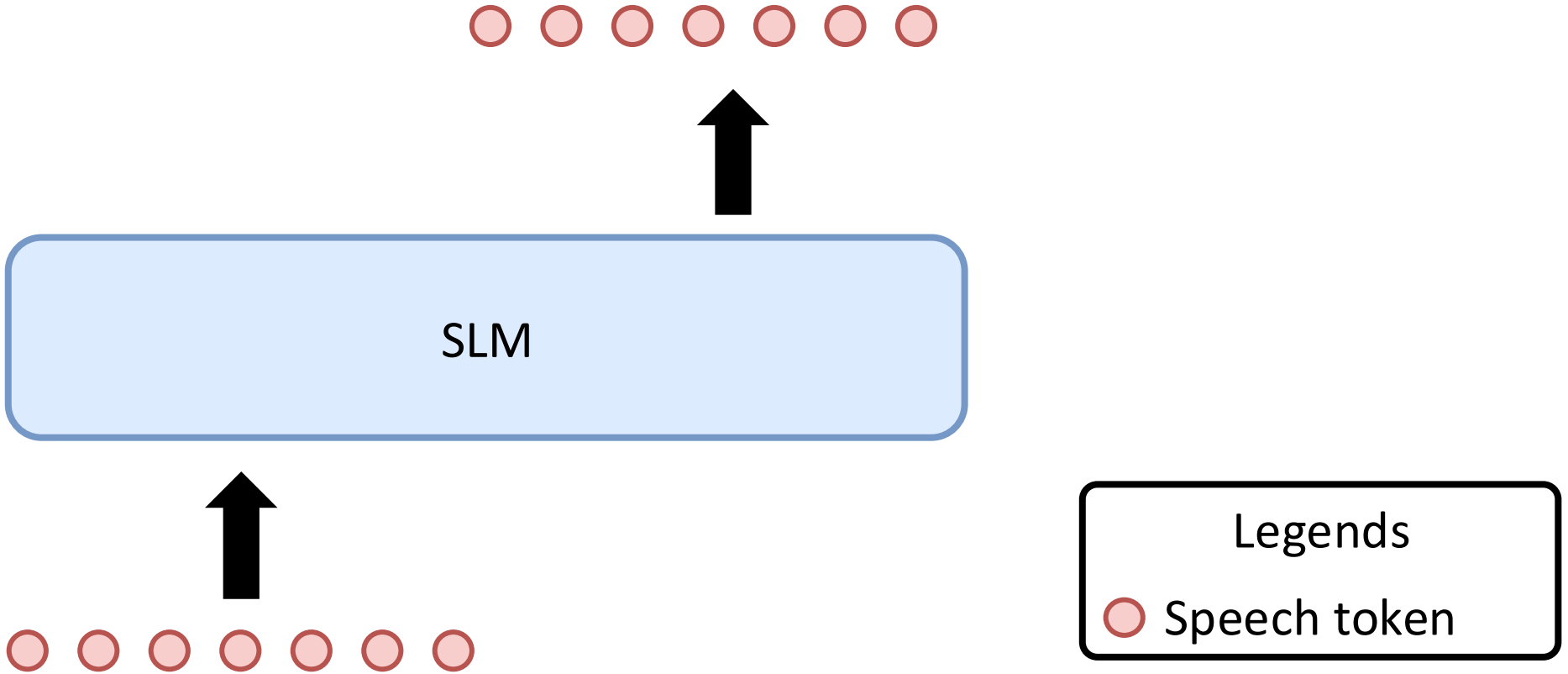
STITCH: Simultaneous Thinking and Talking with Chunked Reasoning for Spoken Language Models

Cheng-Han Chiang[1,2], Xiaofei Wang[2], Linjie Li[2], Chung-Ching Lin[2], Kevin Lin[2], Shujie Liu[2], Zhendong Wang[2], Zhengyuan Yang[2], Hung-yi Lee[1], Lijuan Wang[2]

[1]National Taiwan University
[2]Microsoft

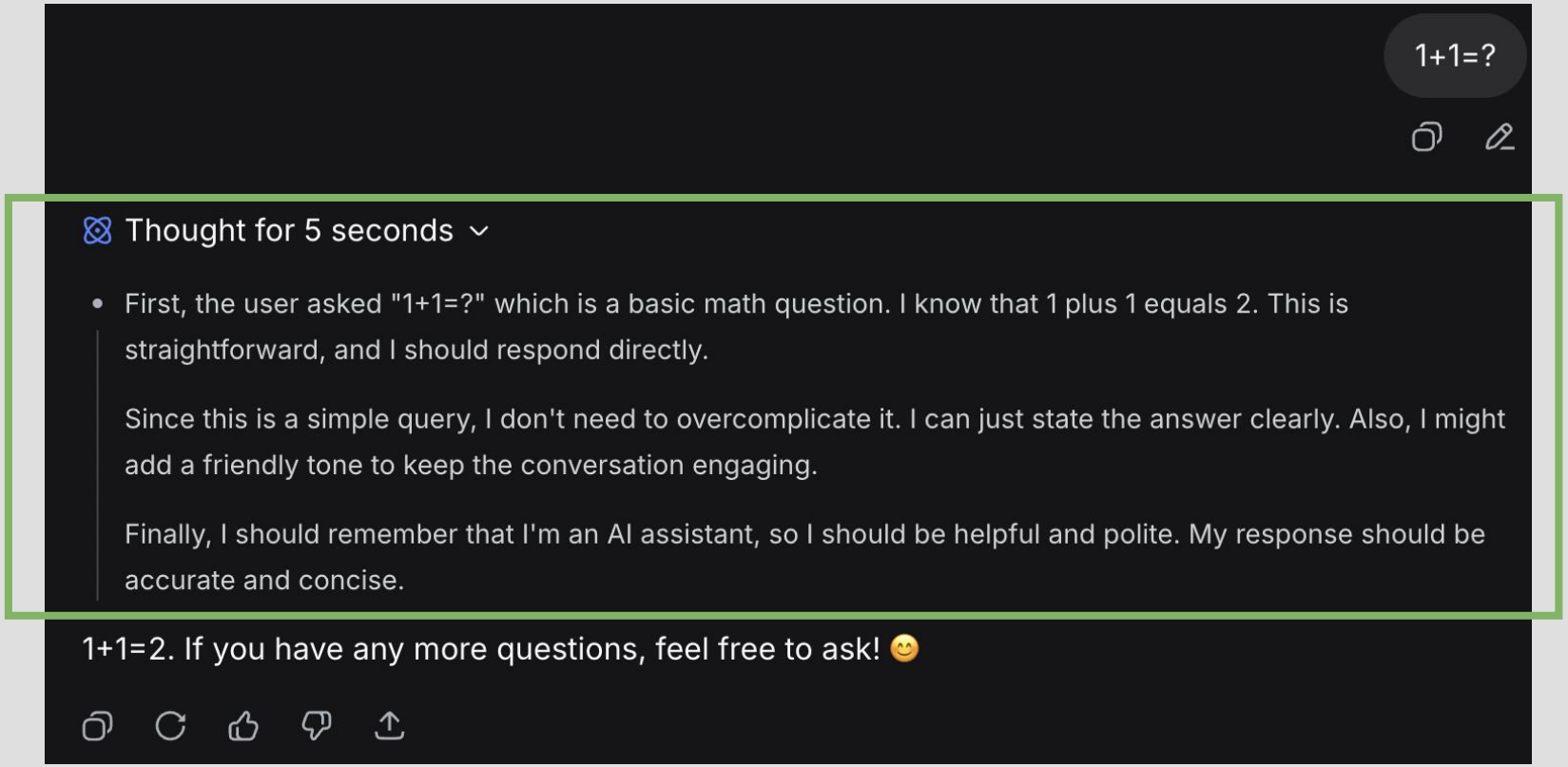
Motivation: Reasoning for SLMs

- SLMs generates speech tokens without any intermediate reasoning
- This is different from “reasoning LLMs”



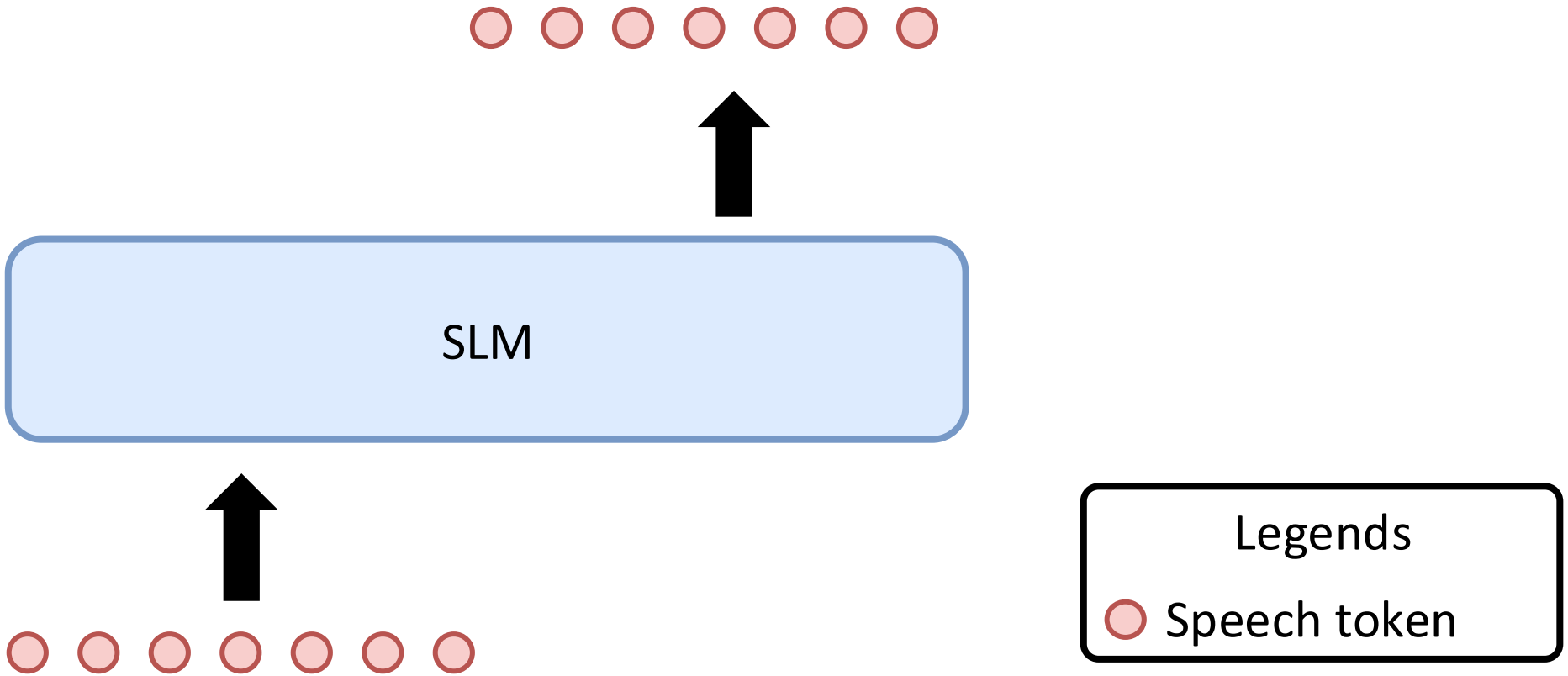
Motivation: Reasoning for SLMs

- Reasoning LLMs (e.g., Deepseek-r1, GPT o-series models) can generate reasoning before emitting the final response



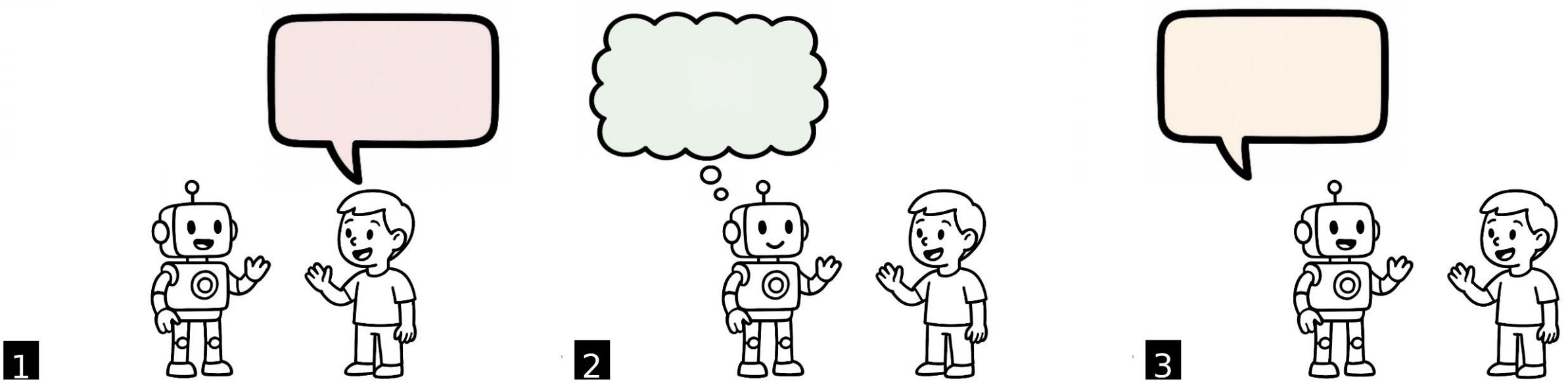
Motivation: Reasoning for SLMs

- SLMs generates speech tokens without any intermediate reasoning
- How can we make the SLMs reason to improve the final response?



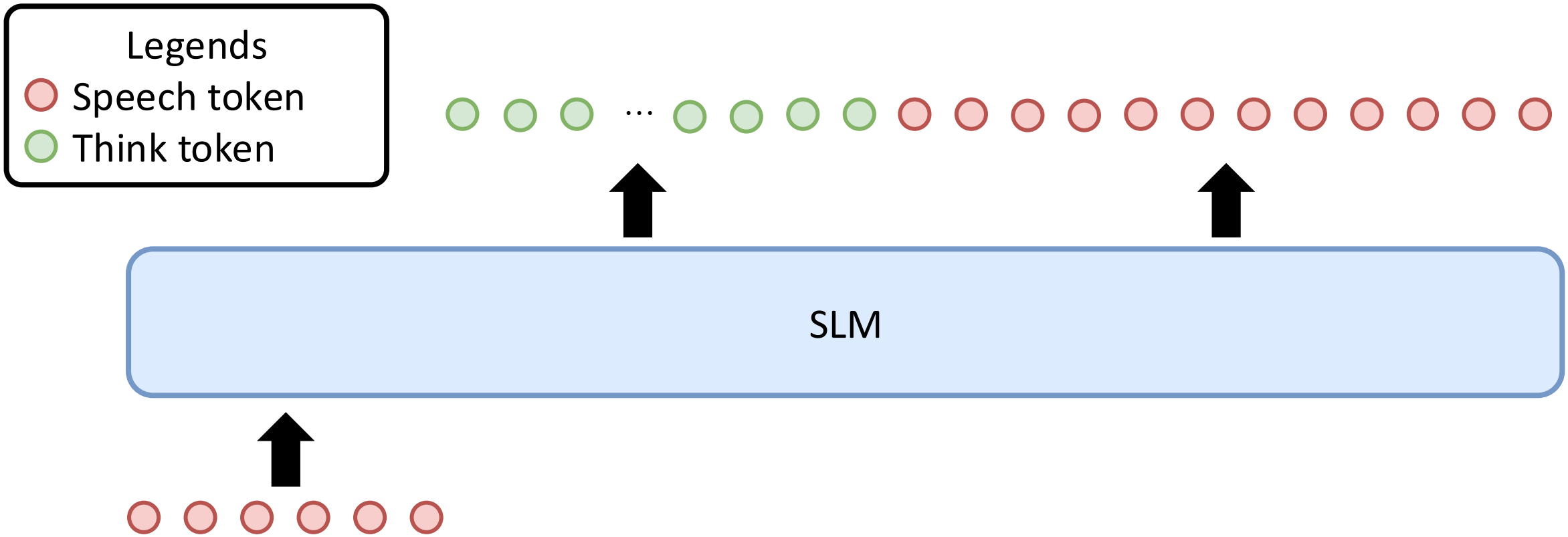
Naïve Baseline: Listen → Think → Speak

- This directly borrows the idea from reasoning LLM: receive the user input, think silently (hidden from the user)



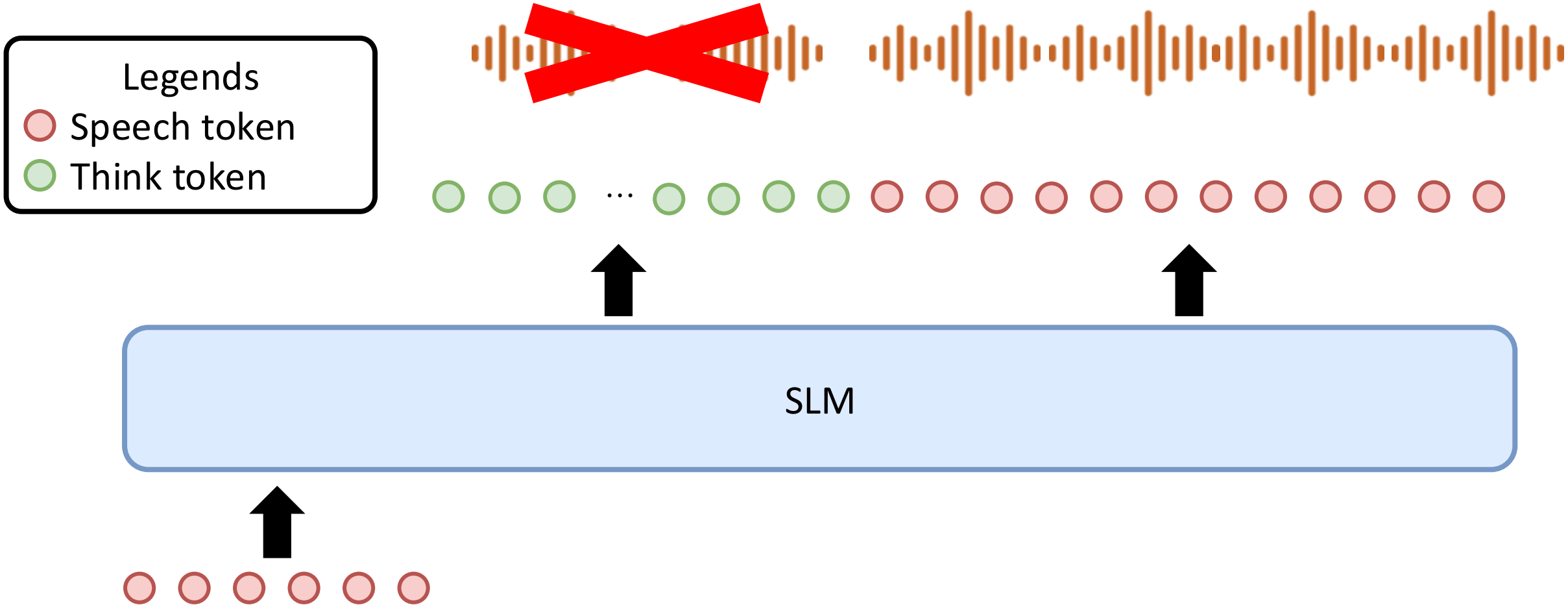
Thinking before Speaking

- Generate some think tokens before generating the speech tokens
  - The think tokens are **text tokens**; they are called think tokens because they are used for thinking



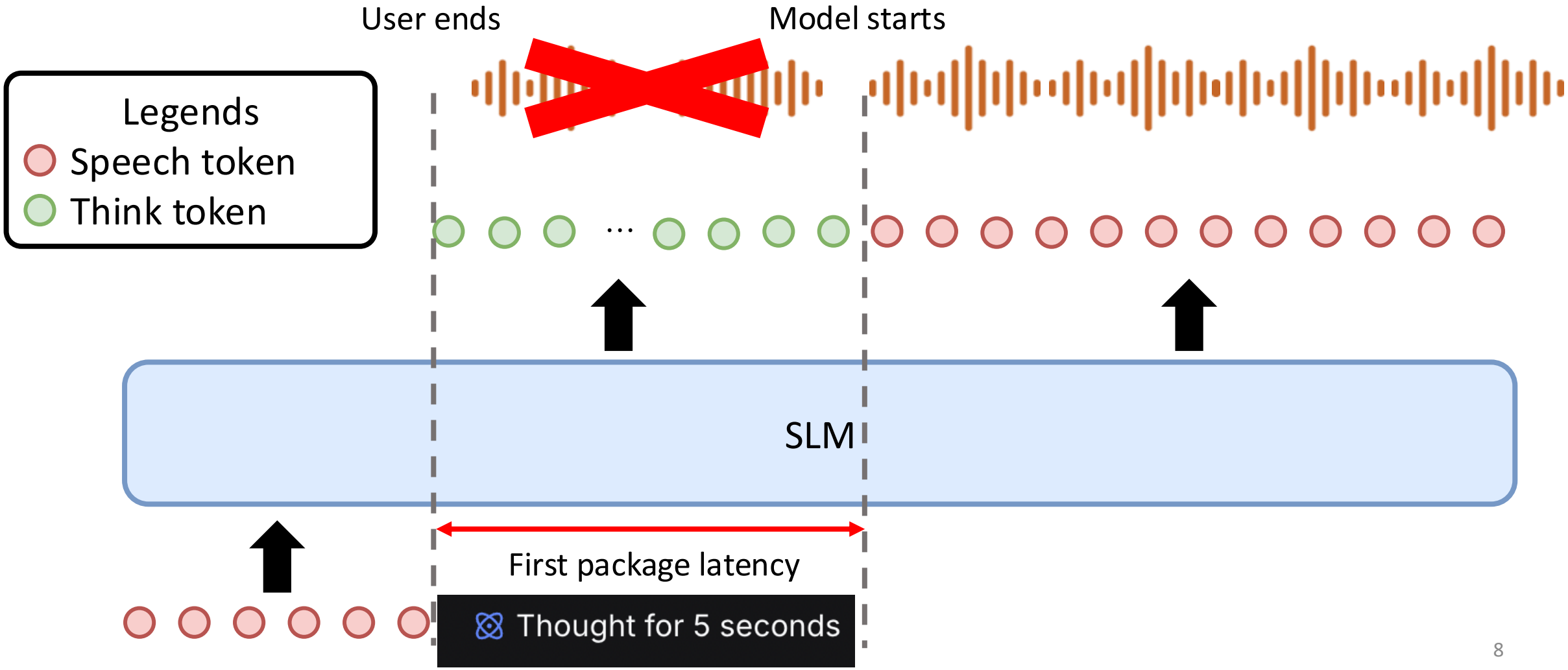
Thinking before Speaking

- Thinking tokens will not be spoken



Thinking before Speaking

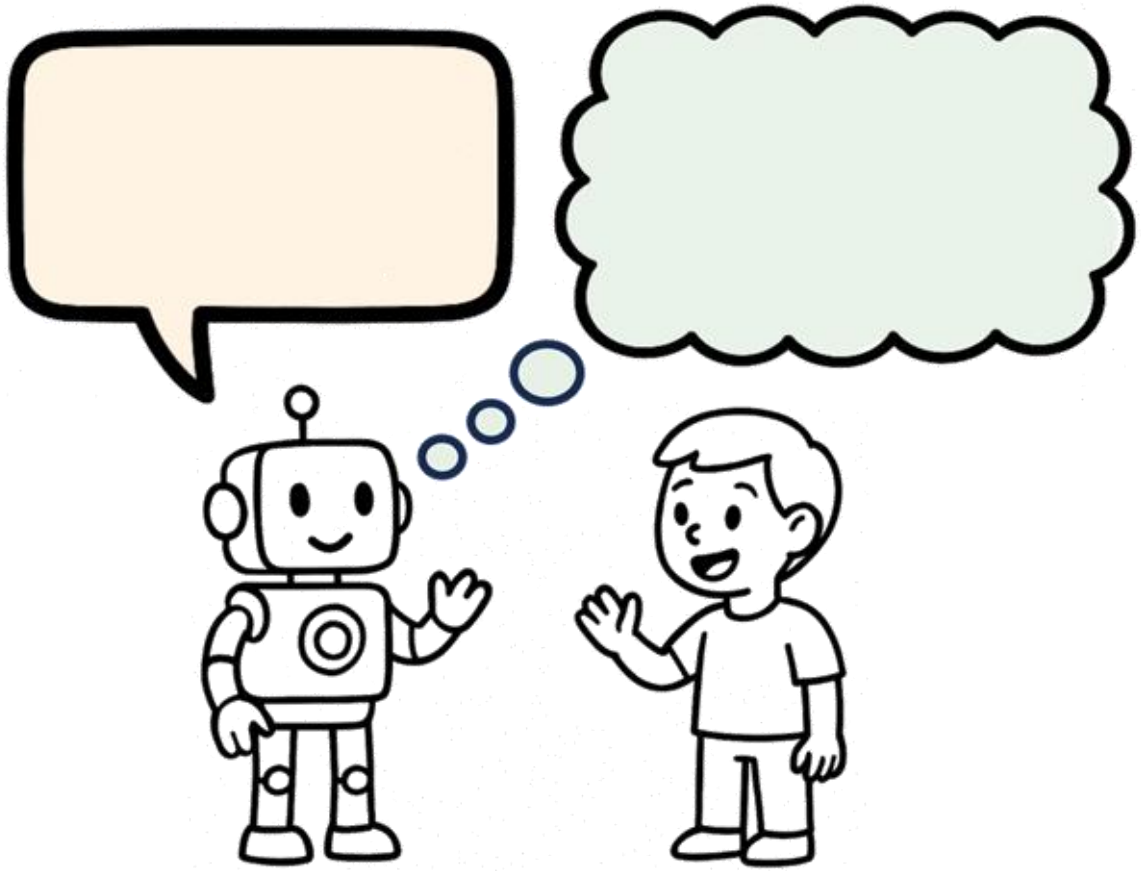
- Thinking before speaking creates a very high latency



“When should SLM think?”

STITCH: Listen → Think + Speak

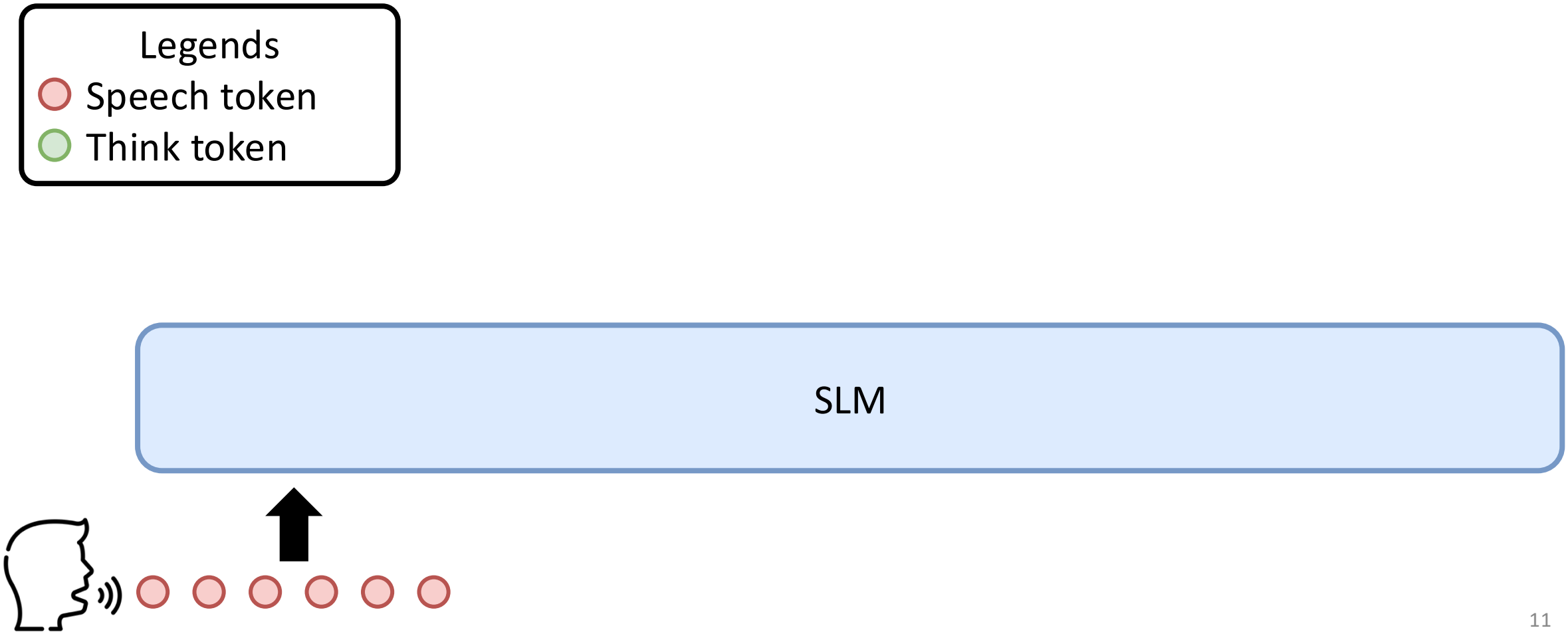
- Humans can think while speaking



STITCH: Simultaneous talking and thinking

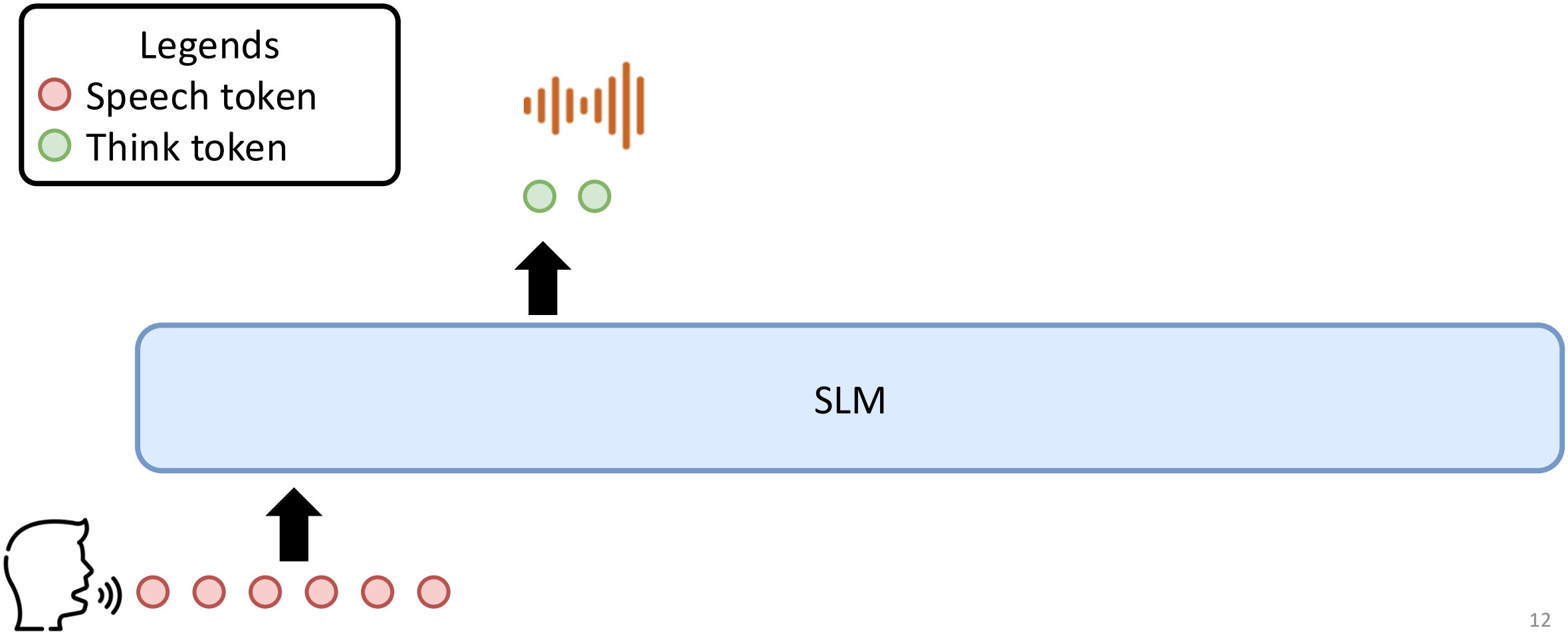
Goal: Thinking while Speaking

- We want to make SLMs think while speaking



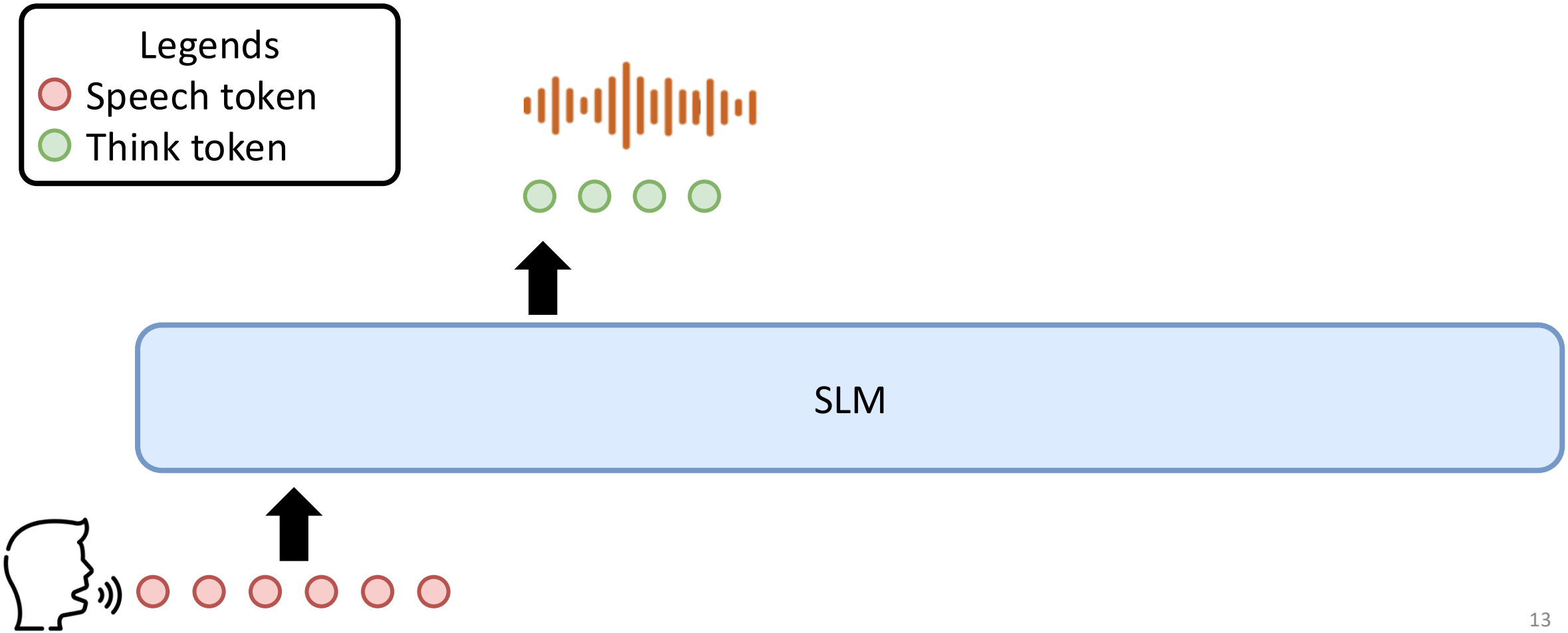
Goal: Thinking while Speaking

- We want to make SLMs think while speaking



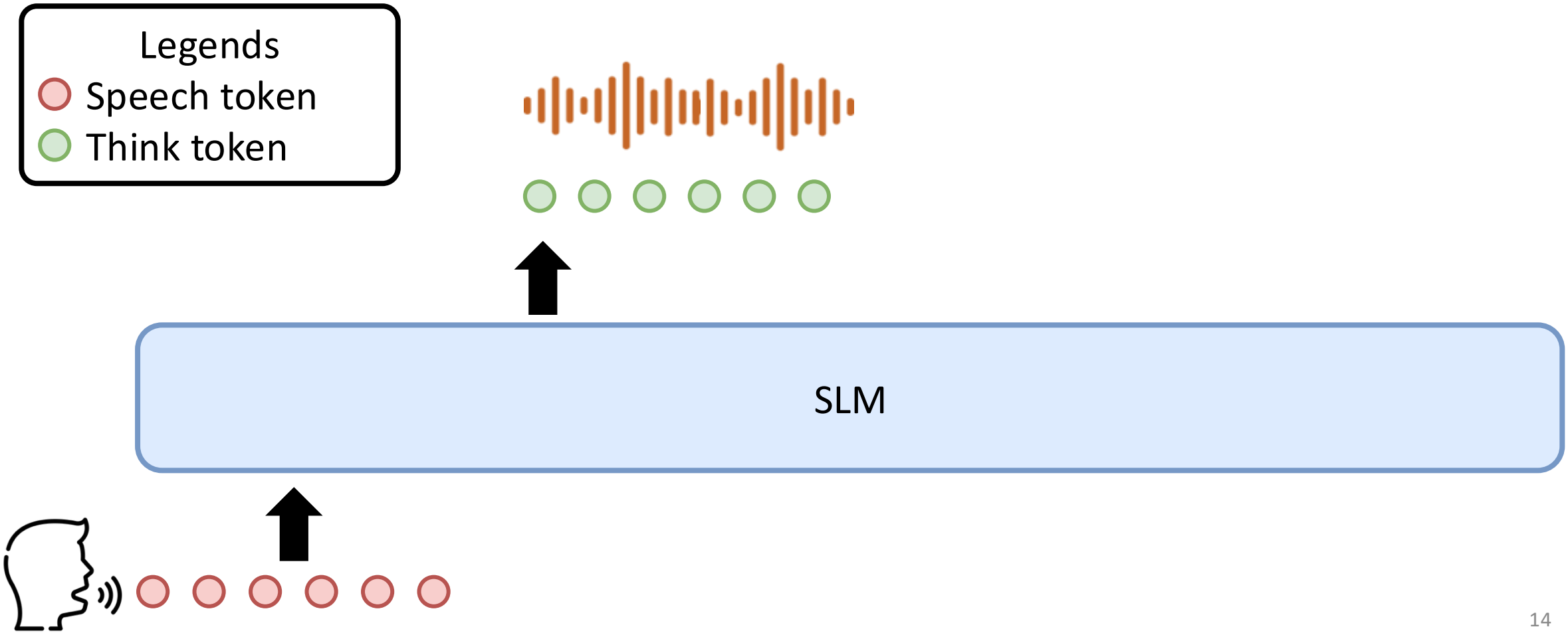
Goal: Thinking while Speaking

- We want to make SLMs think while speaking



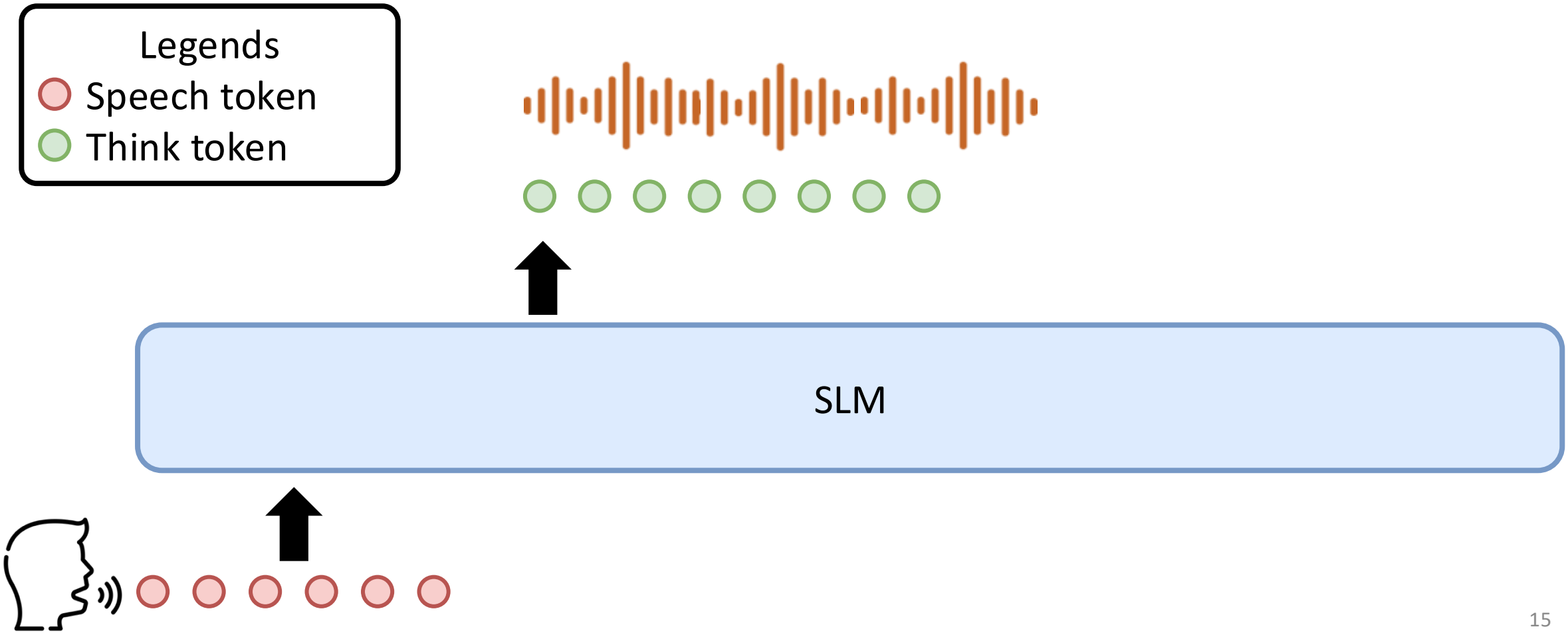
Goal: Thinking while Speaking

- We want to make SLMs think while speaking



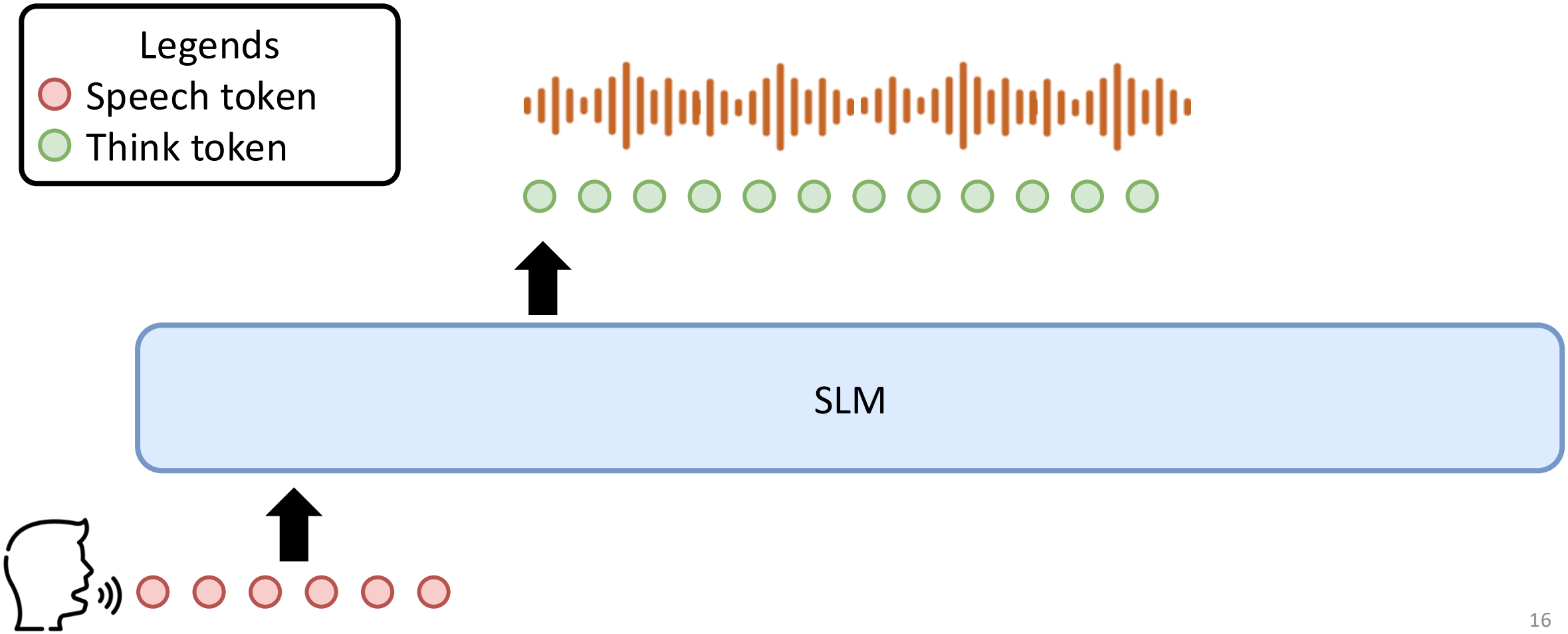
Goal: Thinking while Speaking

- We want to make SLMs think while speaking



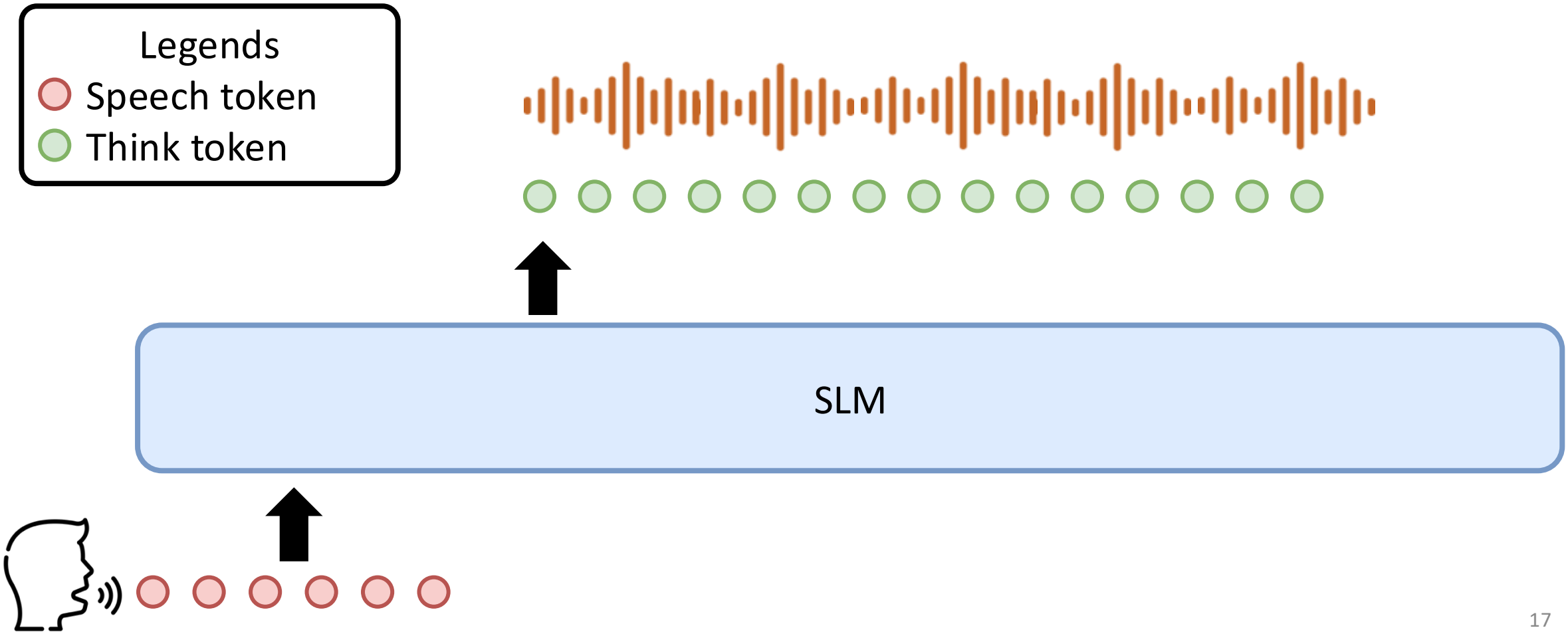
Goal: Thinking while Speaking

- We want to make SLMs think while speaking



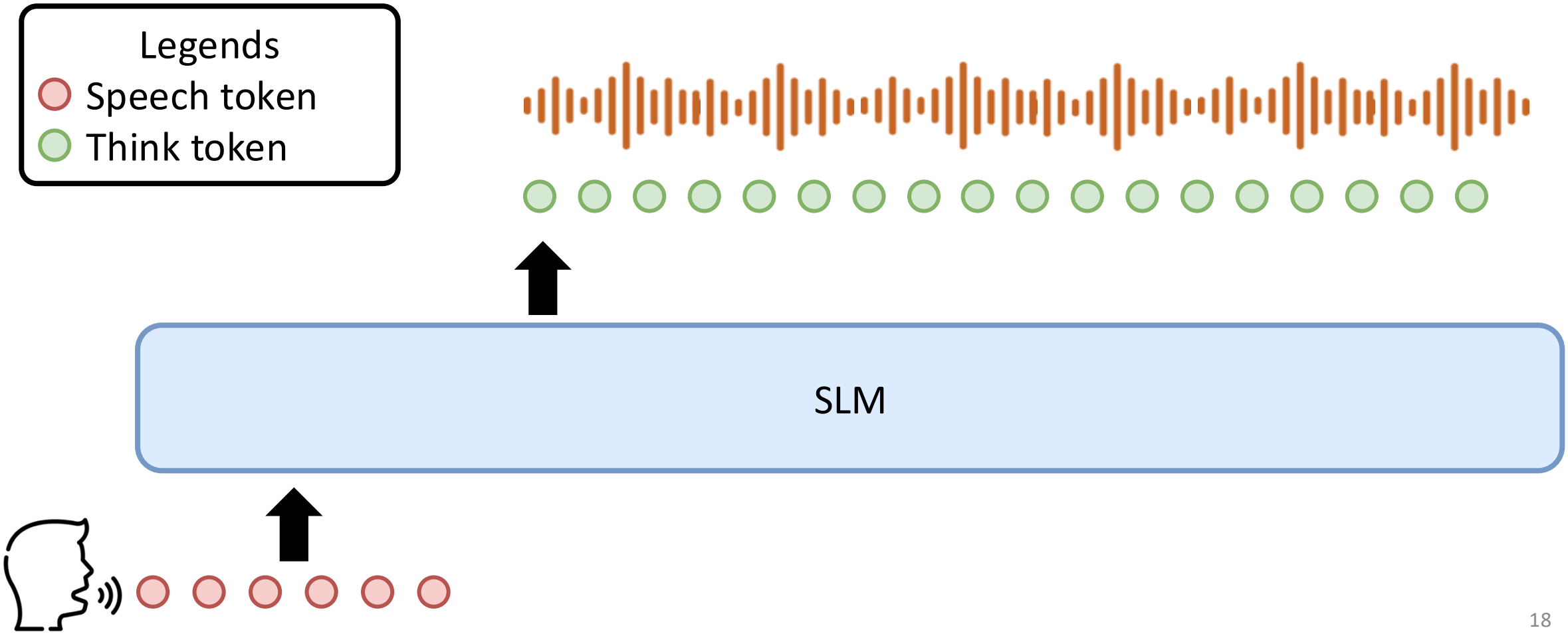
Goal: Thinking while Speaking

- We want to make SLMs think while speaking



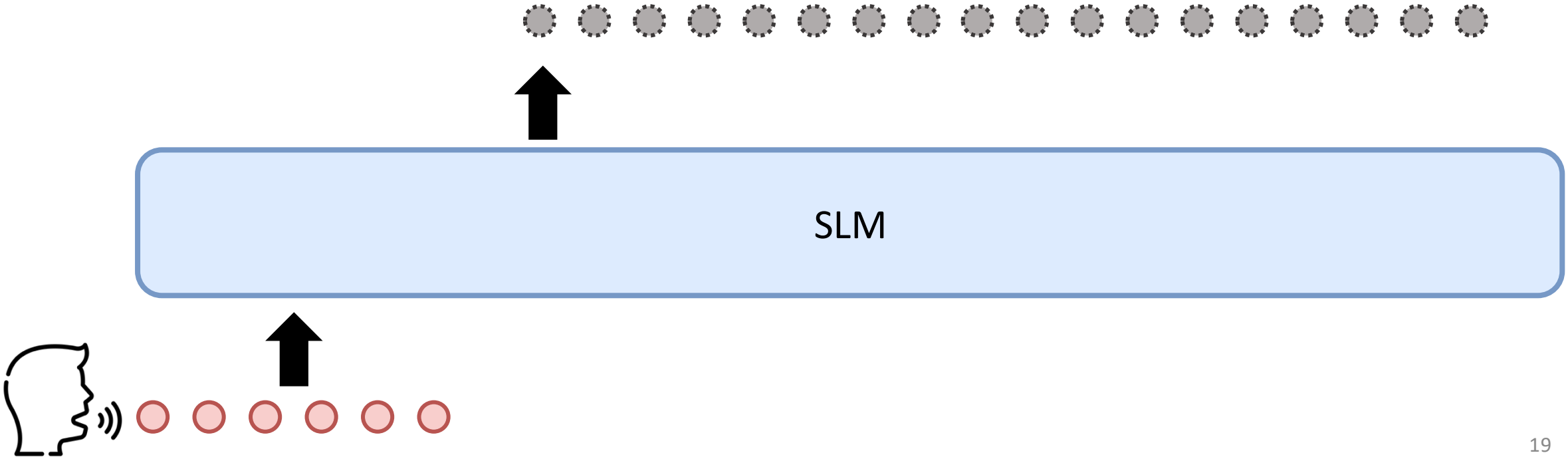
Goal: Thinking while Speaking

- We want to make SLMs think while speaking



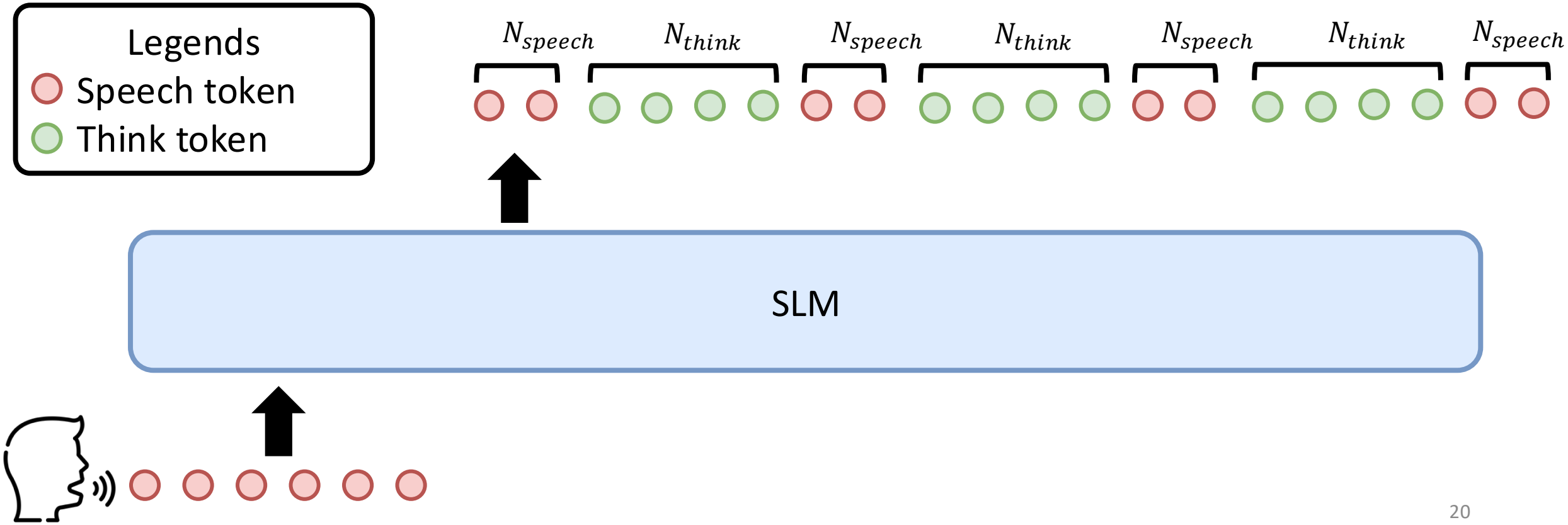
Goal: Thinking while Speaking

- Language model can only generate a single stream of token at a time
- Question: How do we use a single output stream of token but make it as if there are two types of tokens?



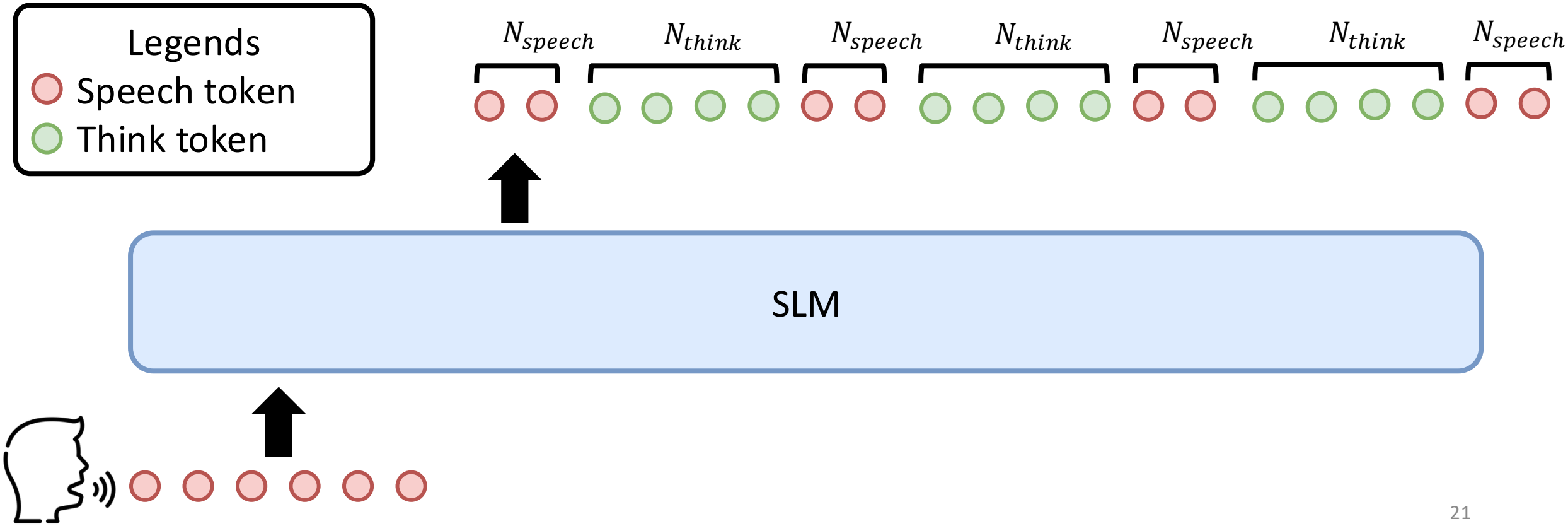
STITCH: Simultaneous Thinking and Talking

- Solution: Generate interleaved ***chunks*** of thinking and speech tokens
  - Thinking tokens: For ***unspoken*** internal thinking
  - Speech tokens: For the ***spoken*** responses



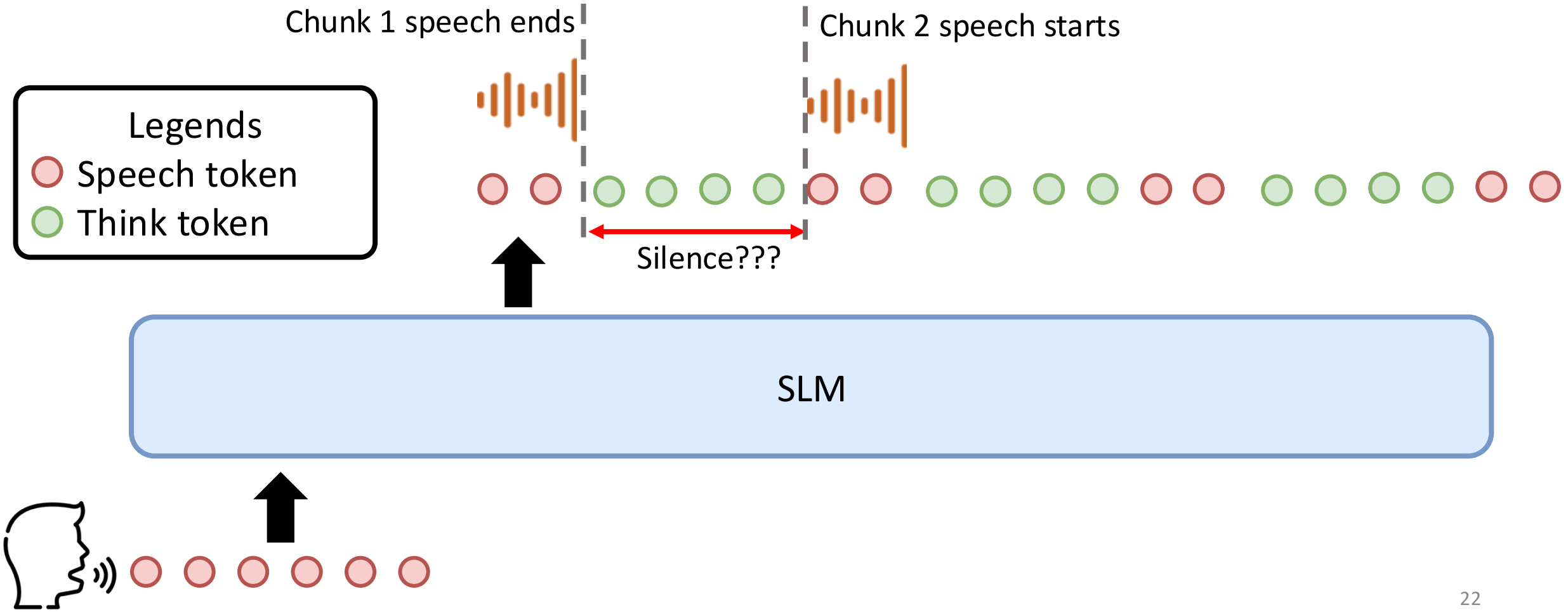
STITCH: Simultaneous Thinking and Talking

- The number of tokens in each chunk is fixed
  - N_{speech}: number of speech tokens in a speech chunk
  - N_{think}: number of thinking tokens in a thinking chunk



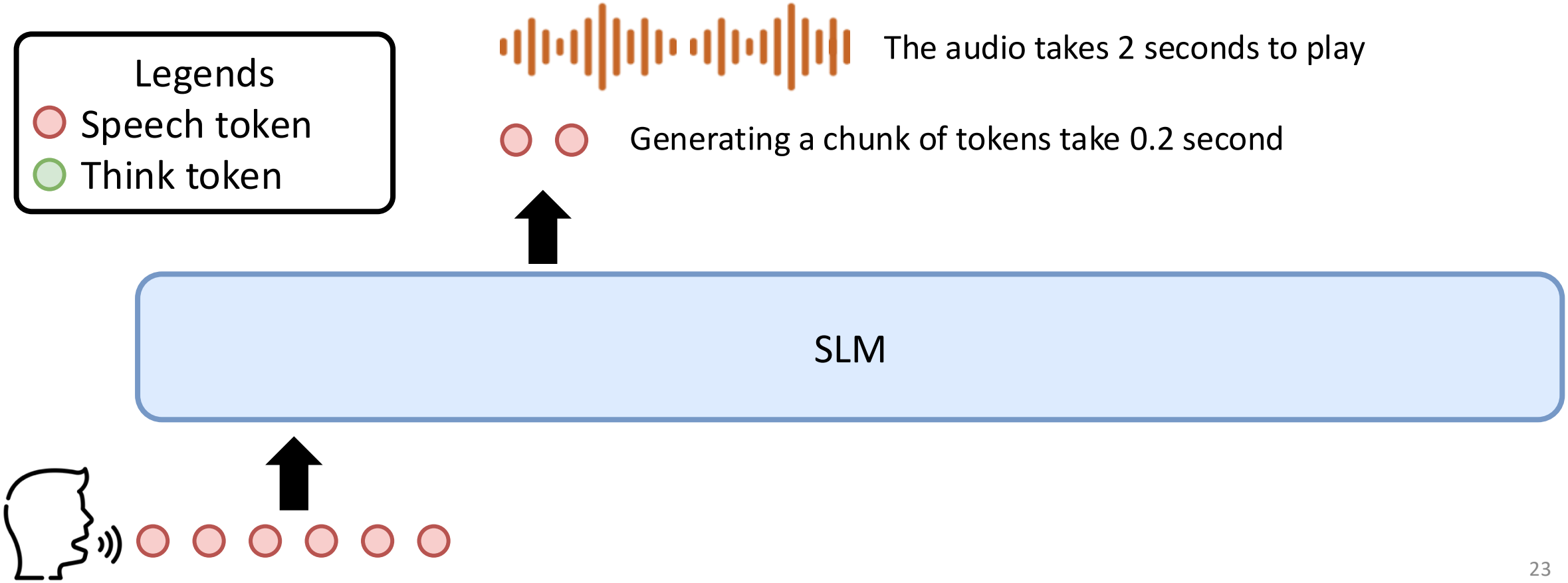
STITCH: Simultaneous Thinking and Talking

- Question: Does the unspoken thinking chunks break the speech response?



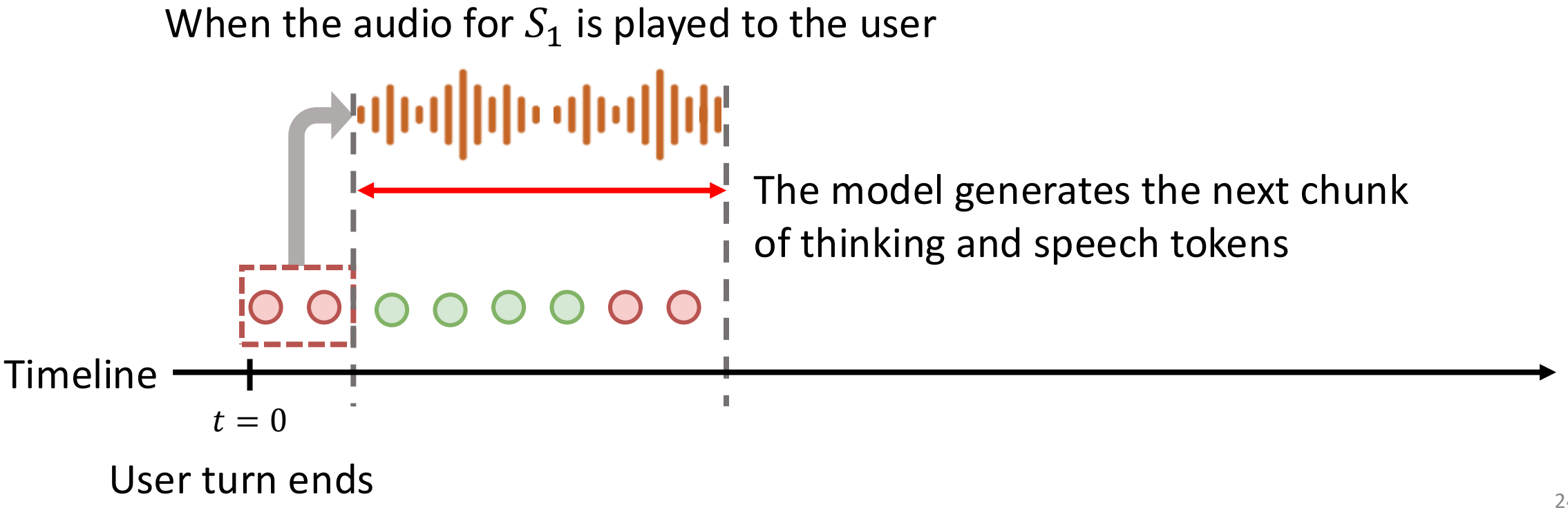
STITCH: Simultaneous Thinking and Talking

- Answer: No. Because the time for generating the speech tokens in a chunk is less than the time of the audio for that chunk



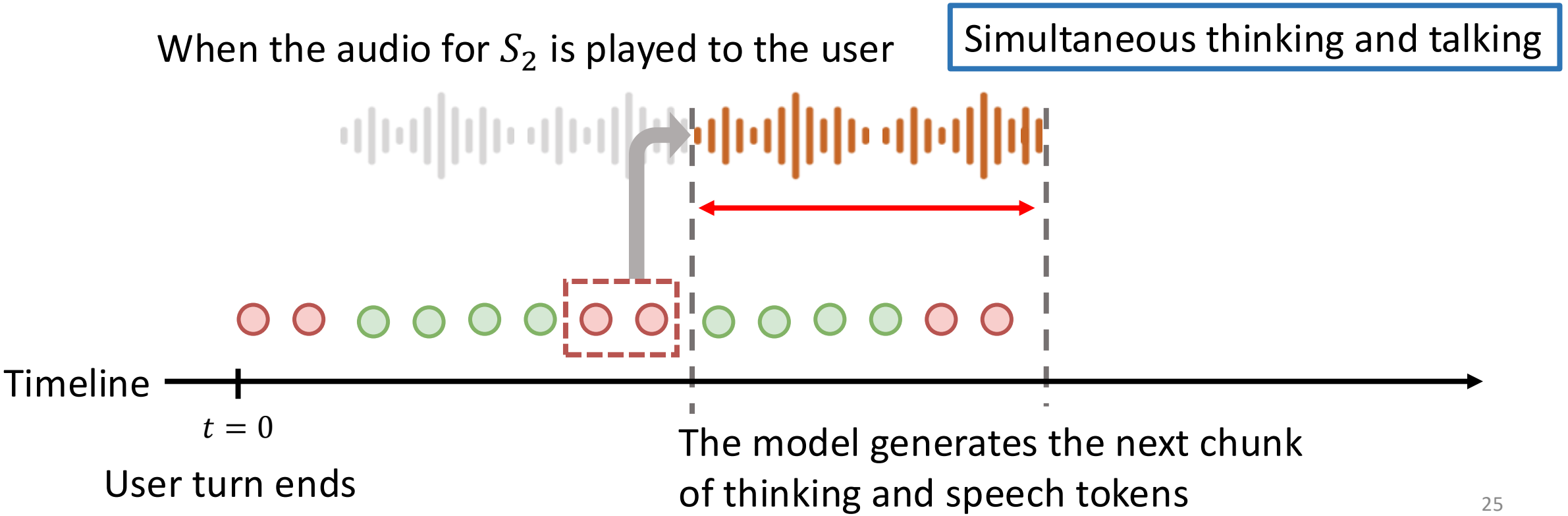
STITCH: Simultaneous Thinking and Talking

- By carefully selecting the ratio of N_{think} and N_{speech}, the duration of the audio for a speech token chunk can cover the time to generate $N_{think} + N_{speech}$ tokens



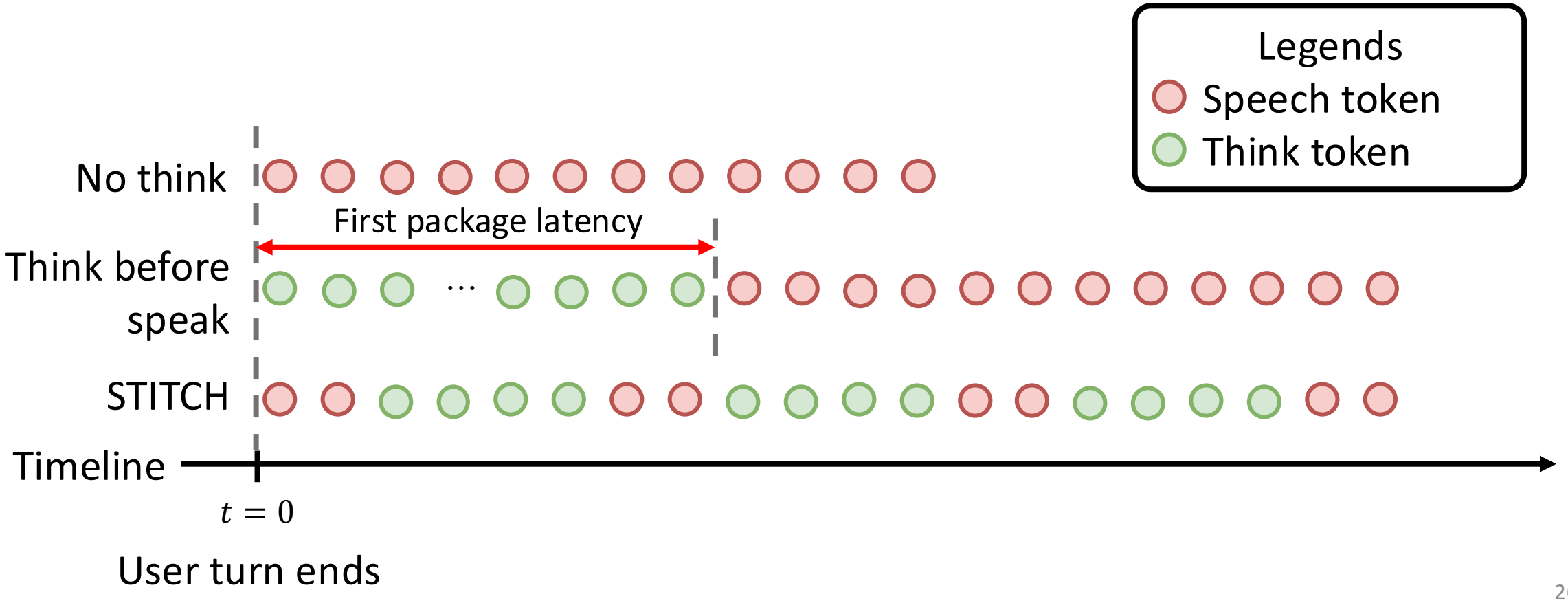
STITCH: Simultaneous Thinking and Talking

- By carefully selecting the ratio of the : N_{think}, N_{speech}, the time of the speech of N_{speech} speech token can cover the time to generate $N_{think} + N_{speech}$ tokens



STITCH: Simultaneous Thinking and Talking

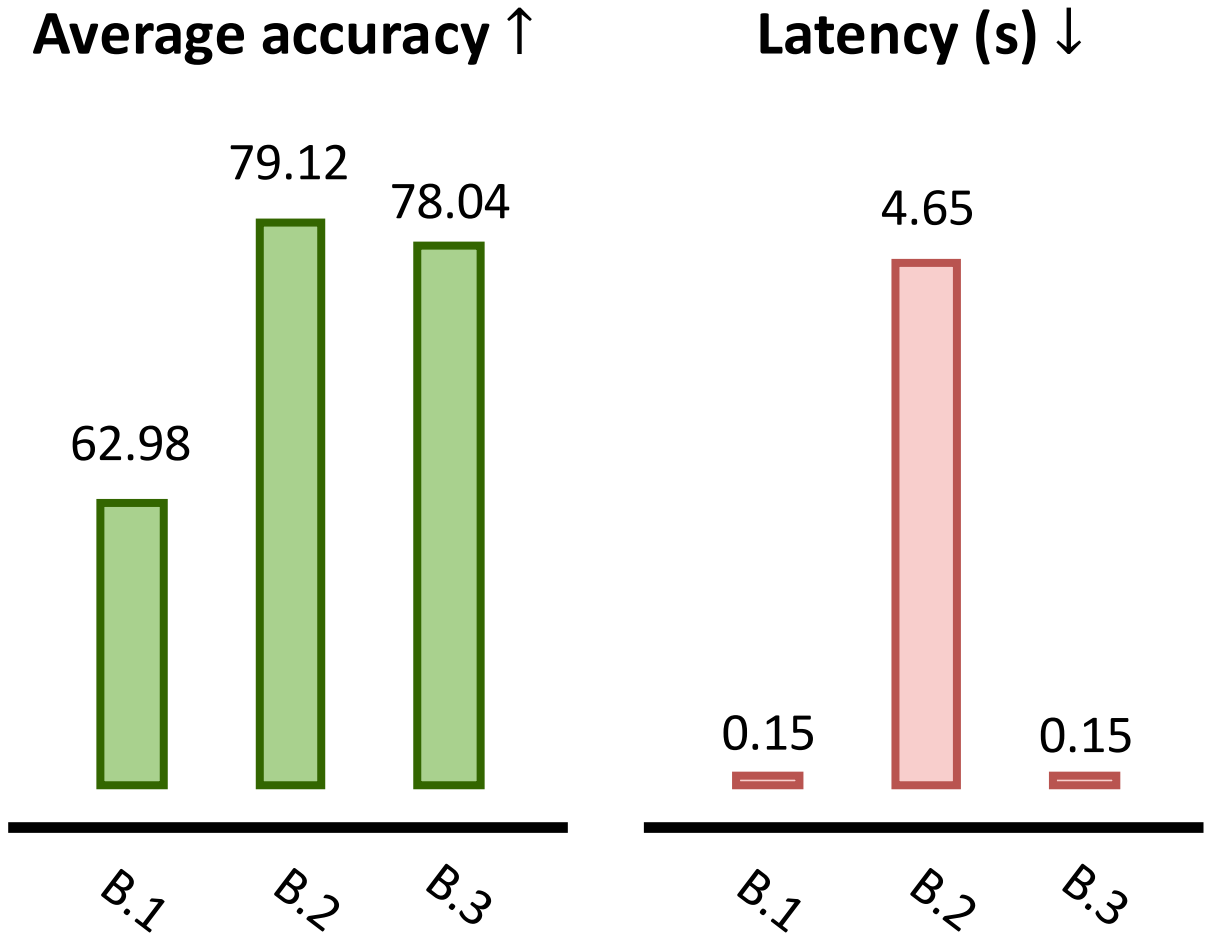
- First package latency (the time user needs to wait) comparison
  - Think before speak ≫ No think = STITCH



STITCH: Experiment Results

Compared methods

- B.1: No think
- B.2: Think before speaking
- B.3: STITCH (thinking while speaking)

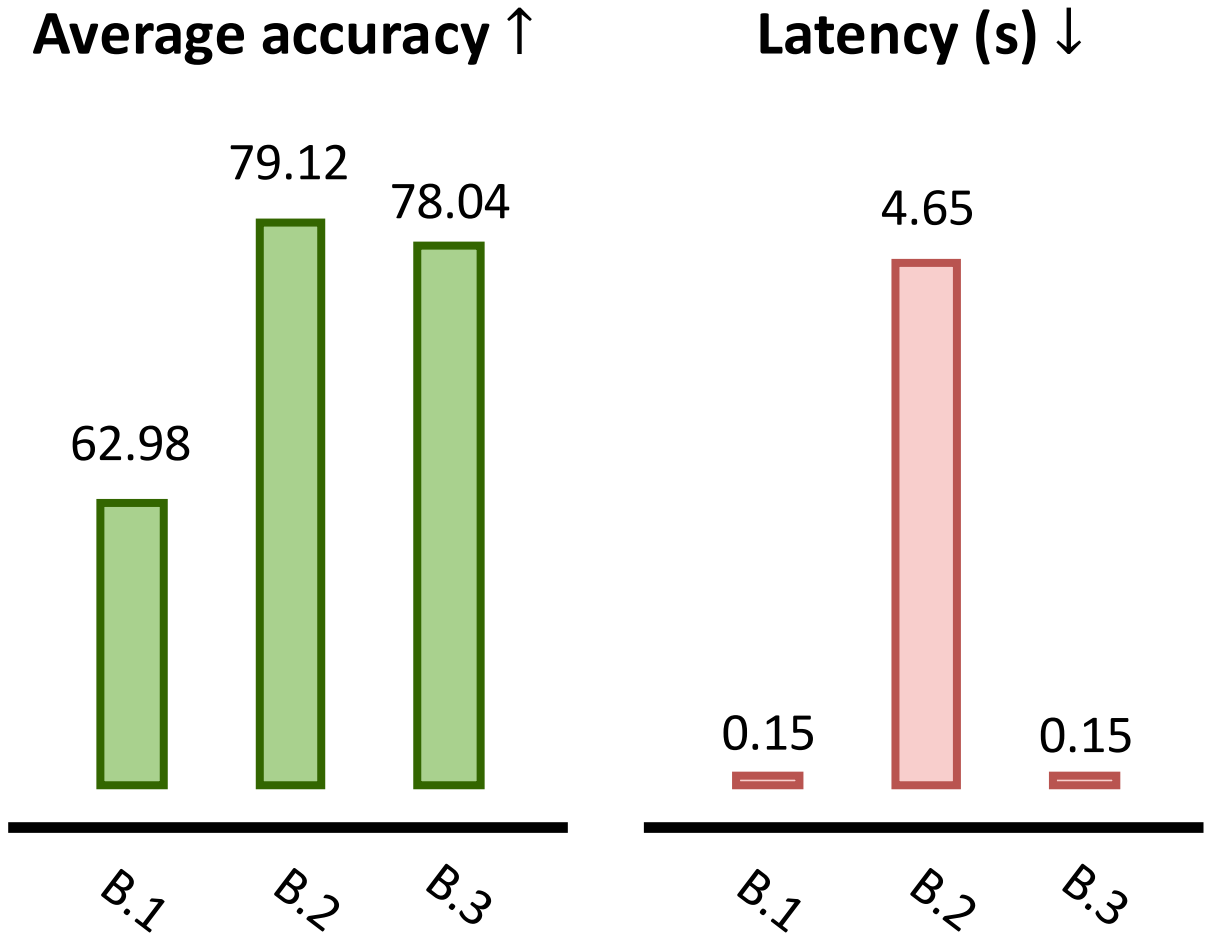


STITCH: Experiment Results

Compared methods

- B.1: No think
- B.2: Think before speaking
- B.3: STITCH (thinking while speaking)

- No thinking has a low latency and low accuracy

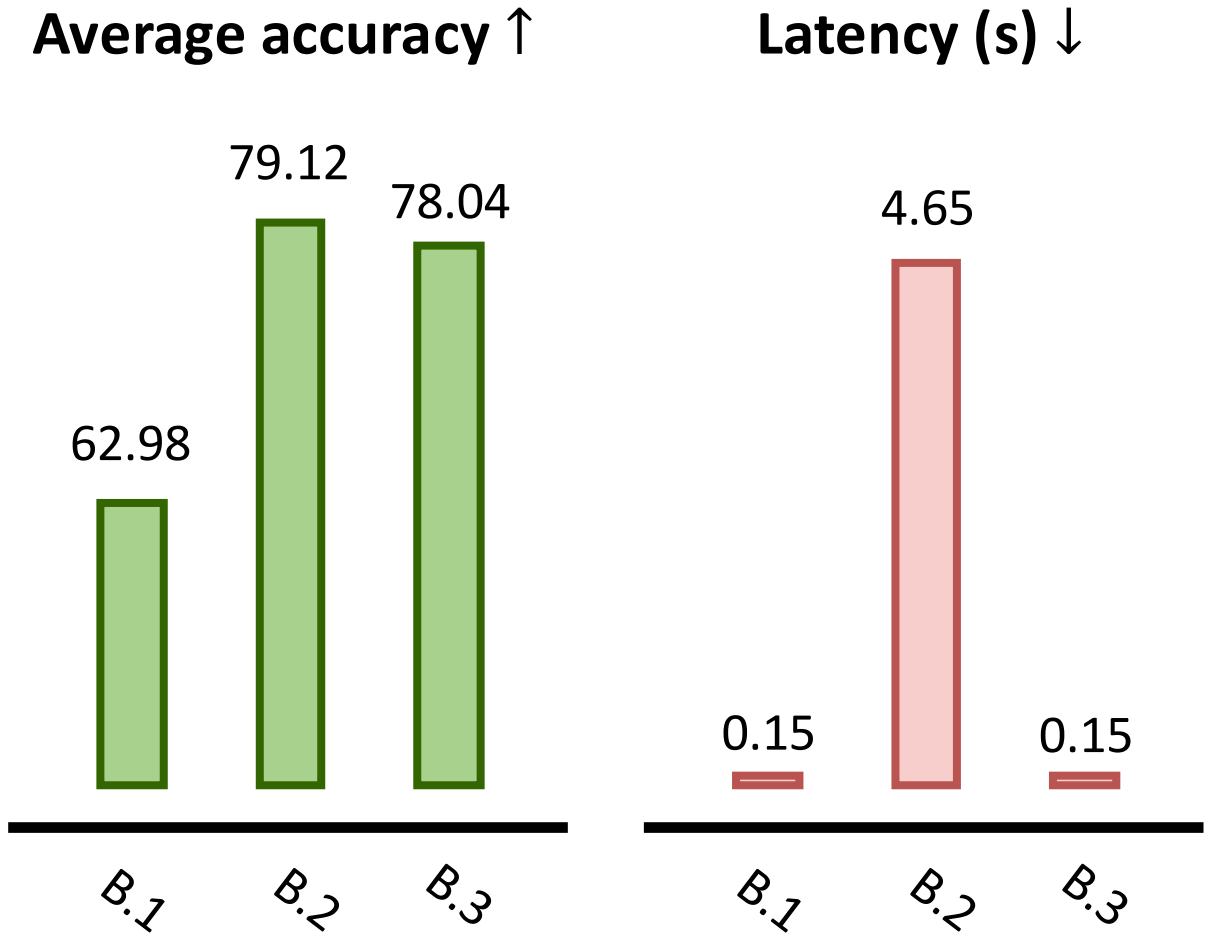


STITCH: Experiment Results

Compared methods

- B.1: No think
- B.2: Think before speaking
- B.3: STITCH (thinking while speaking)

- Thinking before speaking has a high accuracy but induces high latency

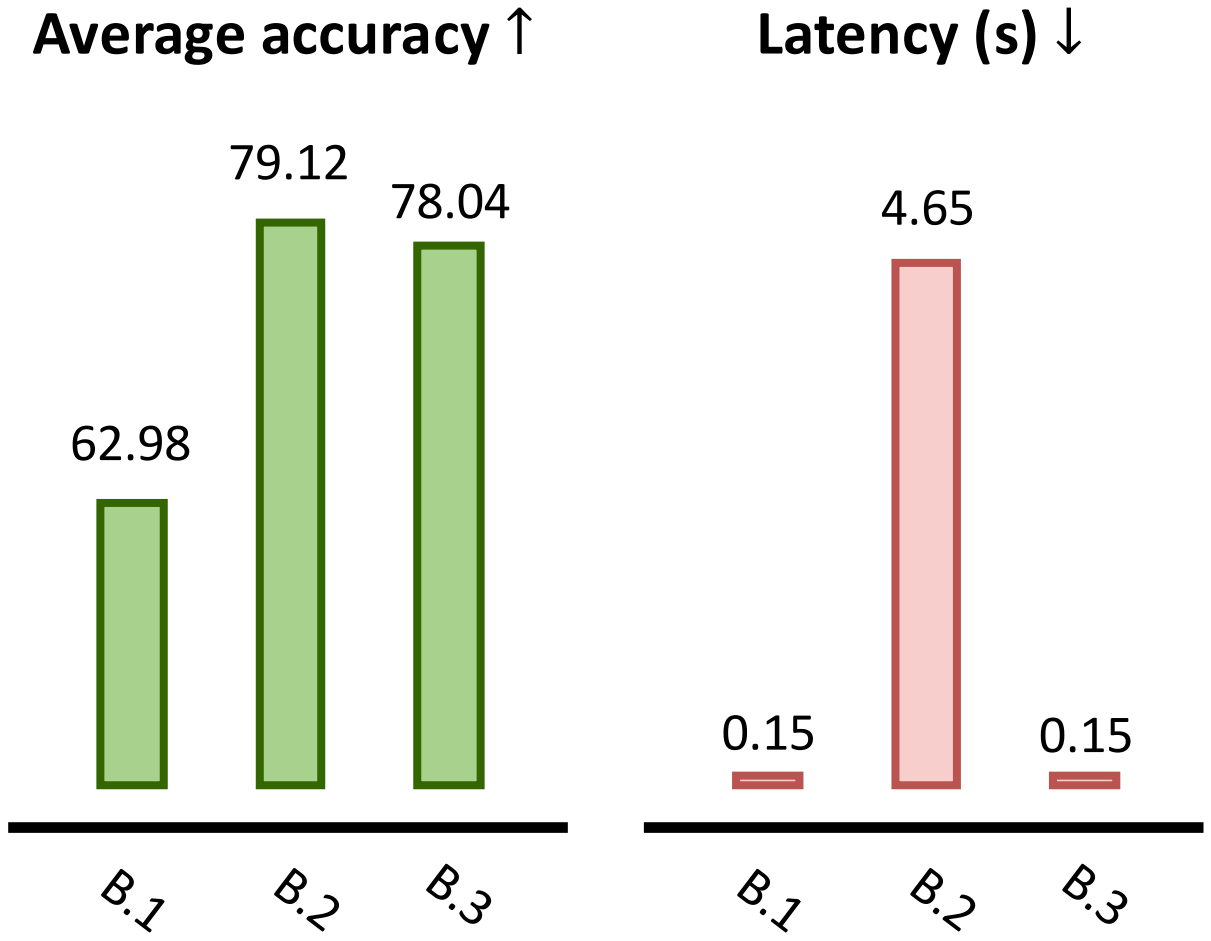


STITCH: Experiment Results

Compared methods

- B.1: No think
- B.2: Think before speaking
- B.3: STITCH (thinking while speaking)

- STITCH has low latency but high accuracy

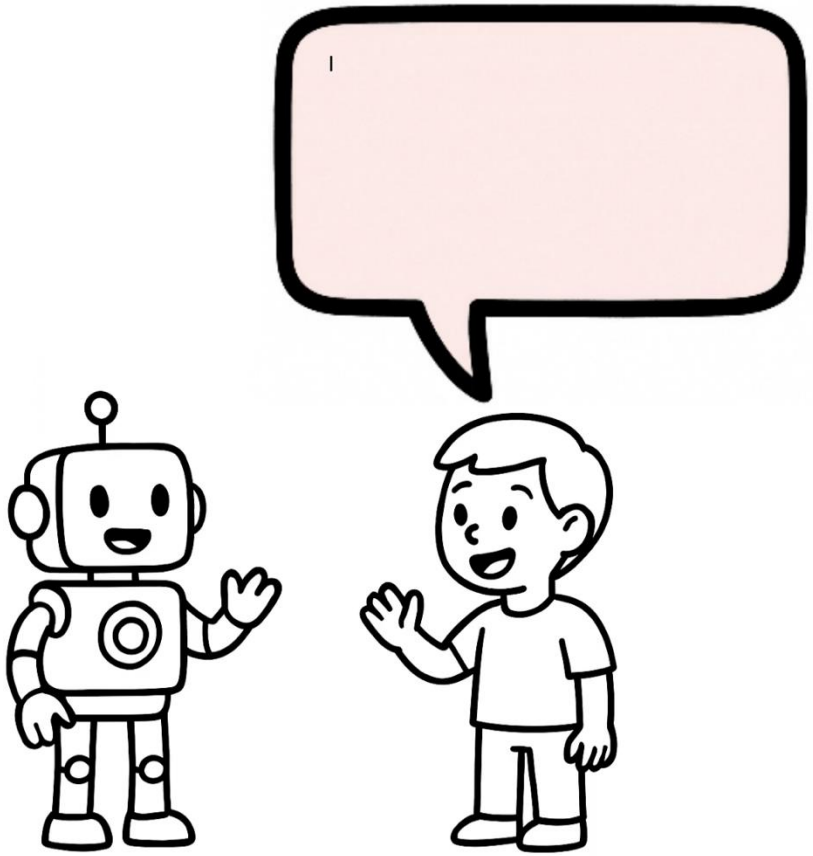


Summary

- Thinking while speaking can has a low latency as the no thinking model but has a much better performance on reasoning tasks
- Check our paper for more examples and details

Setting 3: Listen + Think → Speak

- Humans can also think while listening



https://arxiv.org/abs/2510.06917

SHANKS: Simultaneous listening and thinking

Acknowledgement

- The above work was done during my internship at Microsoft GenAI
  - My mentor: Xiaofei Wang

  - Other amazing team members: Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Lijuan Wang
  - My PhD advisor: Hung-yi Lee