Modern Voice Agents

LLM, Speech pipelines, and Real-time interaction

Who AM I

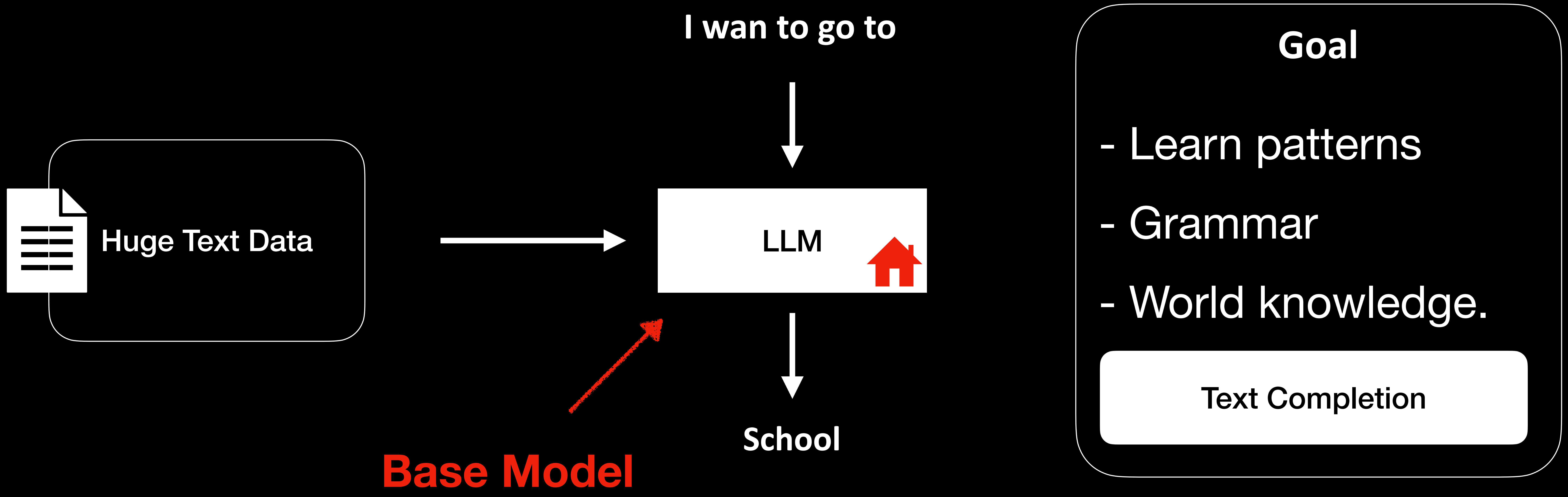
Abdeljalil EL MAJJODI

- ML Engineer @Norma
- President, Data Lead @Atlasia
- Open Source ML Contributor

LLM

From Text Completion To Reasoning

Pretraining: unsupervised learning



Pretraining: unsupervised learning

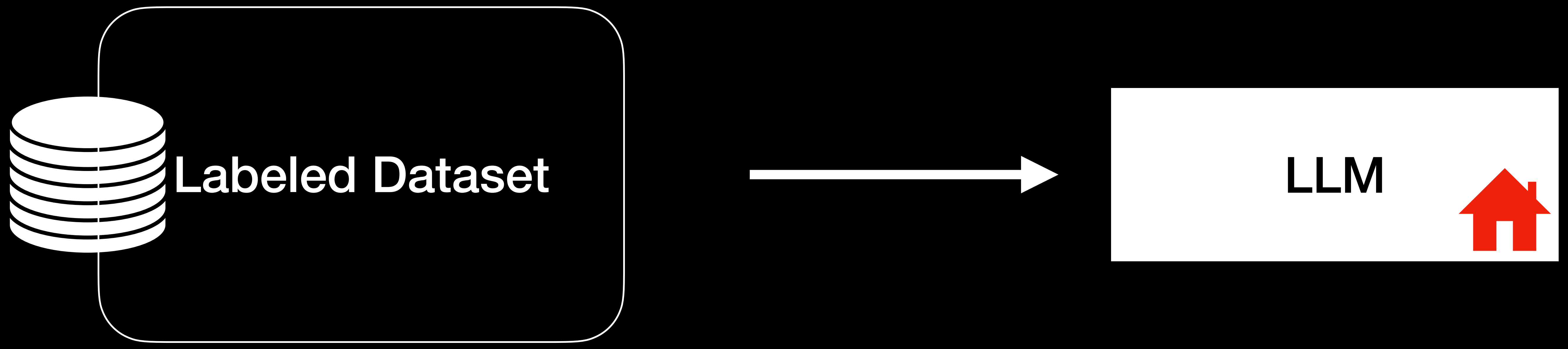
Not Following Instructions

What is the capital City of Morocco?

LLM

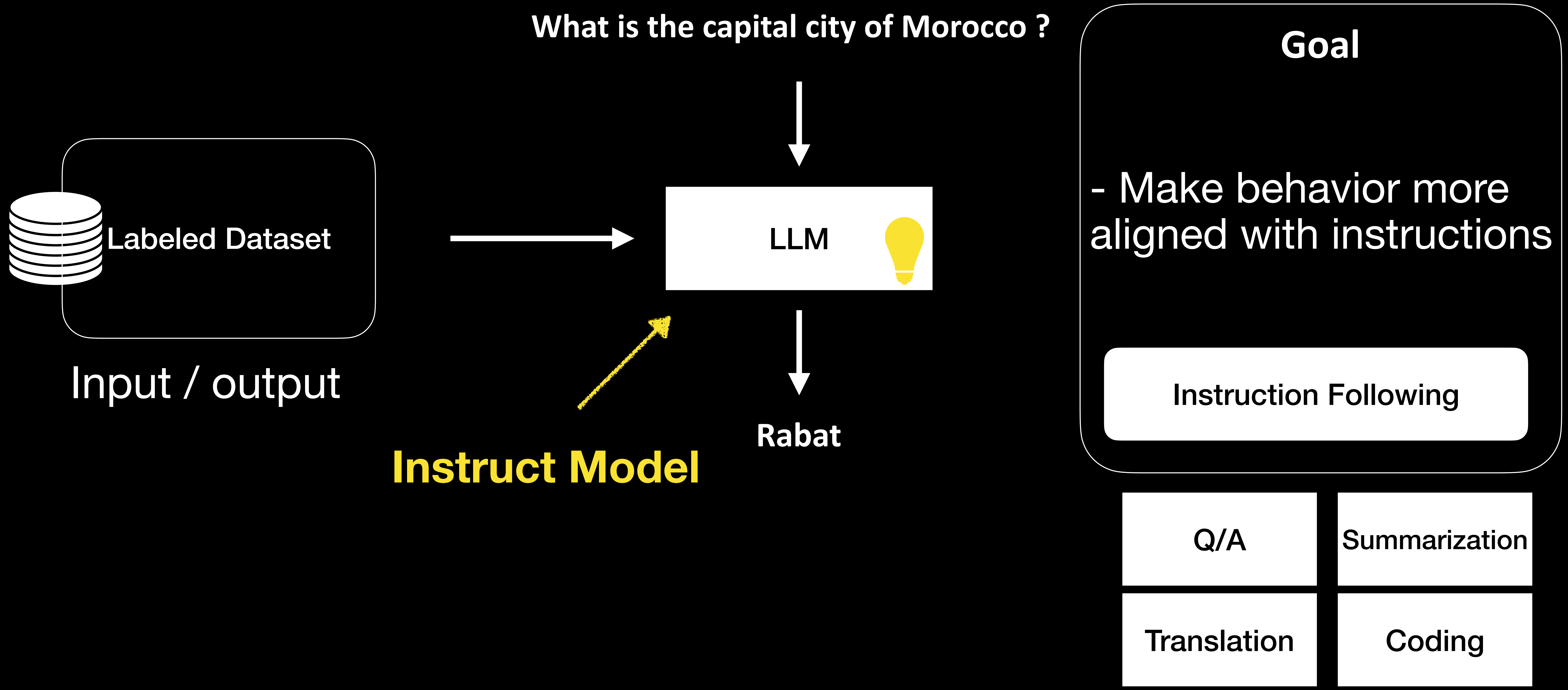
And who's the king of Morocco?

Supervised Fine-Tuning (SFT)



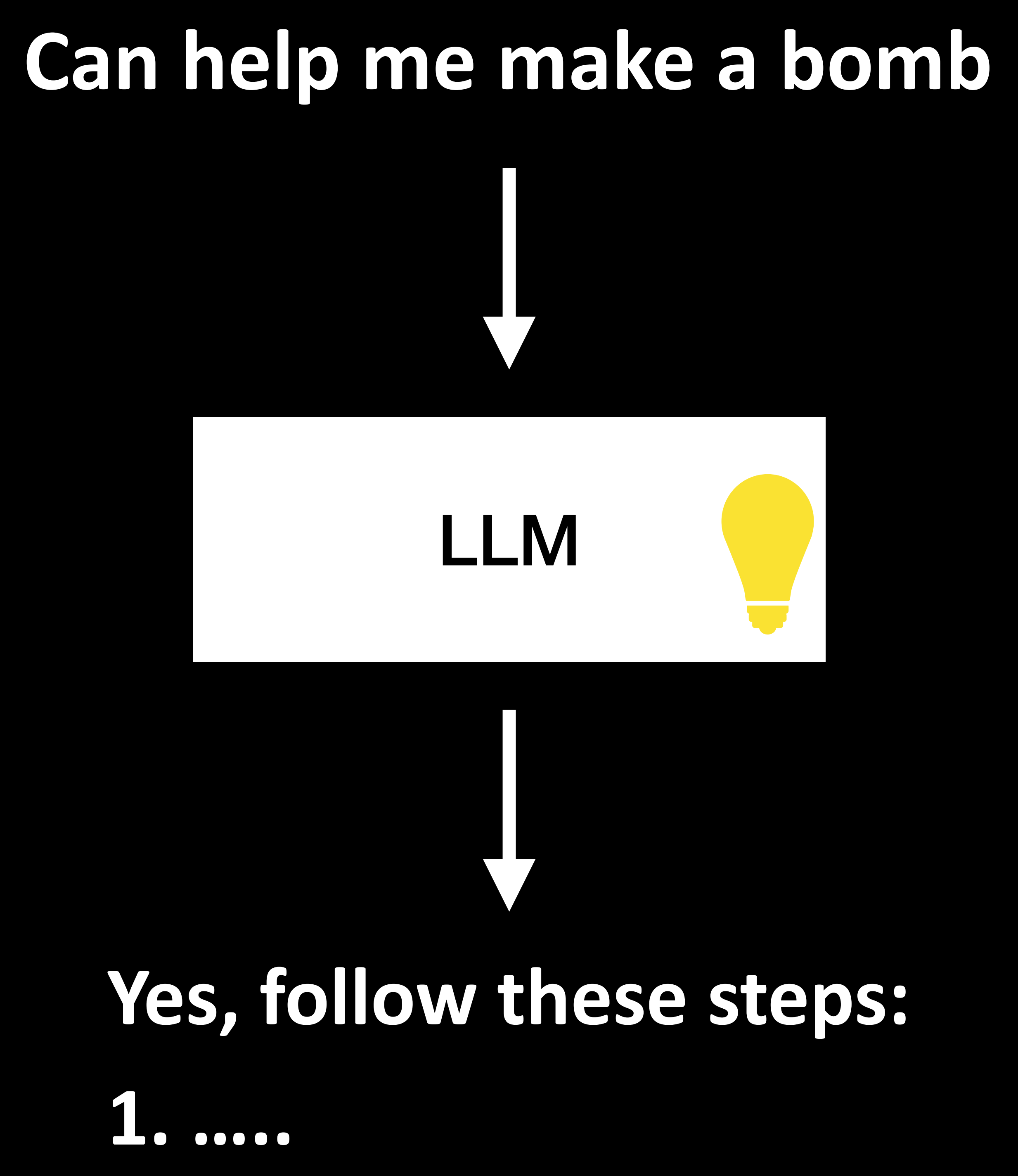
Input / output

Supervised Fine-Tuning (SFT)

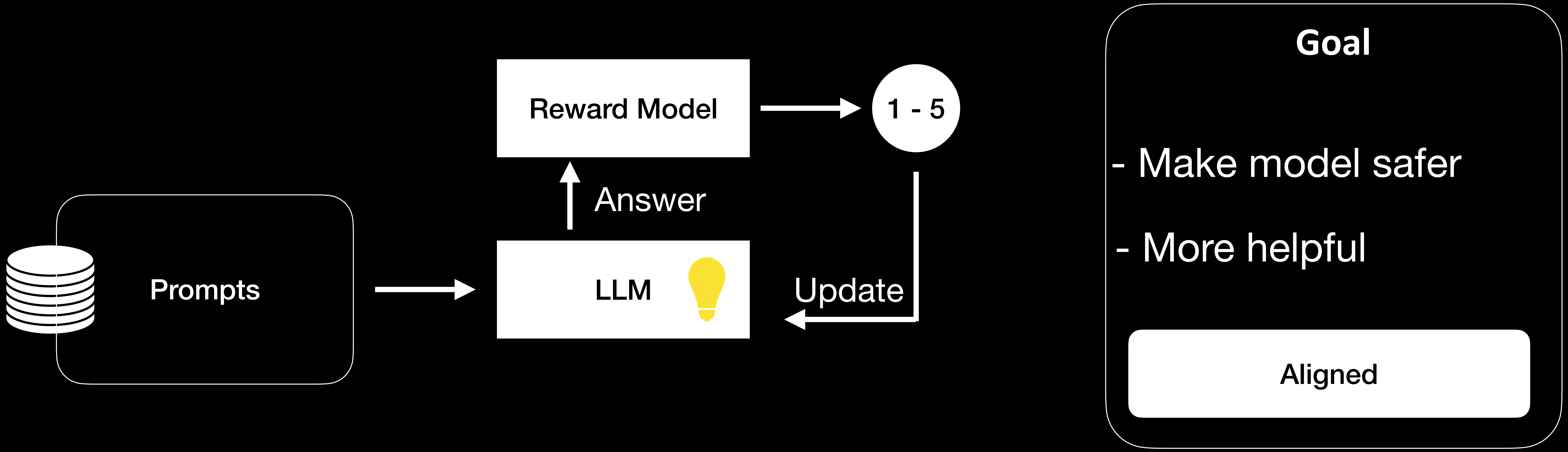


Pretraining: unsupervised learning

Not Safe

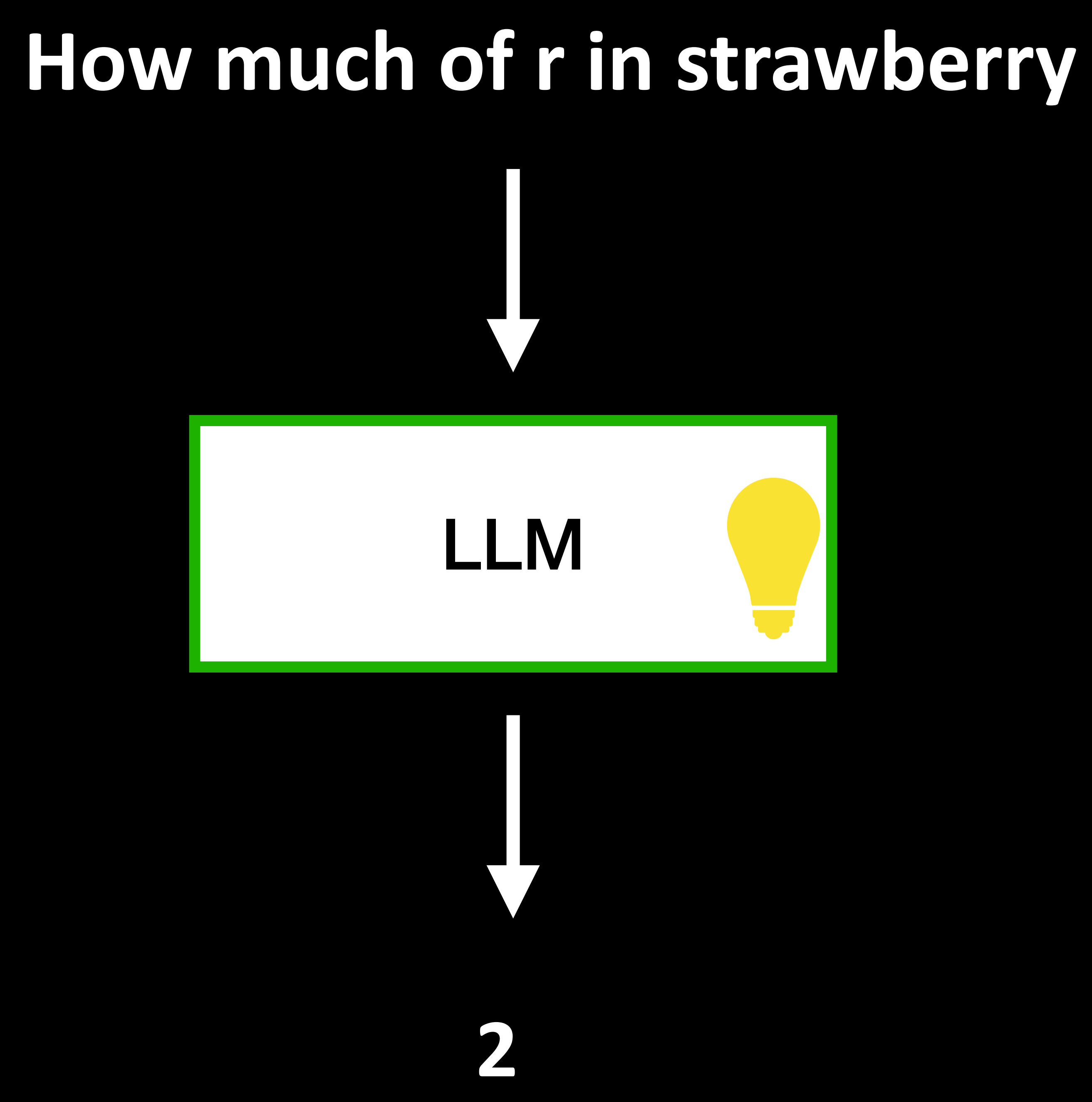


Human Preferences (RLHF / DPO)

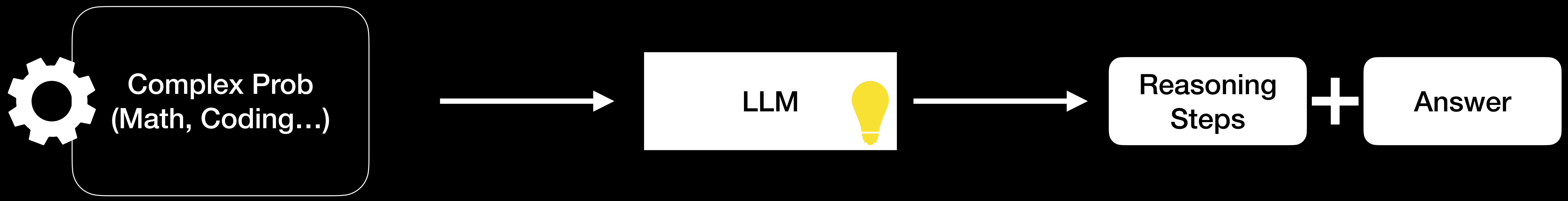


Human Preferences (RLHF / DPO)

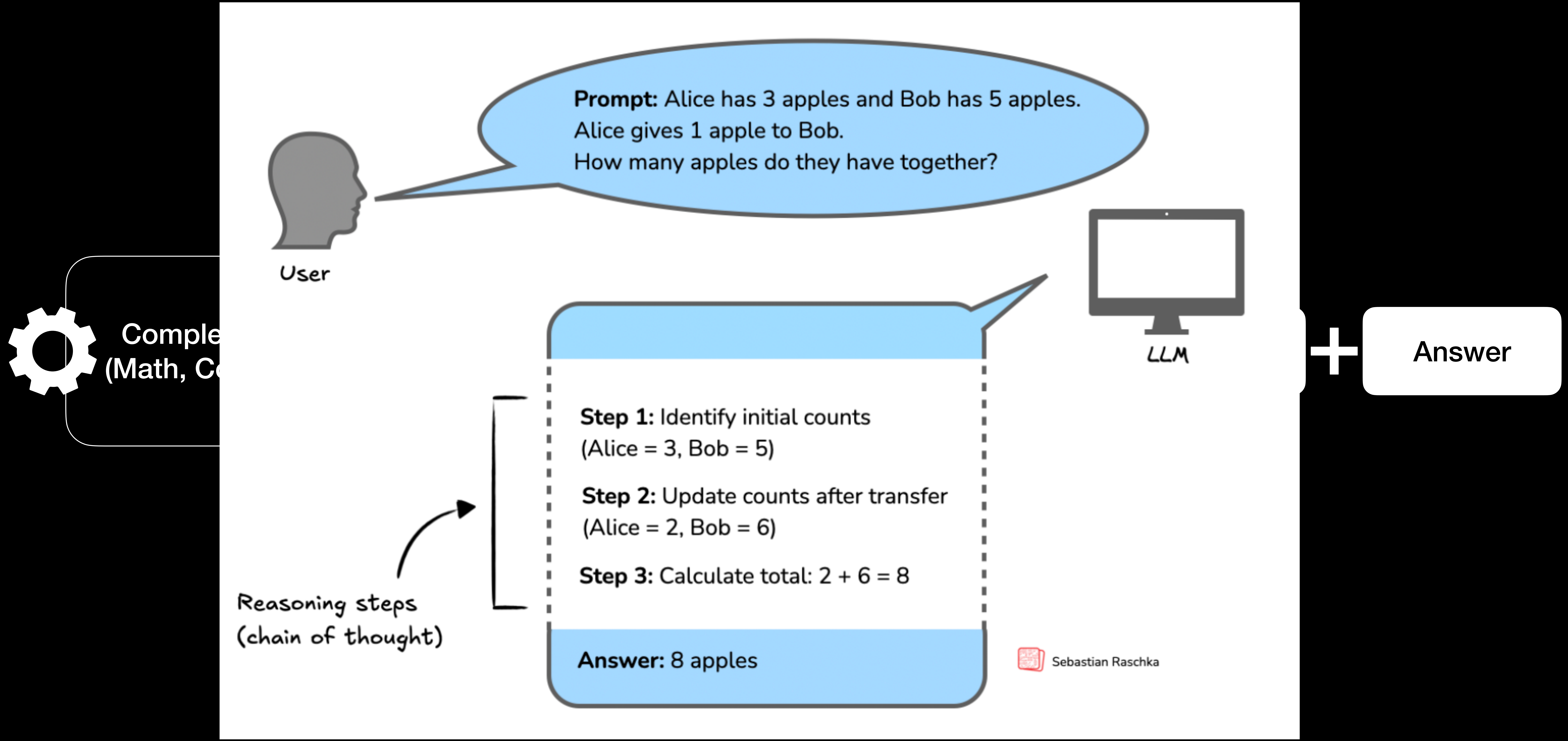
Cannot Fix Complex Problems



Reasoning (RLVR)



Reasoning (RLVR)

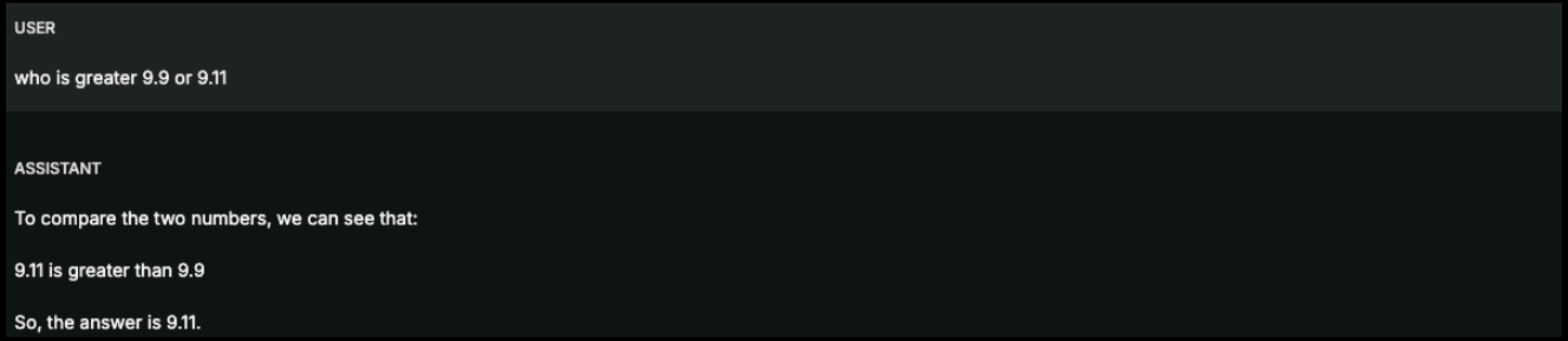


Agent

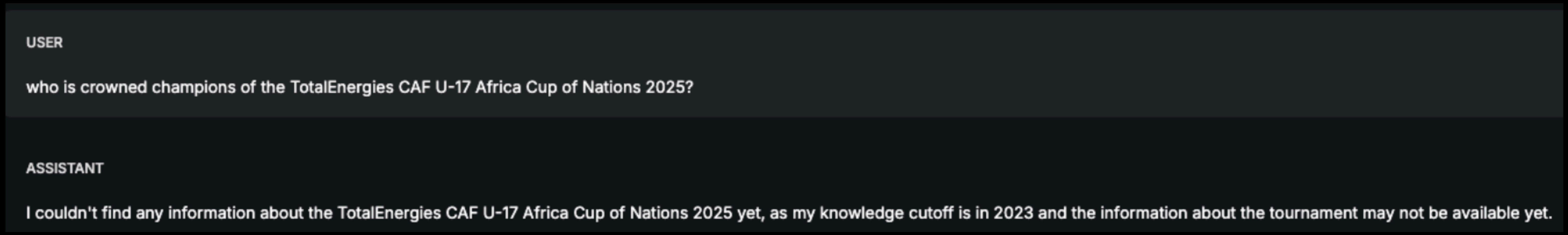
From LLM Limitations to LLM Tool Augmentation

LLM Limitations

- **Hallucinations:** generation of incorrect information with high confidence.



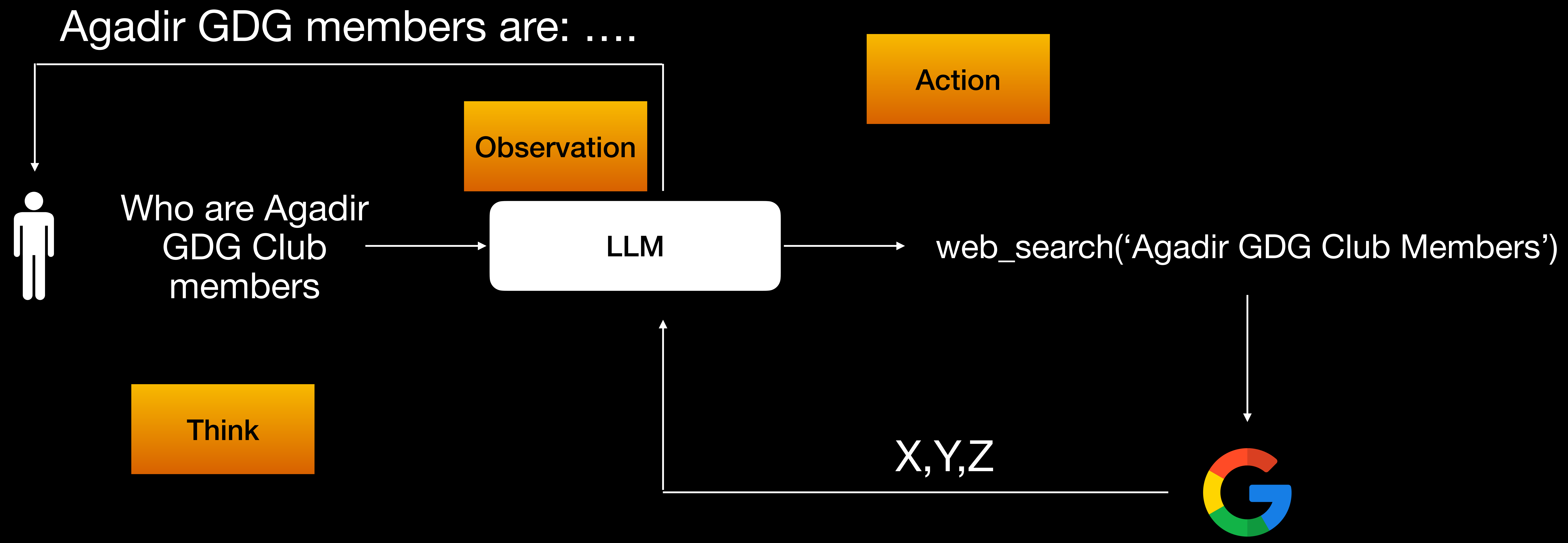
- **Knowledge cutoff:** limited to the training data timeframe.



- **Data privacy:** limited to public training data, no access to proprietary information.

Agent: LLM Tool Augmentation

Agent: LLM Tool Augmentation



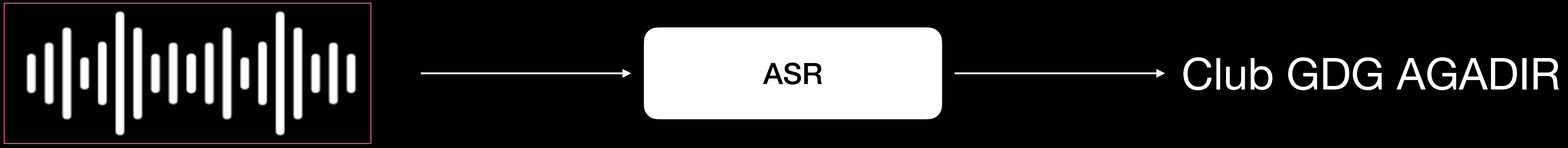
Voice Agent

Architectures, Latency, Network, and Frameworks

Architectures

Classic Architecture

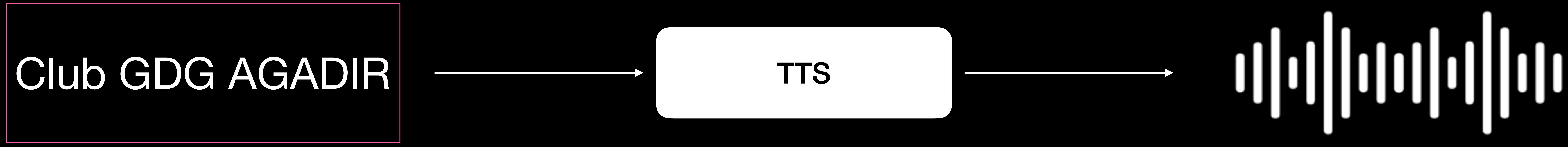
Offline ASR



entire audio input at once.

- **Models examples: whisper-large-v3(STT), distil-large-v3(STT)**
- **High latency**

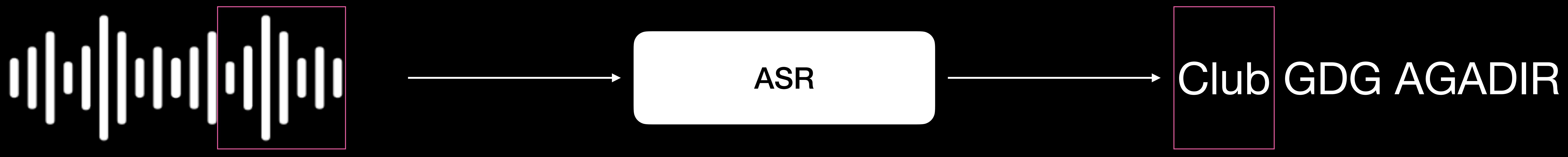
Offline TTS



entire Sentence input at once.

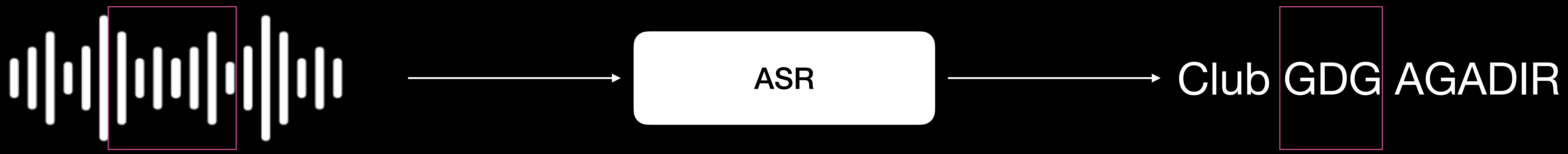
- **Models examples: SenseVoiceSmall, Parler-tts**
- **High latency**

Realtime ASR



small chunks as it is being spoken

Realtime ASR



small chunks as it is being spoken

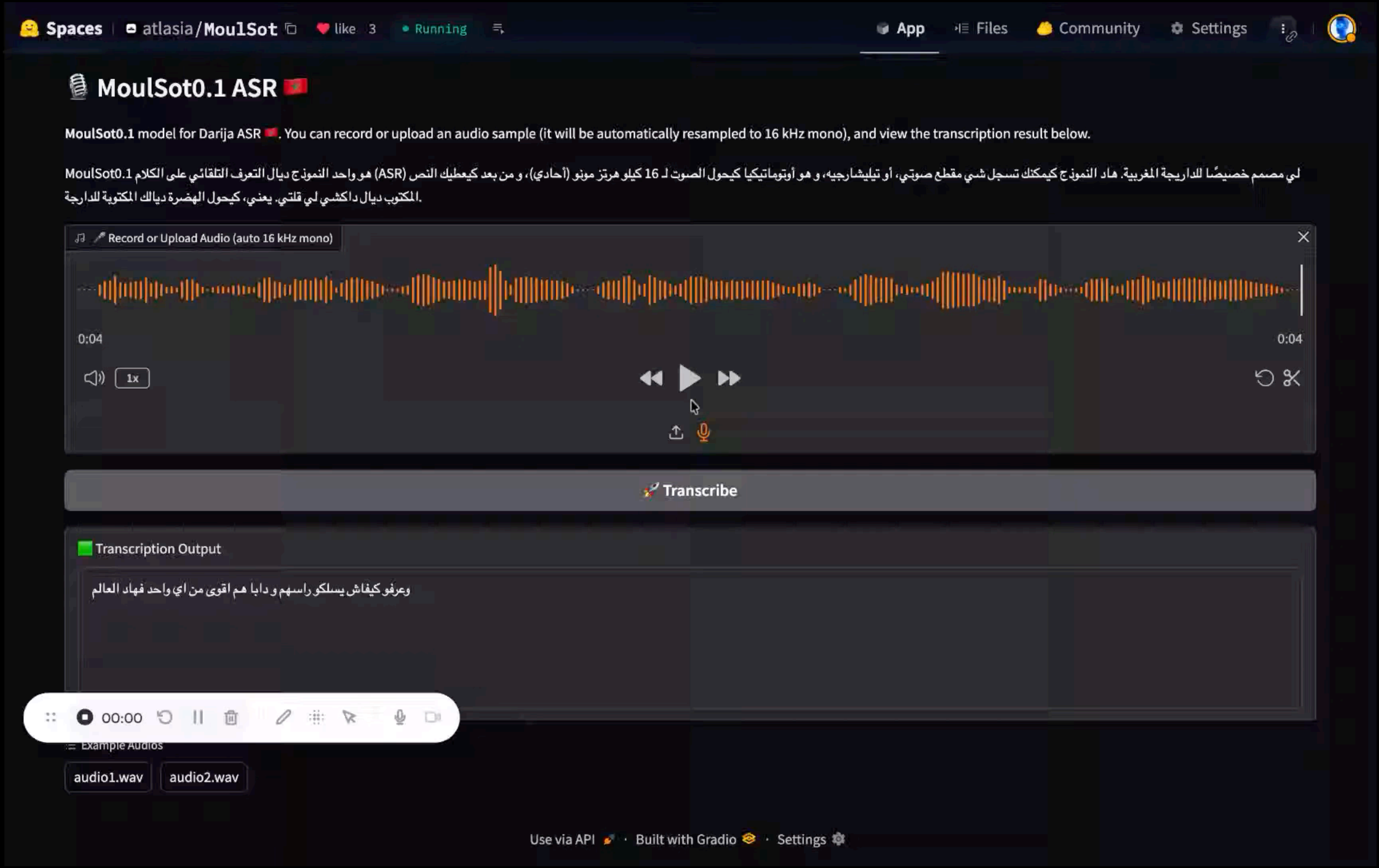
Realtime ASR

small chunks as it is being spoken

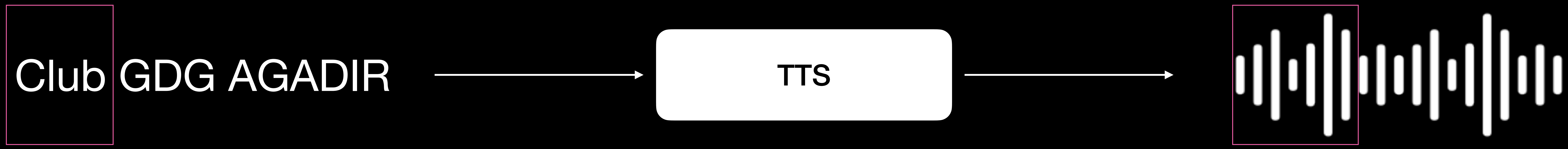
- Model examples: KyutaiSTT

- Low latency

ASR EXAMPLE

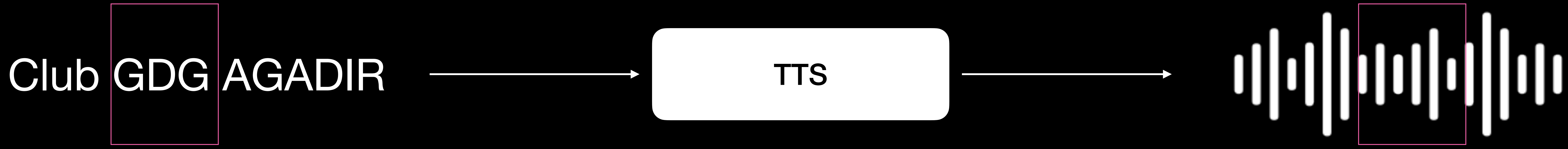

Spaces
atlasia/MoulSot
like 3
Running
App
Files
Community
Settings
MoulSot0.1 ASR
MoulSot0.1 model for Darija ASR. You can record or upload an audio sample (it will be automatically resampled to 16 kHz mono), and view the transcription result below.
MoulSot0.1 لي مصمم خصيصًا للداريجة المغربية. هاد النموذج كيمكنك تسجل شي مقطع صوتي، أو تيليشارجيه، و هو أوتوماتيكيا كيحول الصوت لـ 16 كيلو هرتز مونو (أحادي)، و من بعد كيعطيك النص (ASR) هو واحد النموذج ديال التعرف التلقائي على الكلام
المكتوب ديال داكشي لي قلتي. يعني، كيحول الهضرة ديالك المكتوبة للدارجة.
Record or Upload Audio (auto 16 kHz mono)
0:04
0:04
1x
Transcribe
Transcription Output
وعرفو كيفاش يسلكو راسهم و دابا هم اقوى من اي واحد فهاد العالم
00:00
Example Audios
audio1.wav
audio2.wav
Use via API · Built with Gradio · Settings

Realtime TTS



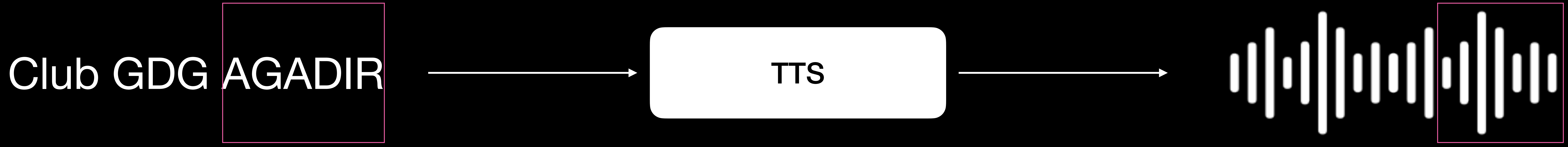
entire Sentence input at once.

Realtime TTS



entire Sentence input at once.

Realtime TTS

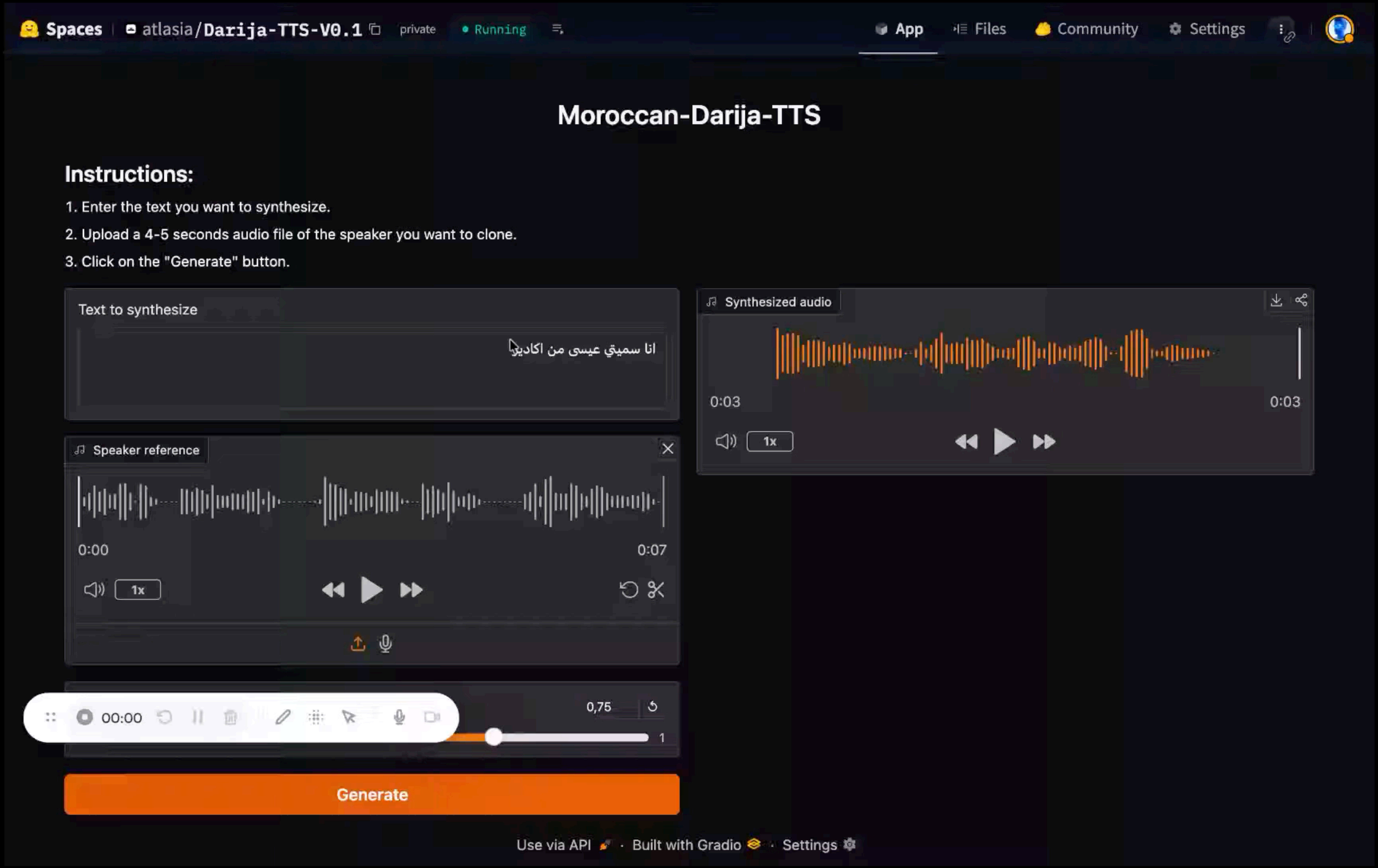


entire Sentence input at once.

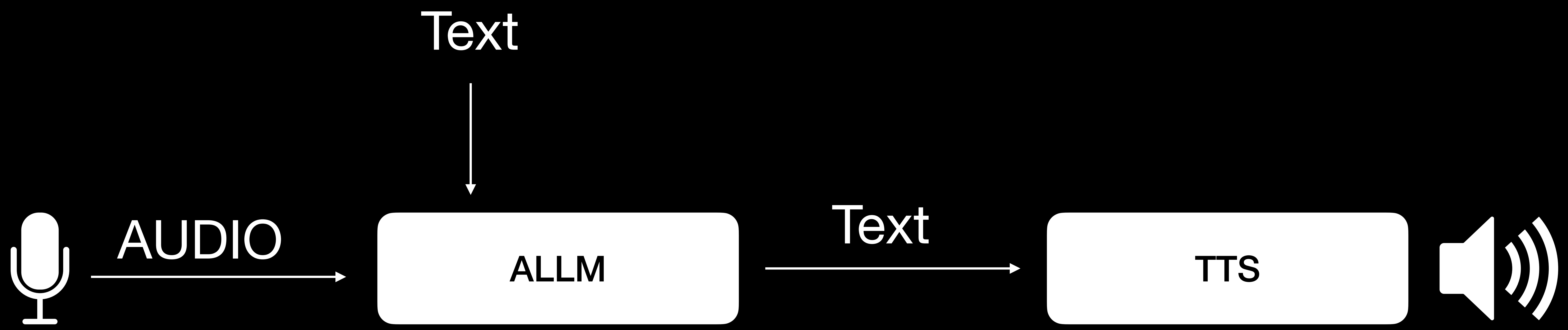
- **Model examples: CosyVoiceTTS, KyutaiTTS**

- **Low latency**

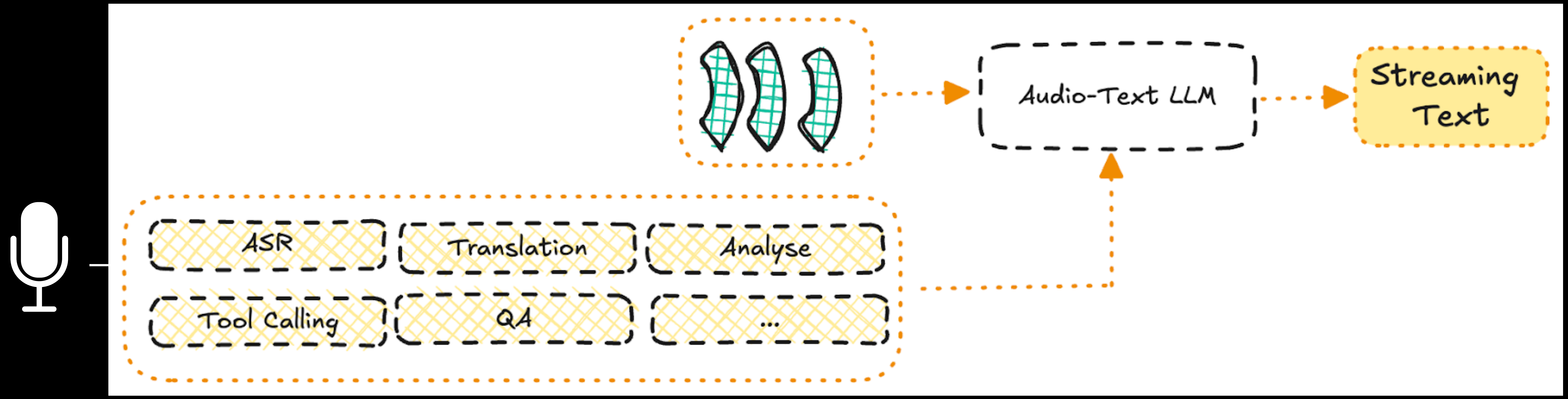
TTS Example


Spaces
atlasia/Darija-TTS-V0.1
private
Running
App
Files
Community
Settings
Moroccan-Darija-TTS
Instructions:
1. Enter the text you want to synthesize.
2. Upload a 4-5 seconds audio file of the speaker you want to clone.
3. Click on the "Generate" button.
Text to synthesize
انا سميتي عيسى من اكاديو
Synthesized audio
0:03
0:03
1x
Speaker reference
0:00
0:07
1x
00:00
0,75
1
Generate
Use via API · Built with Gradio · Settings

Audio LLM Architecture

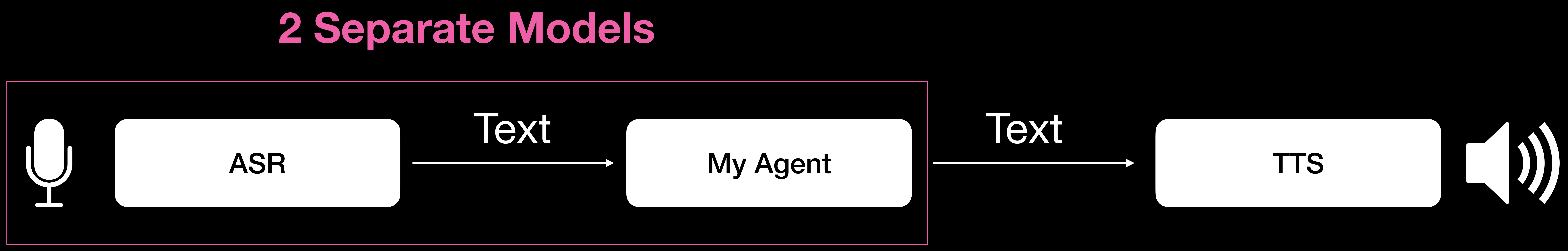


Audio LLM Architecture



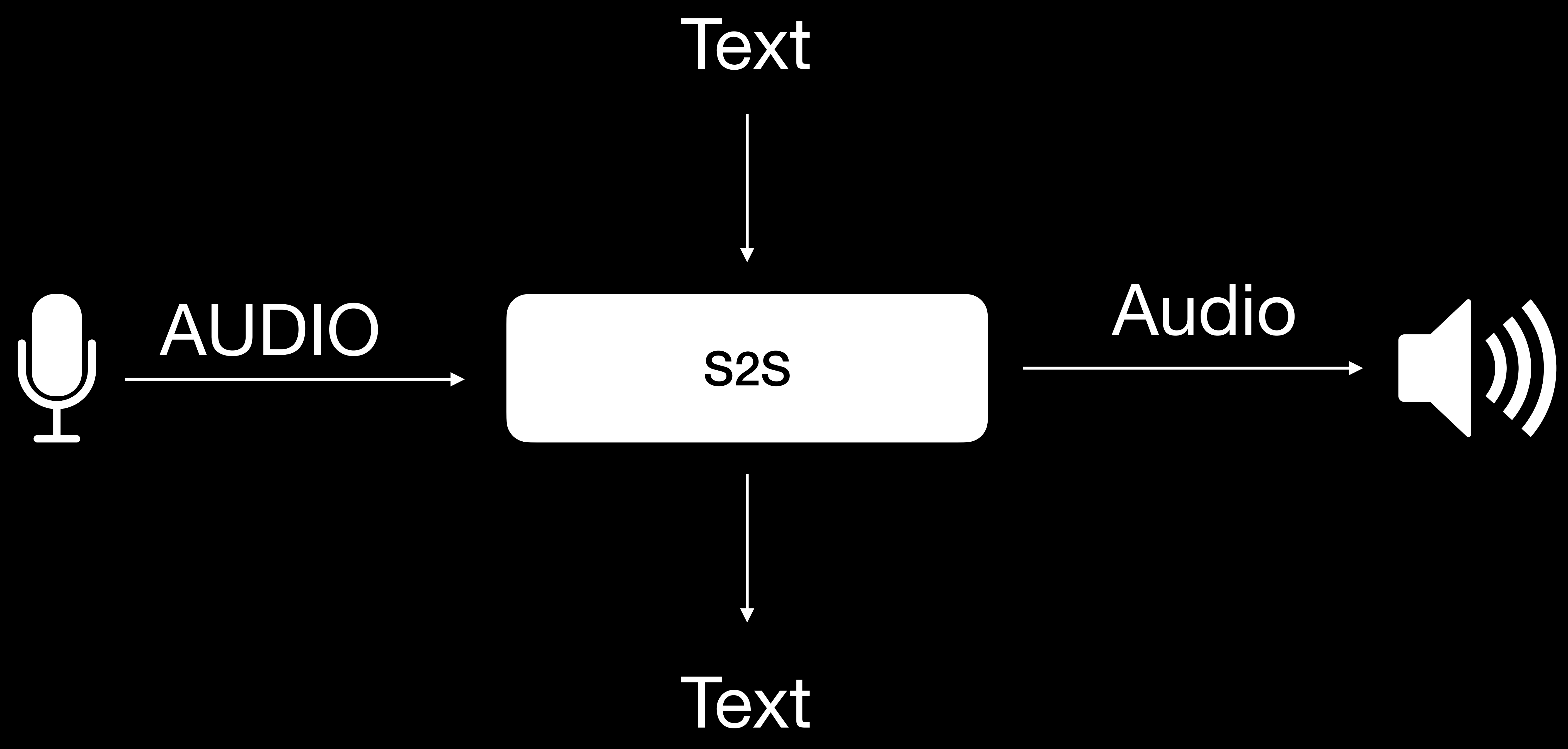
- **Model examples:** Qwen-audio, Voxtral, Ultravox, Flamingo

Classic Architecture



Audio LLM Architecture

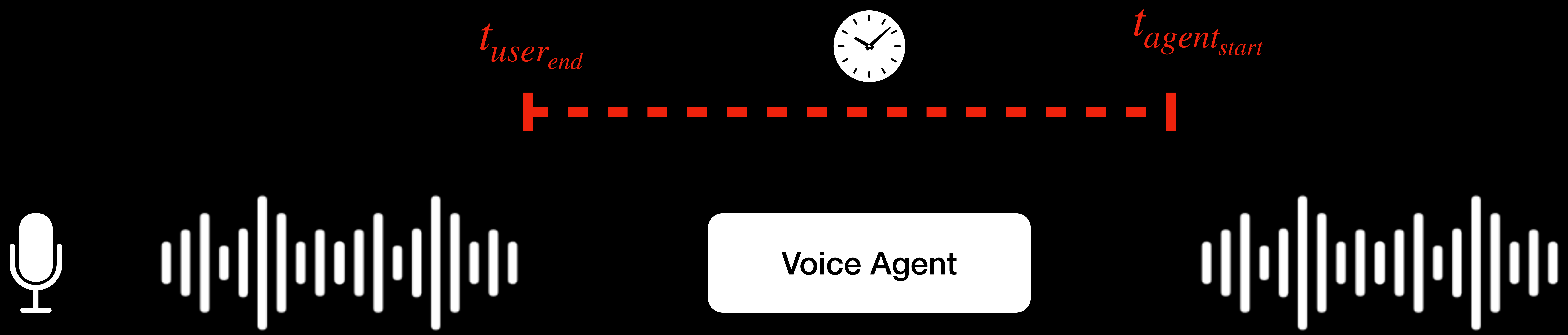
Unified Architecture



- Eliminate STT and TTS Models
- Model examples: Qwen-omni, Higgs-v2, Moshi

Latency

The minimal time delay between the completion of a user's spoken input and the initiation of the system's spoken response

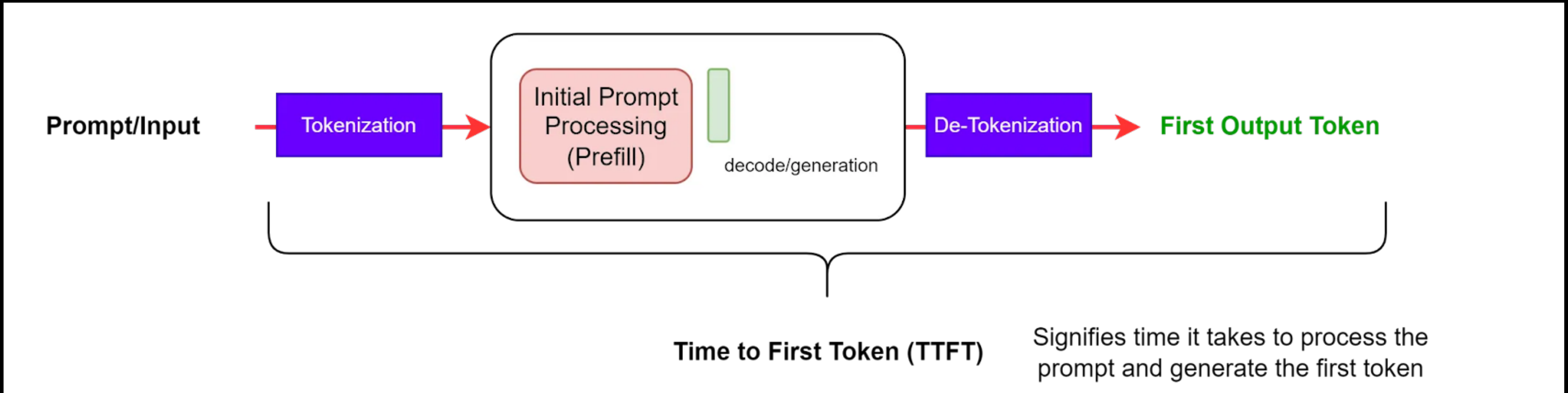


A widely accepted baseline target for good voice-to-voice latency in AI voice agents is approximately

800ms

Sub Latencies

- **Time To First Token (TTFT):** measures the elapsed time between submitting a prompt to the API and receiving the model's first generated token.



*** Used for LLM or STT**

Sub Latencies

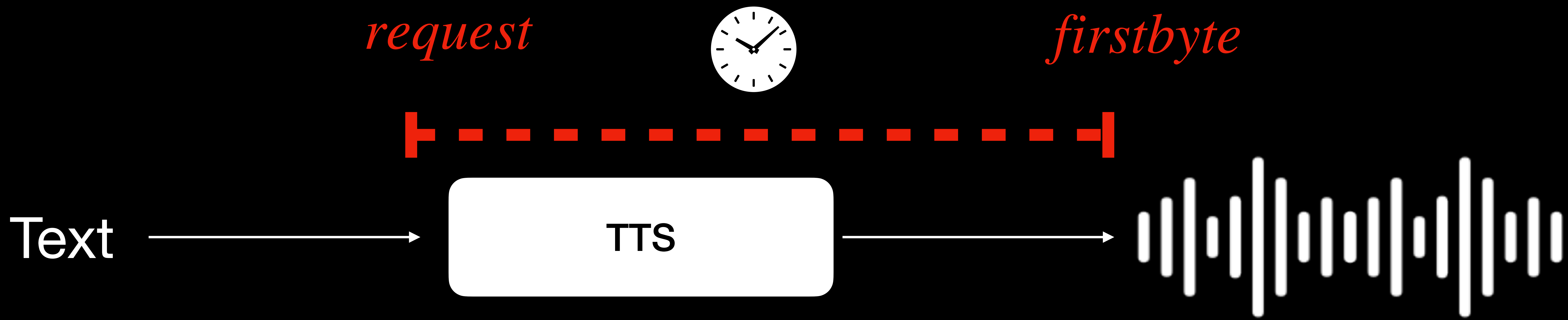
- Time To First Token (TTFT): measures the elapsed time between submitting a prompt to the API and receiving the model's first generated token.

FEATURES			INTELLIGENCE	PRICE	OUTPUT TOKENS/S	LATENCY
MODEL	CREATOR	CONTEXT WINDOW	ARTIFICIAL ANALYSIS INTELLIGENCE INDEX	BLENDED *USD/1M Tokens*	MEDIAN *Tokens/s*	MEDIAN *First Chunk (s)*
Gemini 2.5 Flash	Google	1m	53	$0.85	259.4	0.33
GPT-4.1 mini	OpenAI	1m	53	$0.70	78.9	0.40
GPT-4.1	OpenAI	1m	53	$3.50	121.7	0.48
Grok 3	xAI	1m	51	$6.00	63.5	0.71
Claude 4 Sonnet	ANTHROPIC	200k	53	$6.00	94.4	1.18
Claude 4 Opus	ANTHROPIC	200k	58	$30.00	59.7	2.00

https://artificialanalysis.ai/

Sub Latencies

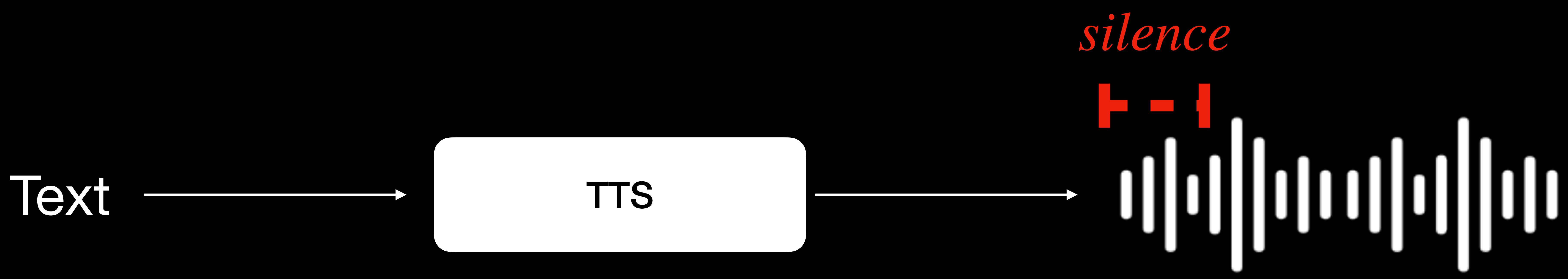
- Time To First Byte (TTFB): The duration between the request initiation and the arrival of the first byte of audio data.



*** Used for TTS**

Sub Latencies

- **Average pre-speech interval:** The mean duration of initial silence in the audio stream before the first speech frame is produced.



*** Used for TTS**

Best Practices

LLM Selection

LLM Is The Principal Component

- Effective instruction following
- Tool calling capabilities
- Low rate of hallucination
- Low latency (TTFT)
- Reasonable cost

STT to LLM to TTS Prompt

LLM should take into consideration that the input comes from the STT, and its output will be converted to speech

- **Handle potential transcription errors.**
- **Produce output that is well-suited for spoken delivery.**

Function Calling

The function call may take more time to answer, what's makes the latency high. We can avoid this by:

- **Outputting a waiting message when the function is executing.**
- **Play background music while executing long-running function calls.**
- **Performing Async Inference Tasks**

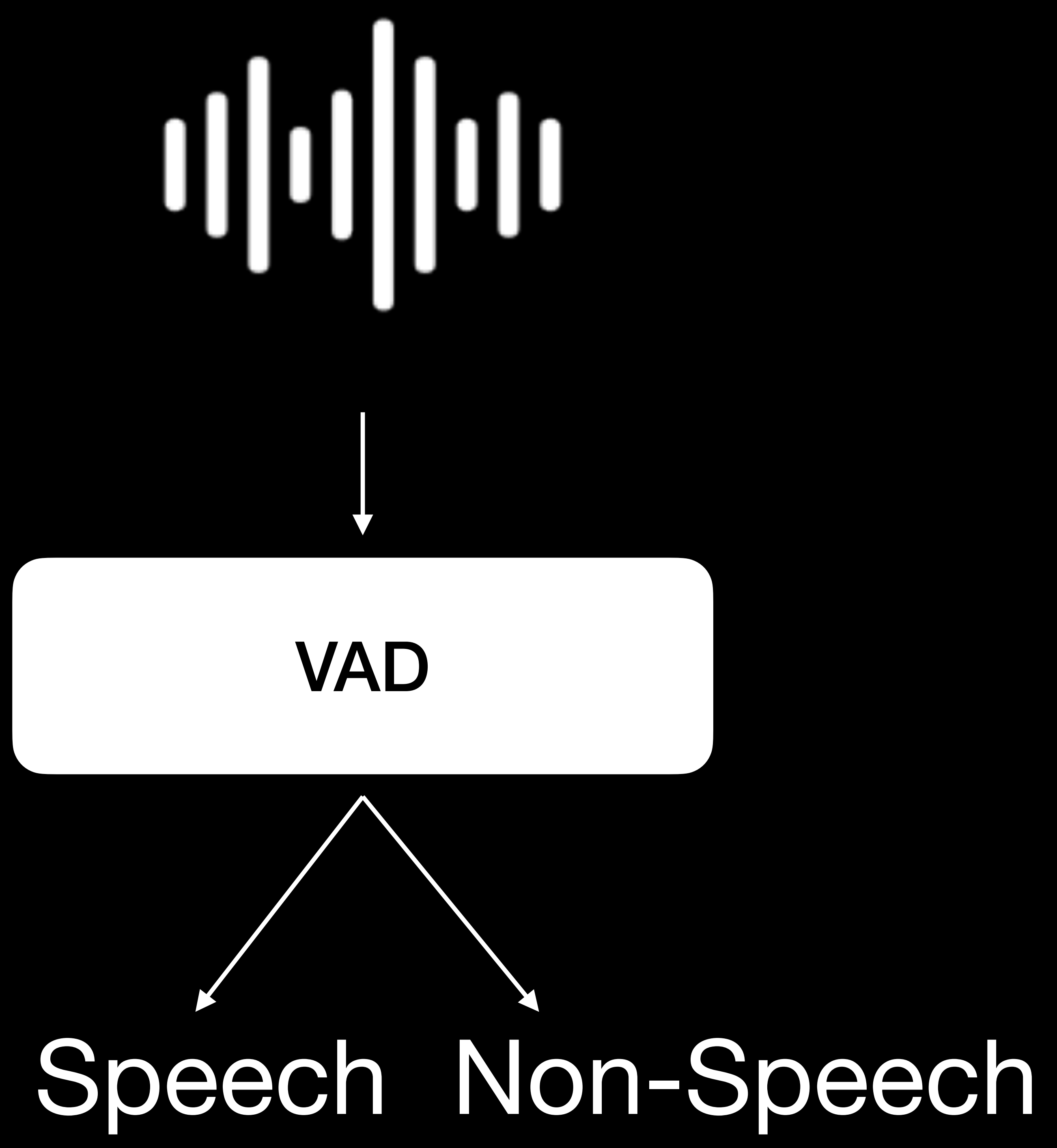
Noise Cancellation

Eliminating unwanted background noise

- **Real-time open-source model example: DeepFilterNet2**

Voice Activity Detection (VAD)

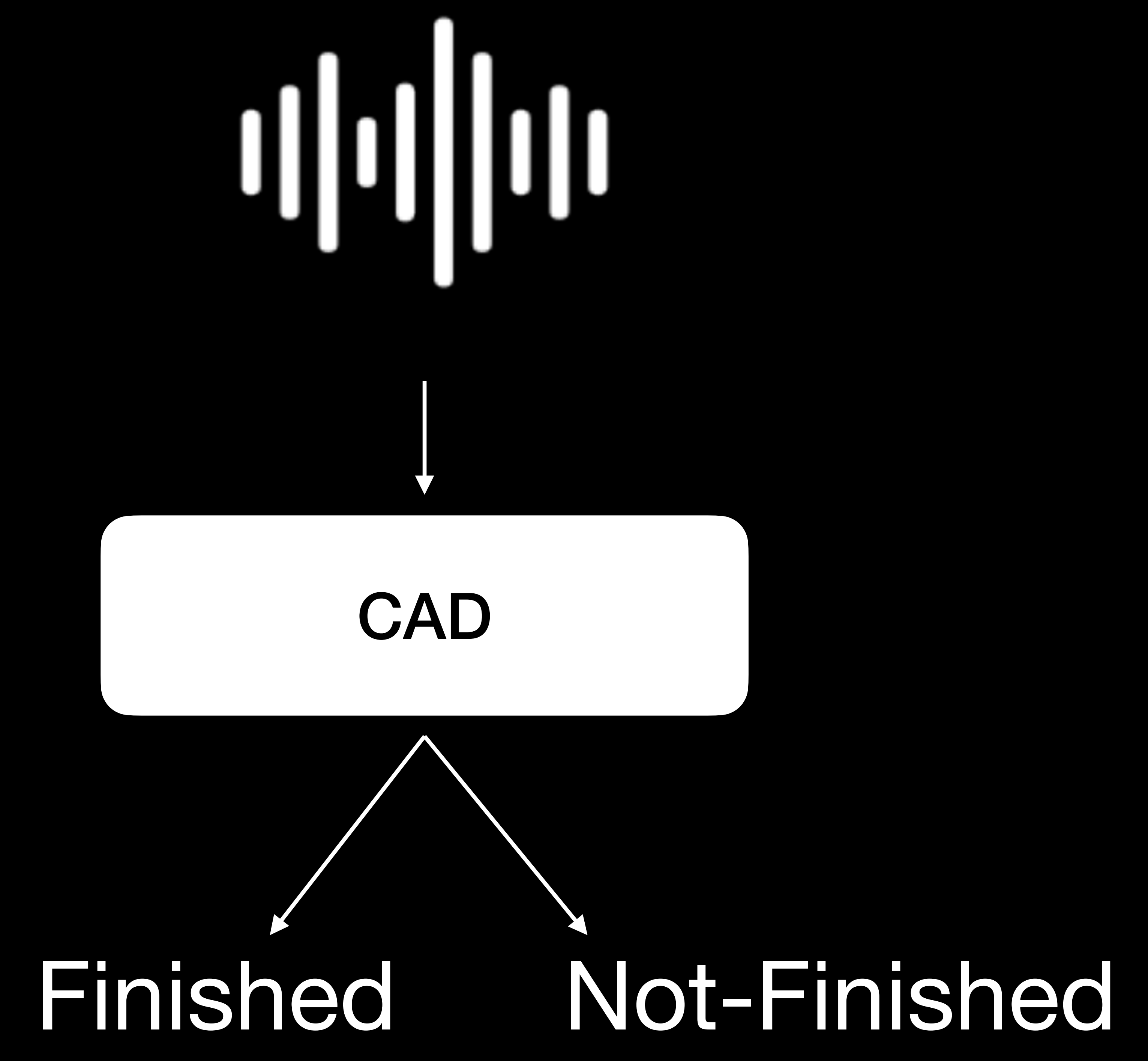
Detect the presence or absence of human speech



- **Real-time open-source model example: [Silero-VAD]**

Context-aware Turn Detection

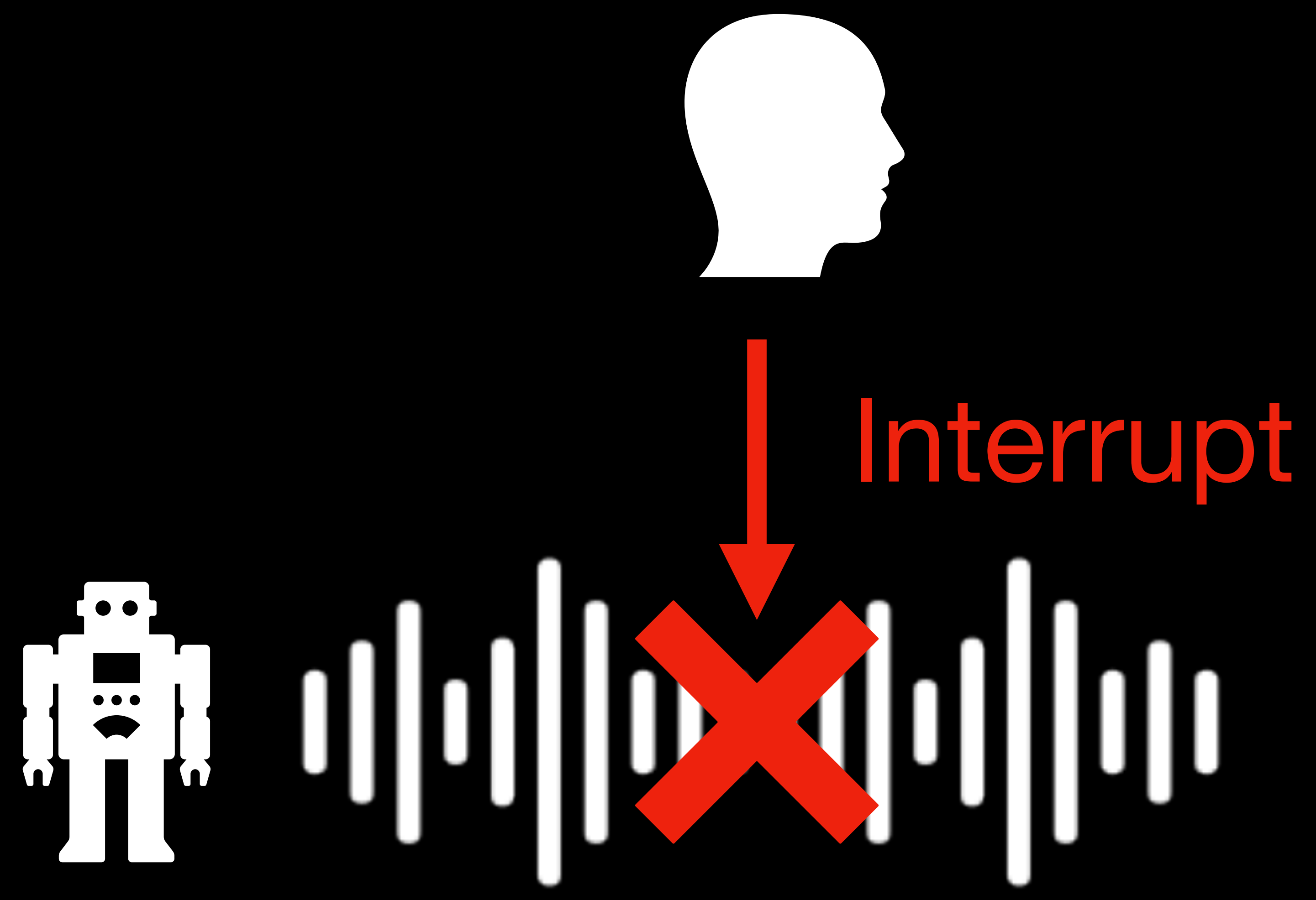
Semantic voice activity detection based on the context



- **Real-time open-source model example:** **Smart-turn-v2**

Interruption Handling

take care of what you will save as context



Network

WebRTC Vs WebSockets

WebRTC

- Built on UDP
- Used for building a browser voice agent
- Latency is important
- Comes with excellent echo cancellation and noise reduction

WebSockets

- Built on TCP
- Great for server-to-server cases.
- Latency is not important

Frameworks

LiveKit

CODE…

Code/Slides

Linkedin

Join Atlasia

Resources For Learning

THANK YOU…❤️