Stream Everything

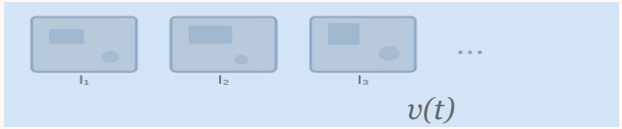
Moving from Request input to Streaming input

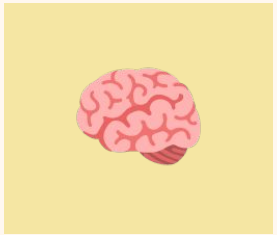
Patrick von Platen, Mistral AI

How humans process the world

Audio (w):

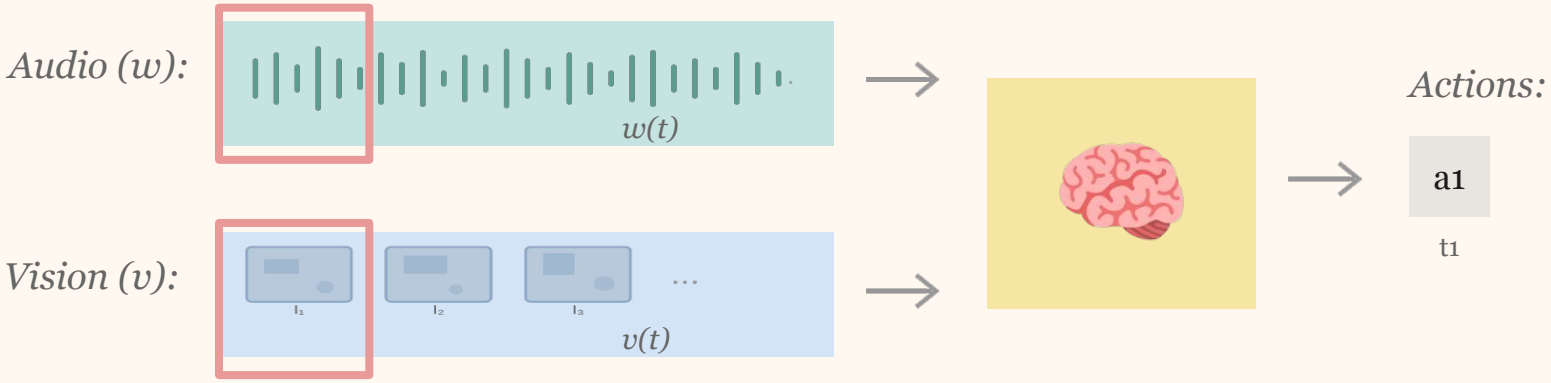
Vision (v):





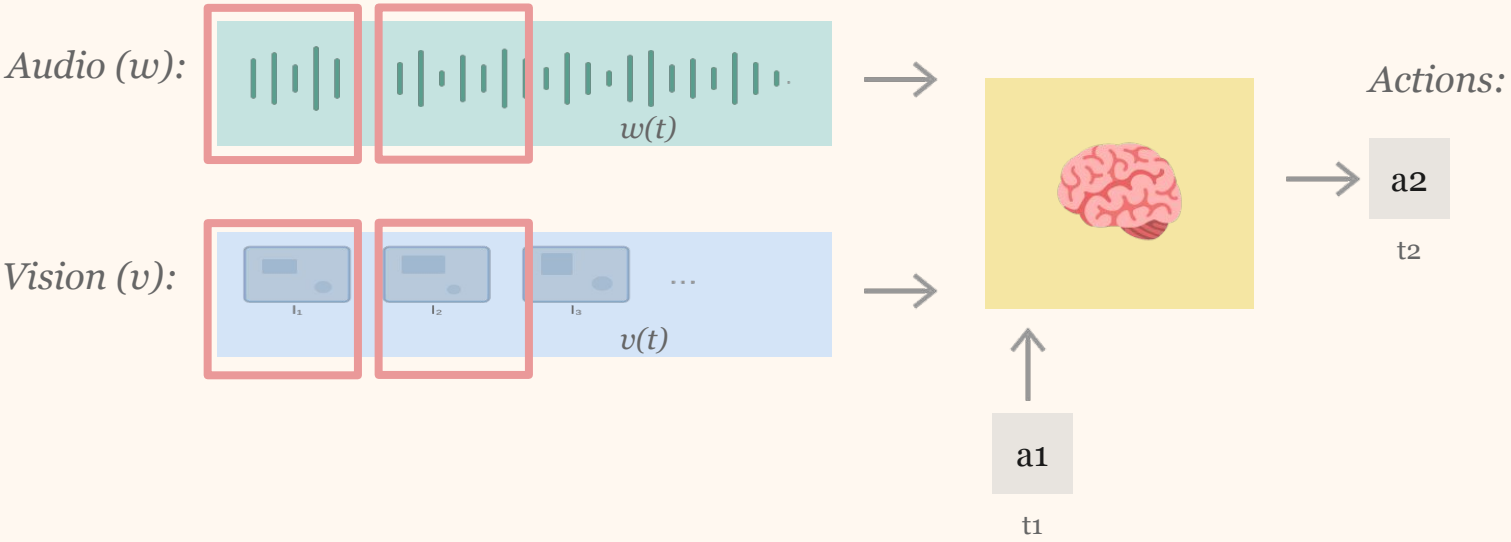
Confidential ©2026 Mistral AI. All rights reserved.

How humans process the world



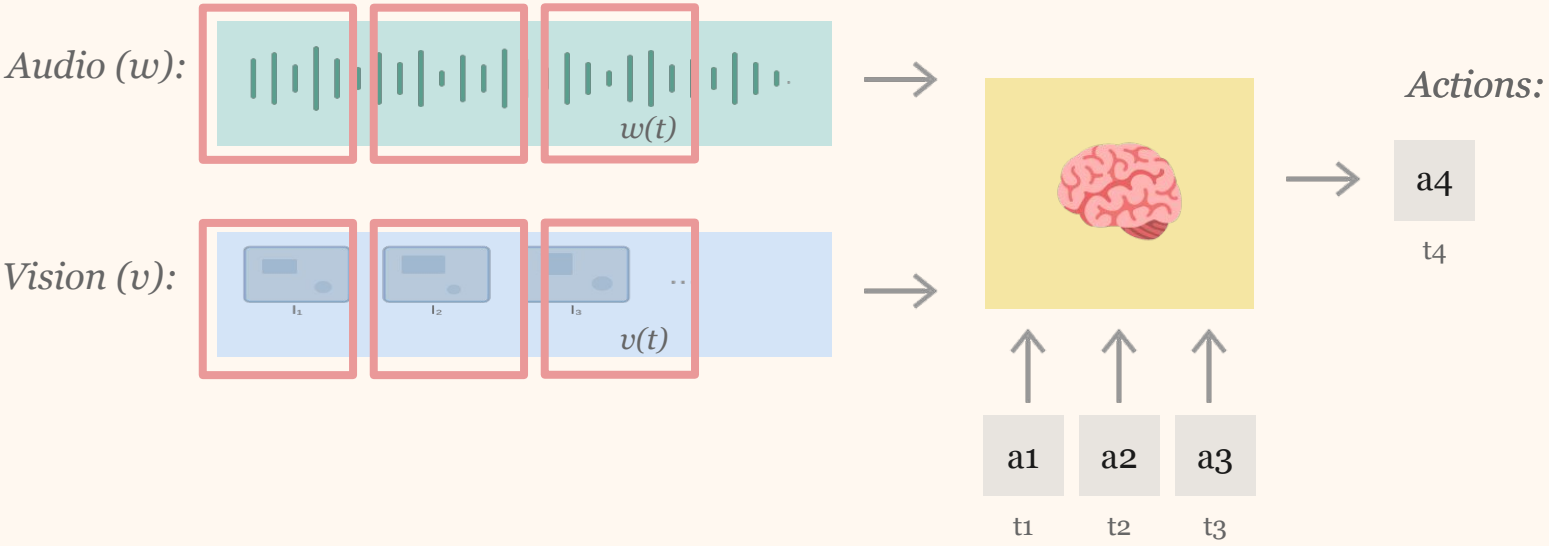
Confidential ©2026 Mistral AI. All rights reserved.

How humans process the world



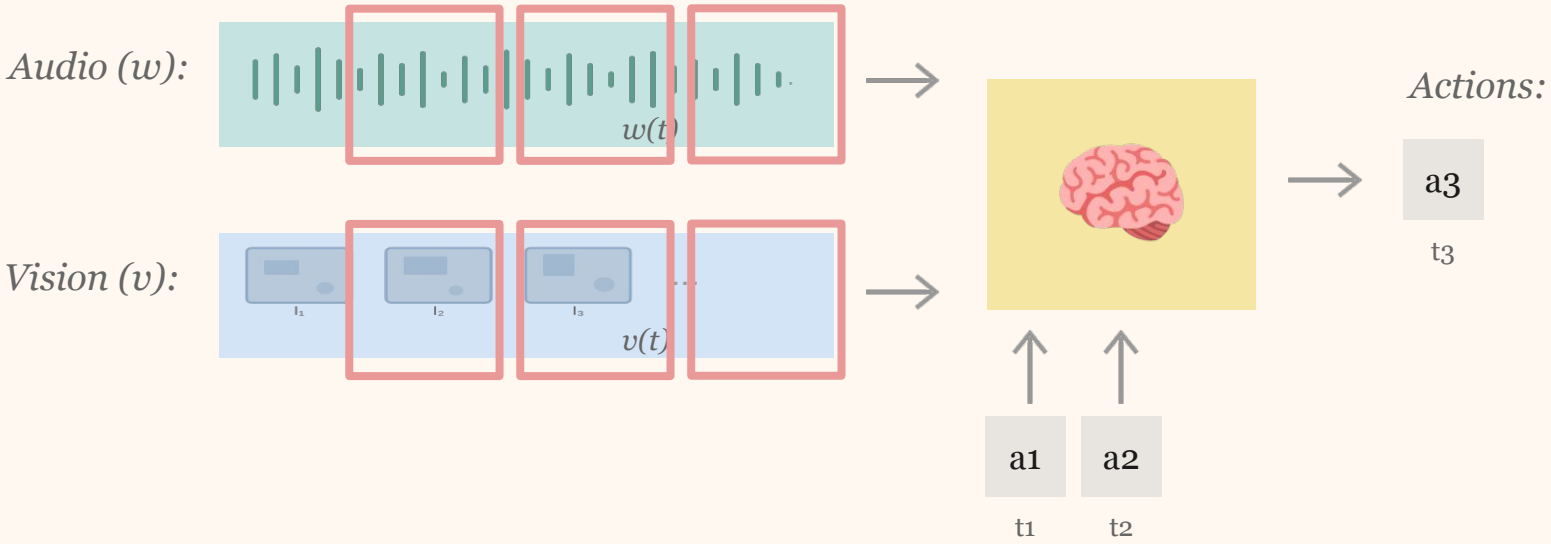
Confidential ©2026 Mistral AI. All rights reserved.

How humans process the world



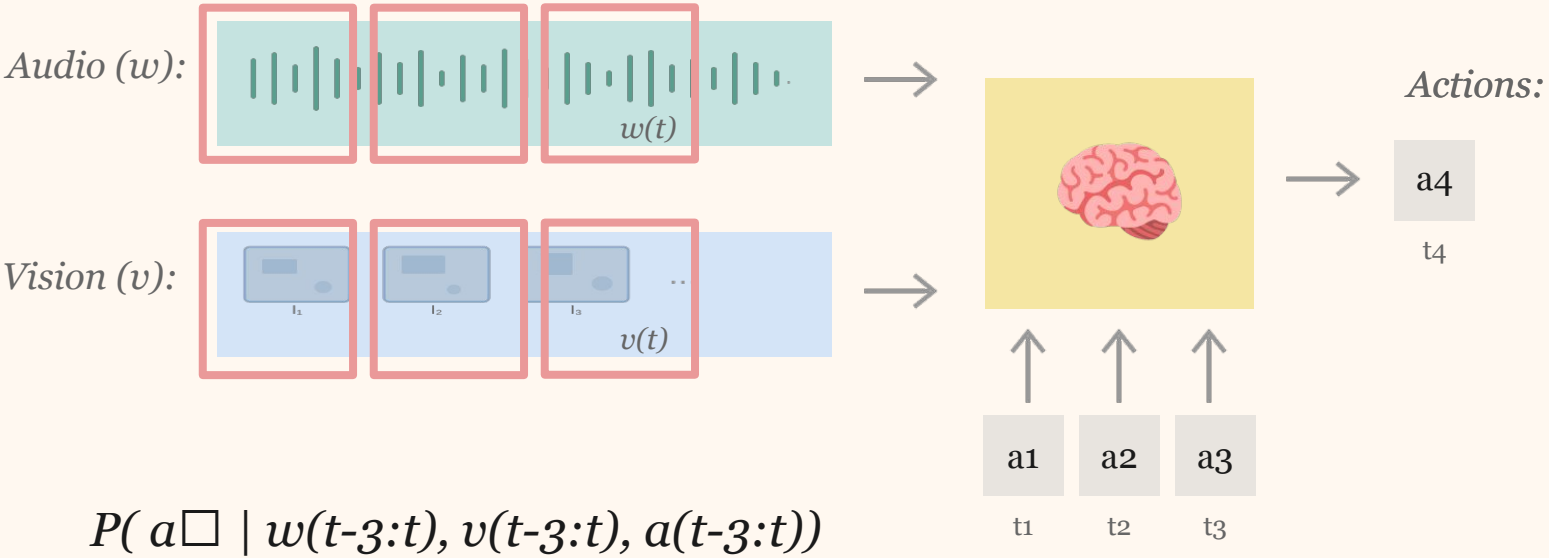
Confidential ©2026 Mistral AI. All rights reserved.

How humans process the world



Confidential ©2026 Mistral AI. All rights reserved.

How humans process the world



$$P(a\square \mid w(t\text{-}3{:}t), v(t\text{-}3{:}t), a(t\text{-}3{:}t))$$

At every time step t, an action a is taken based on:

- *Audio input stream*
- *Visual input stream*
- *Previous actions*

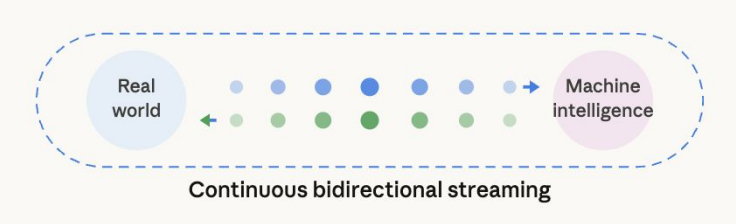
Confidential ©2026 Mistral AI. All rights reserved.

Why does that matter for LLMs?

Confidential ©2026 Mistral AI. All rights reserved.

Why does that matter for LLMs?

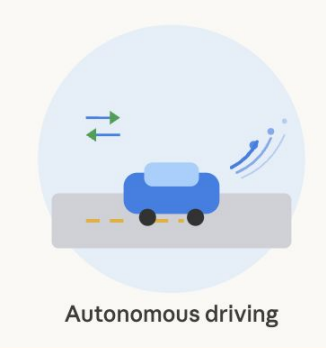
Flawless real world <> machine interaction needs streaming!



Confidential ©2026 Mistral AI. All rights reserved.

Why does that matter for LLMs?

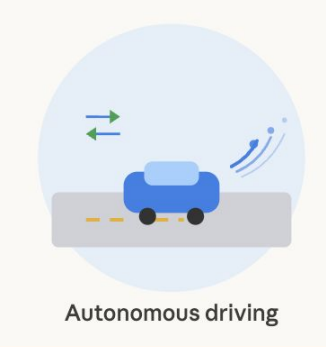
Flawless real world <> machine interaction needs streaming!



Confidential ©2026 Mistral AI. All rights reserved.

Why does that matter for LLMs?

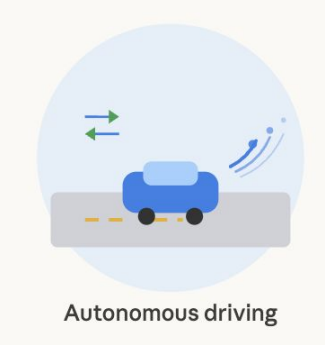
Flawless real world <> machine interaction needs streaming!

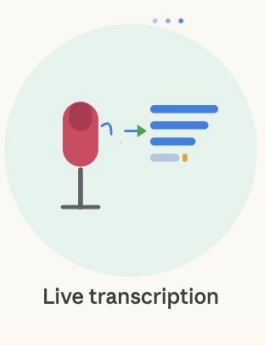


Confidential ©2026 Mistral AI. All rights reserved.

Why does that matter for LLMs?

Flawless real world <> machine interaction needs streaming!

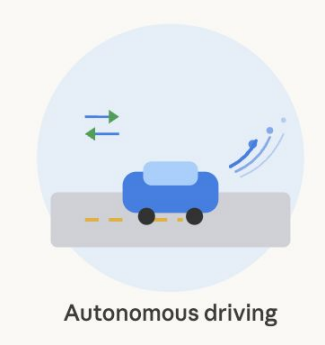


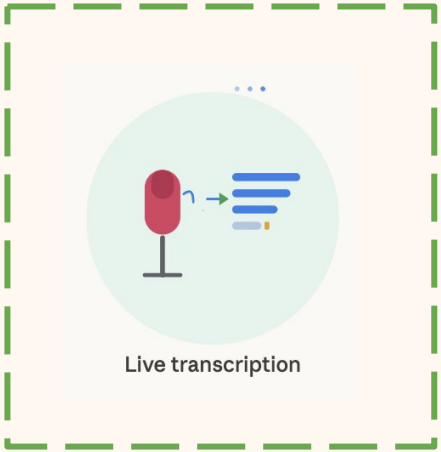


Confidential ©2026 Mistral AI. All rights reserved.

Why does that matter for LLMs?

Flawless real world <> machine interaction needs streaming!





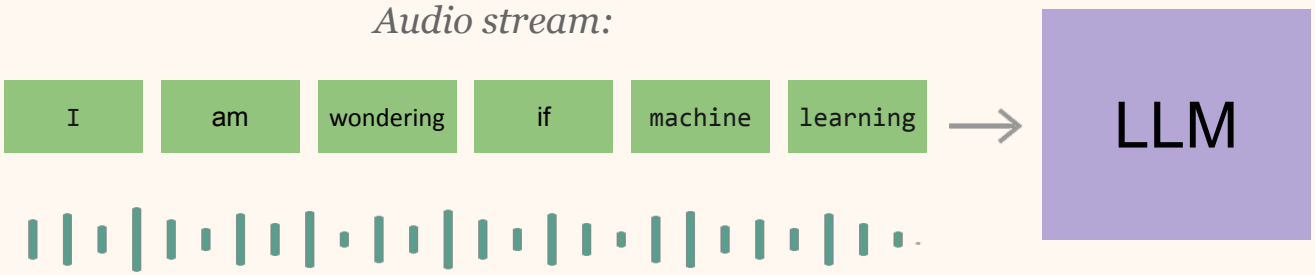
Confidential ©2026 Mistral AI. All rights reserved.

Voxtral Realtime

Abstract

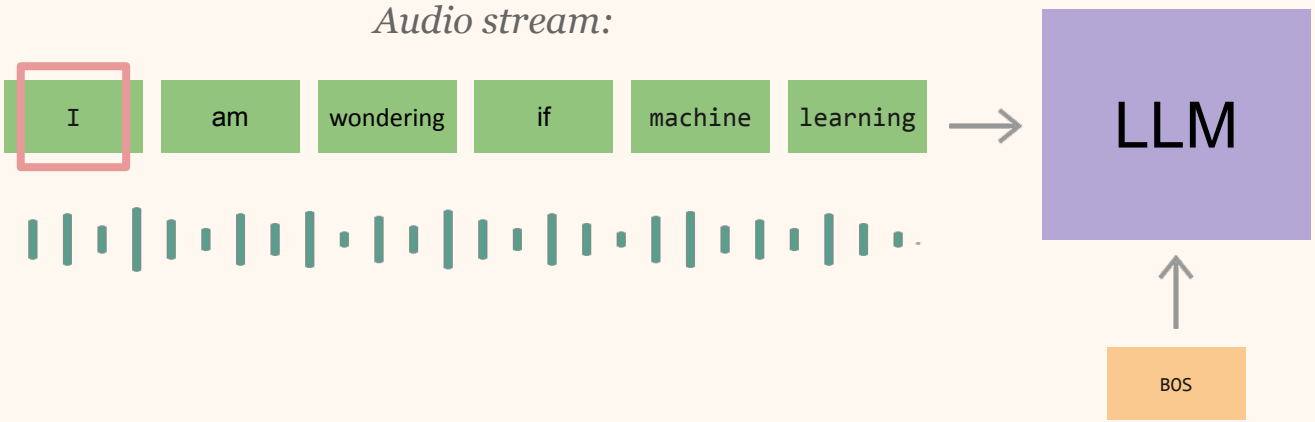
We introduce Voxtral Realtime, a natively streaming automatic speech recognition

Speech Recognition - Realtime



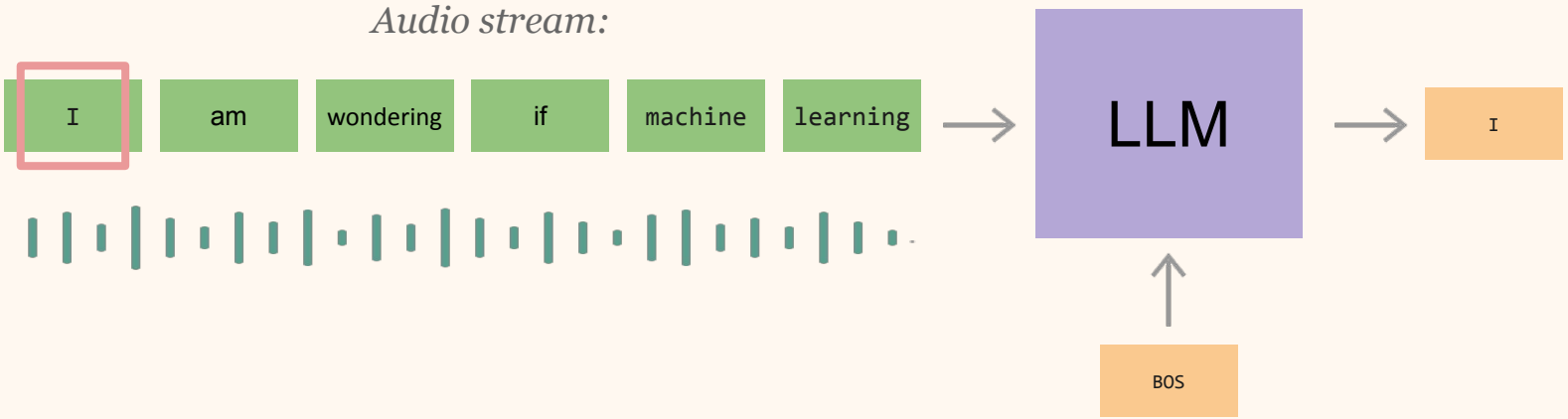
Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime



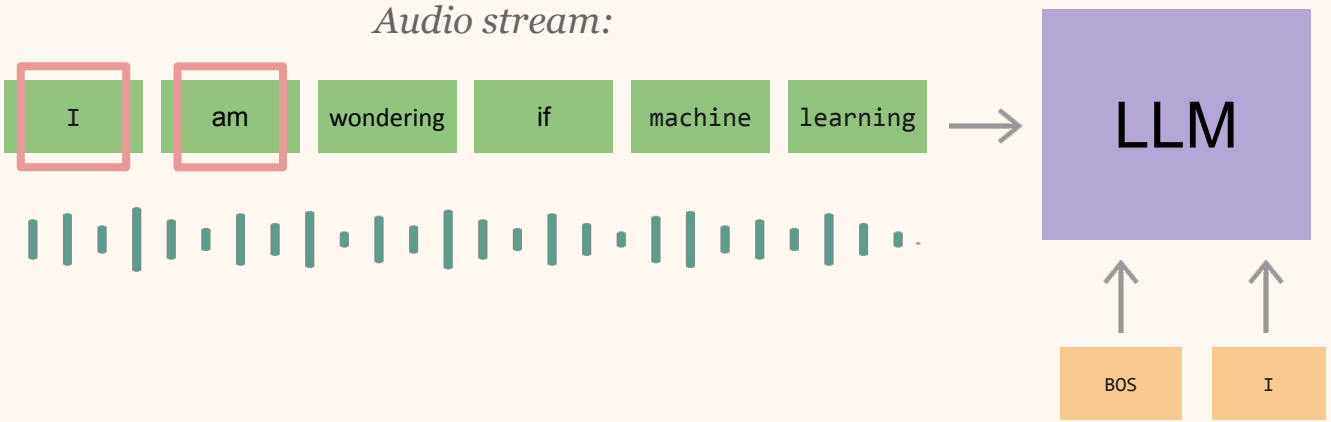
Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime



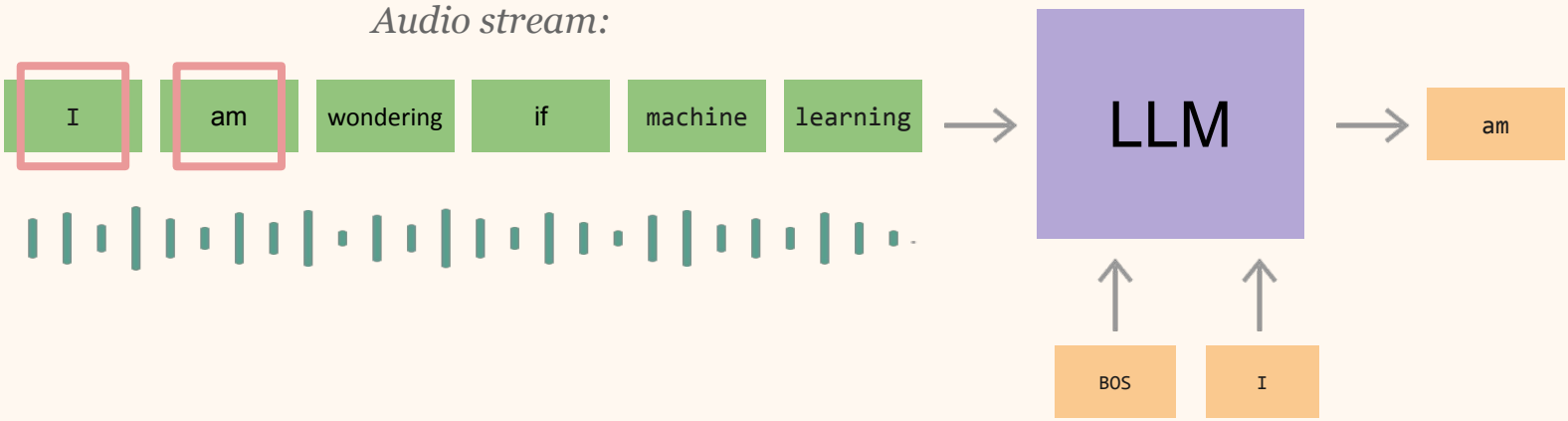
Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime



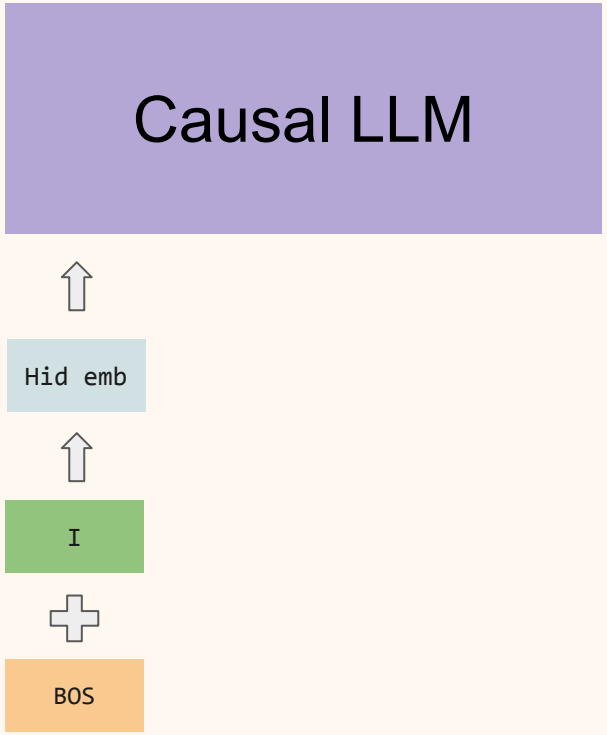
Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime



Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime

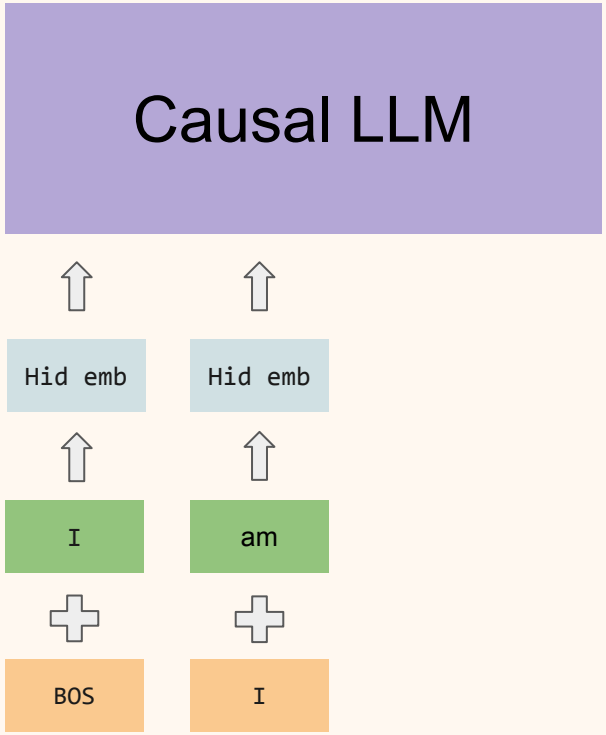


Audio emb

Word emb

Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime

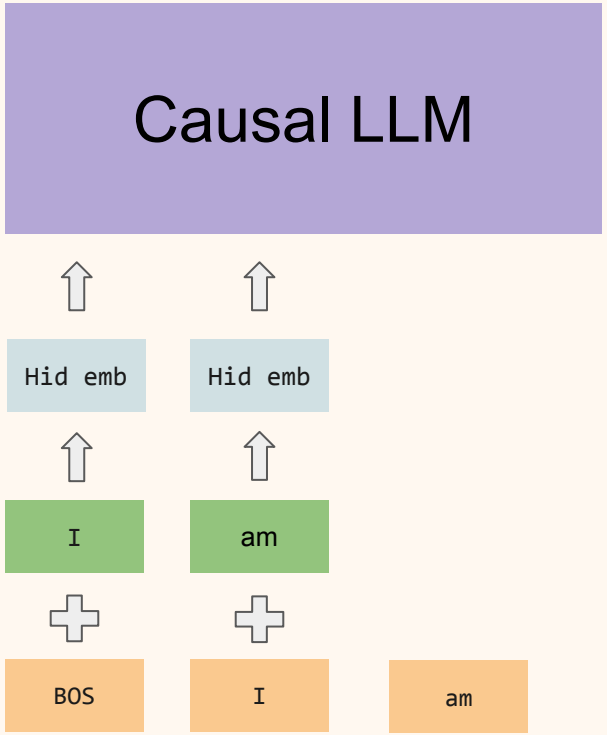


Audio emb

Word emb

Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime



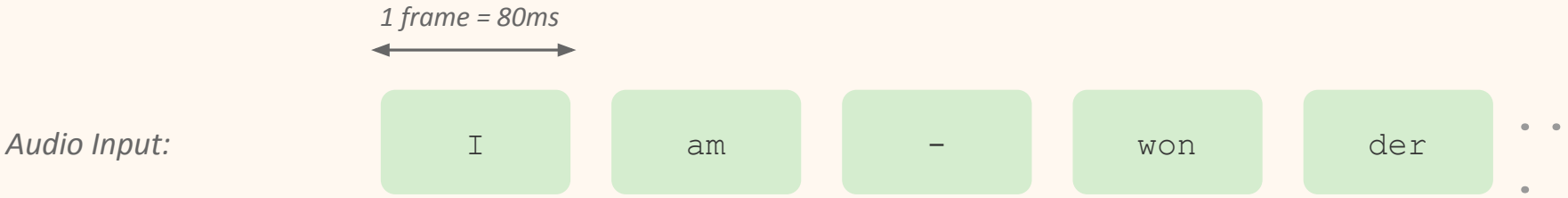
Audio emb

Word emb

Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime

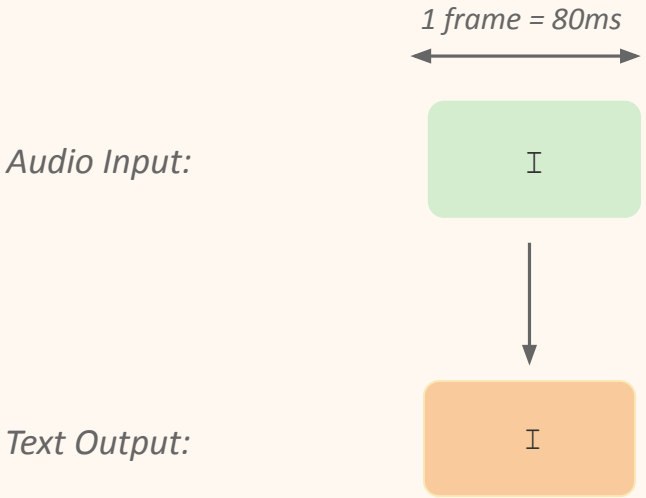
Audio is moving faster than text... what to predict?



Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime

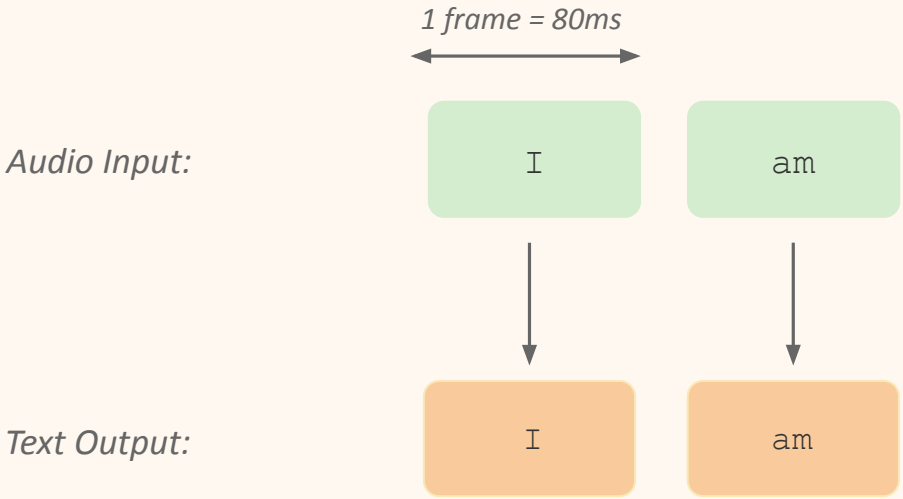
Audio is moving faster than text... what to predict?



Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime

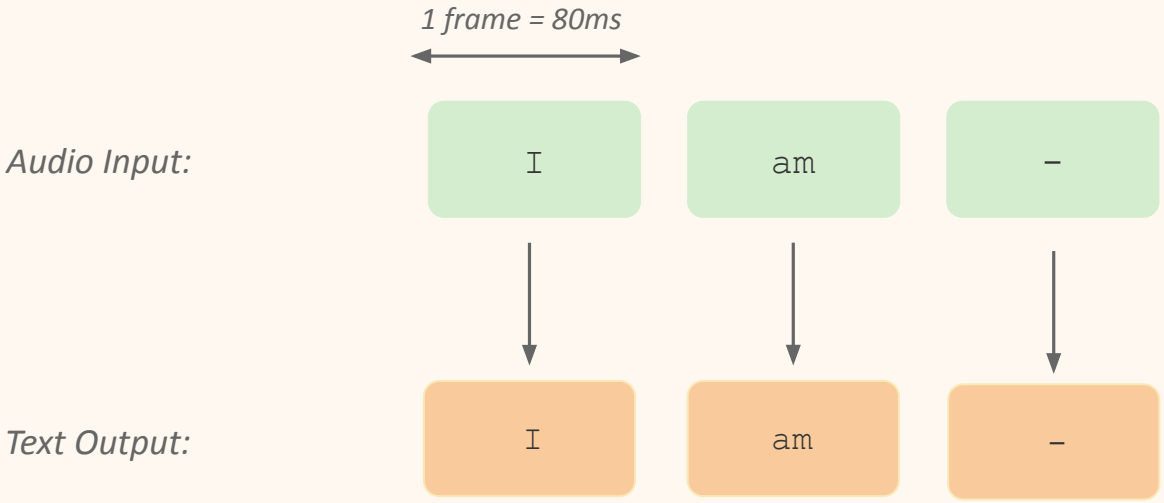
Audio is moving faster than text... what to predict?



Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime

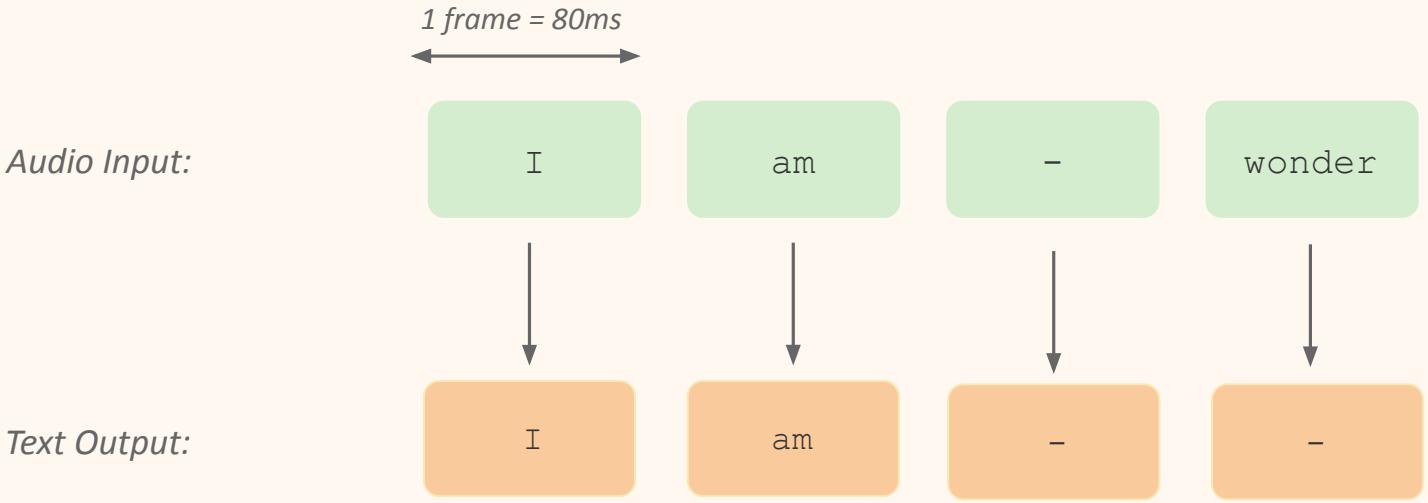
Audio is moving faster than text... what to predict?



Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime

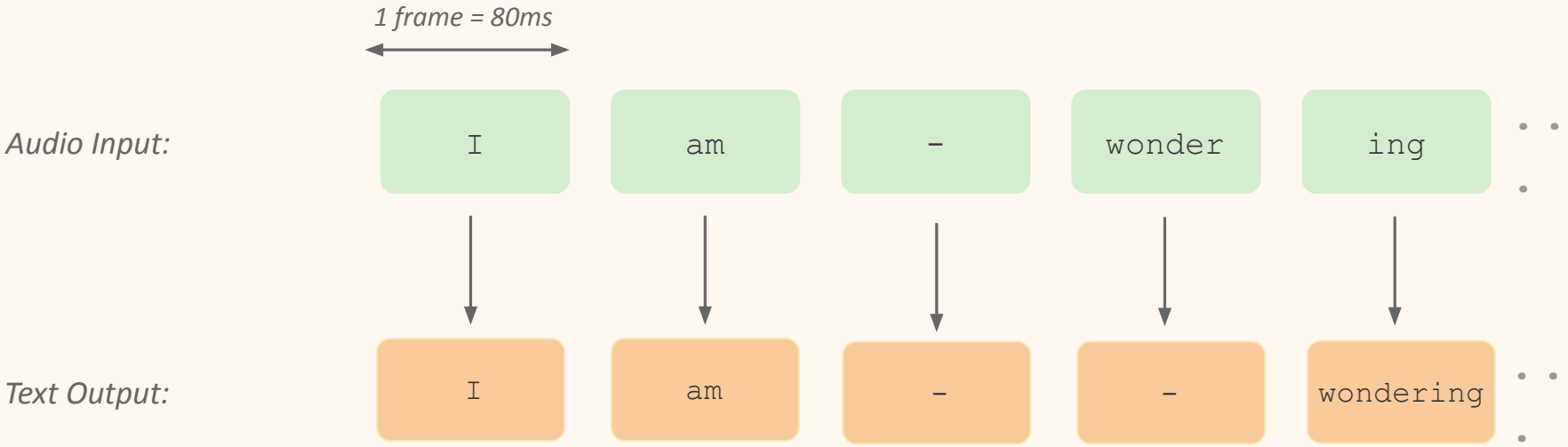
Audio is moving faster than text... what to predict?



Confidential ©2026 Mistral AI. All rights reserved.

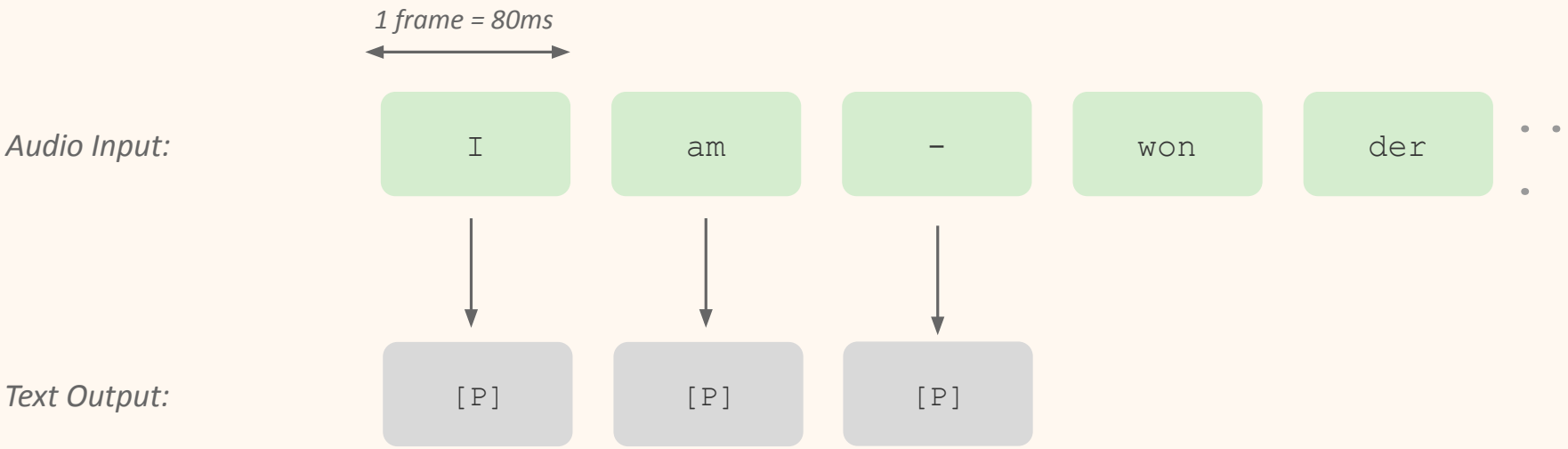
Speech Recognition - Realtime

Audio is moving faster than text... what to predict?



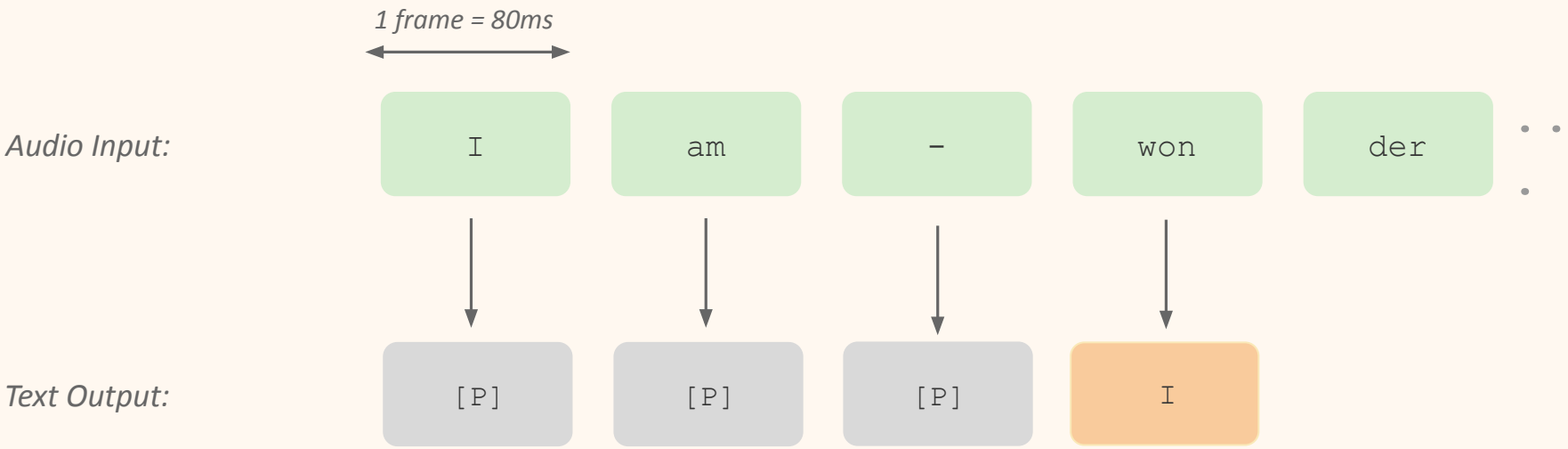
Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime



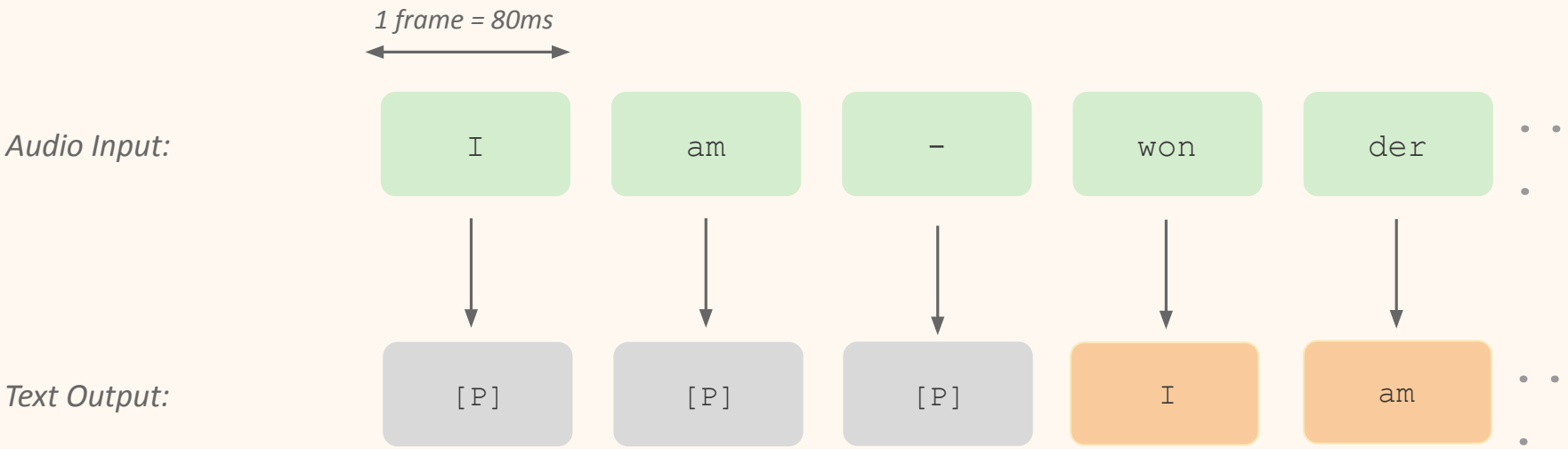
Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime



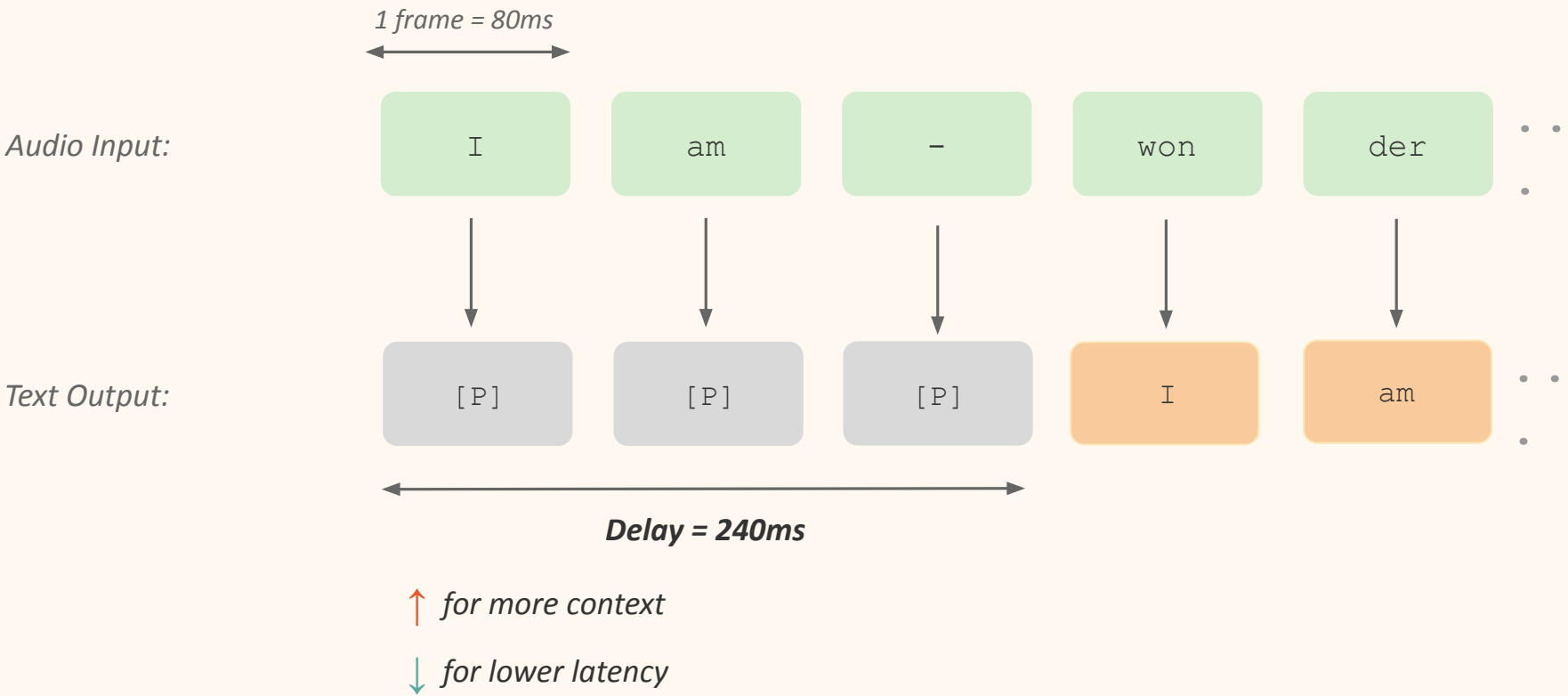
Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime



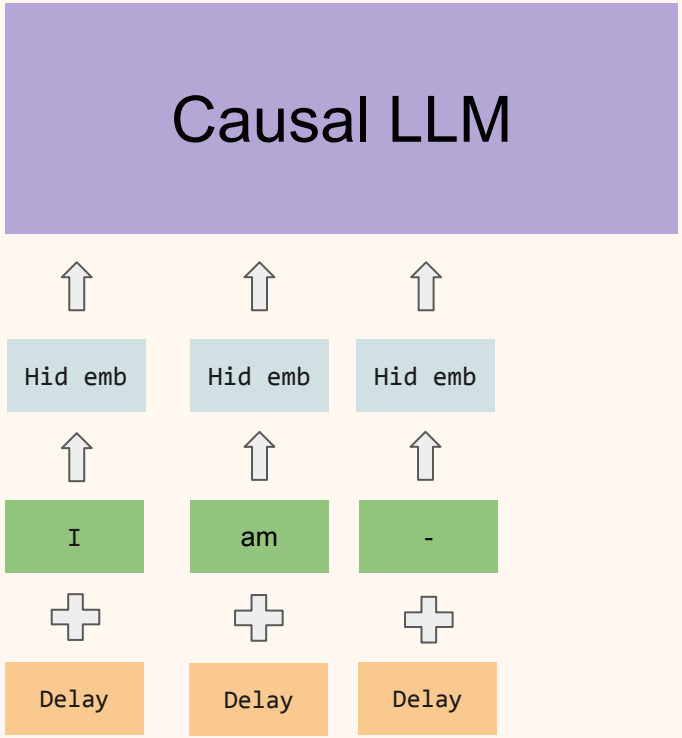
Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime



Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime

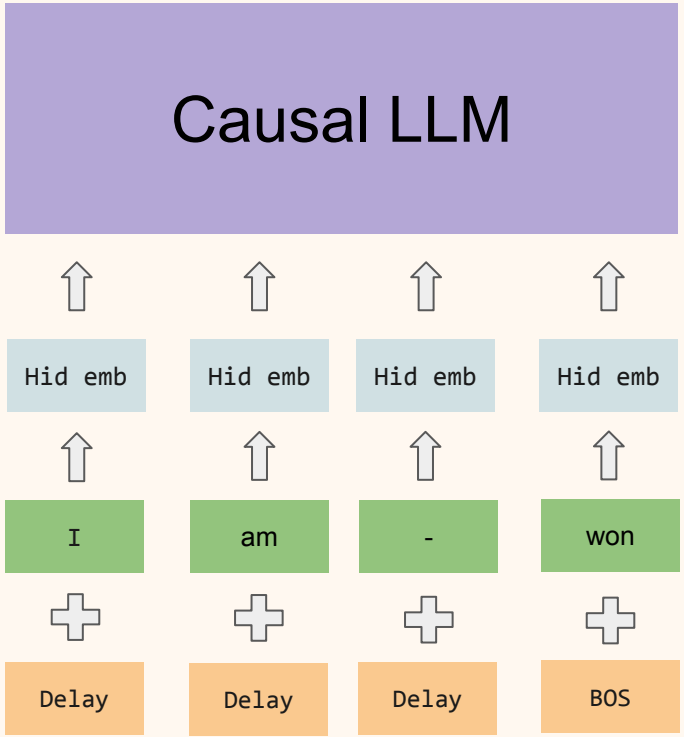


Audio emb

Word emb

Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime

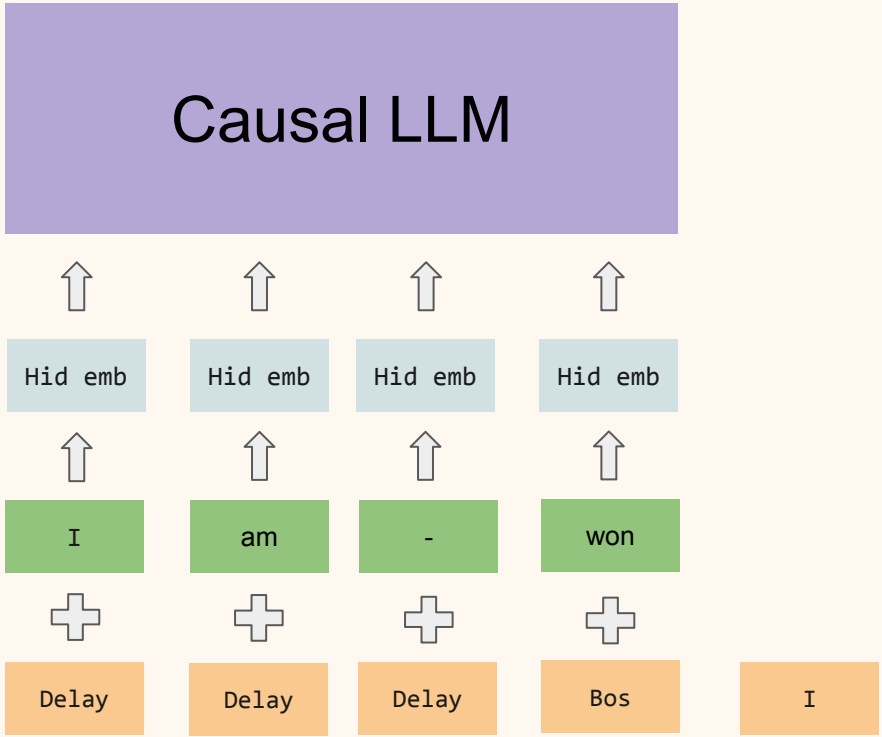


Audio emb

Word emb

Confidential ©2026 Mistral AI. All rights reserved.

Speech Recognition - Realtime

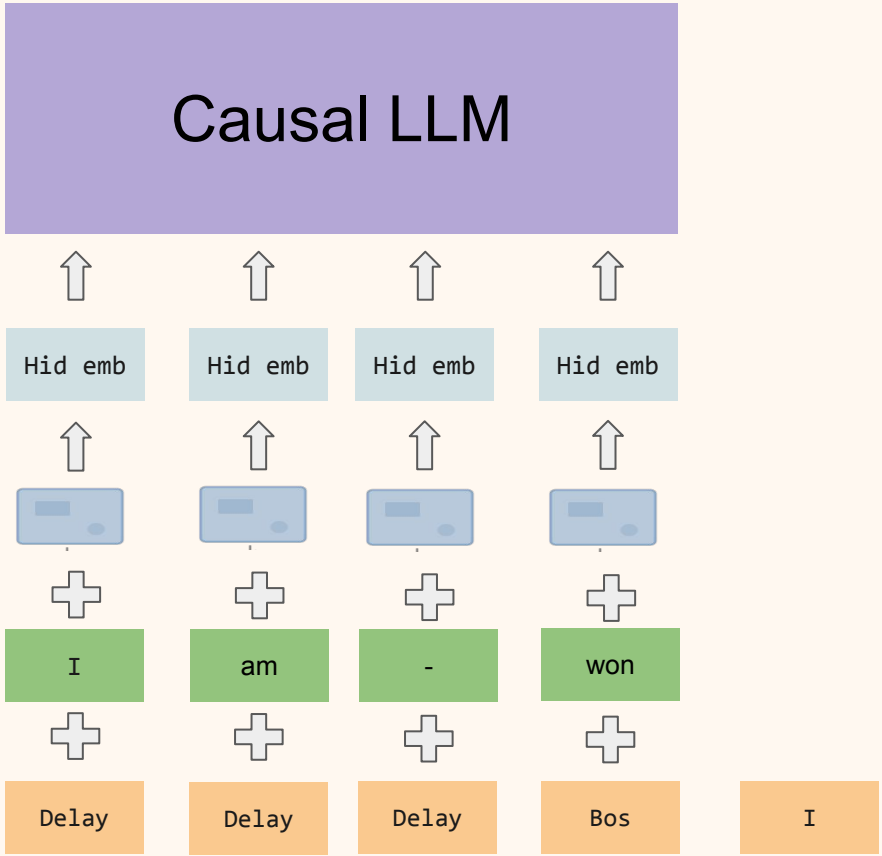


Audio emb

Word emb

Confidential ©2026 Mistral AI. All rights reserved.

Speech & Video - Realtime



Video emb

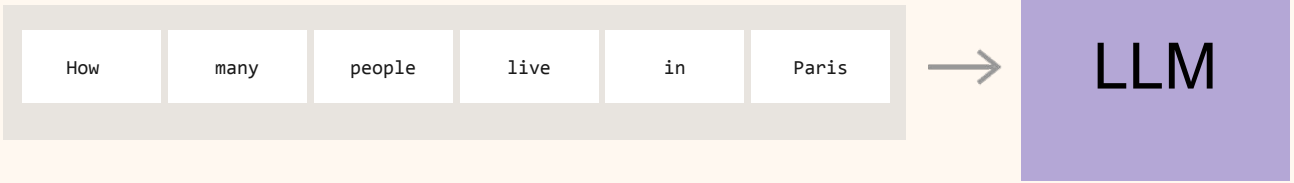
Audio emb

Word emb

Confidential ©2026 Mistral AI. All rights reserved.

How LLMs currently work

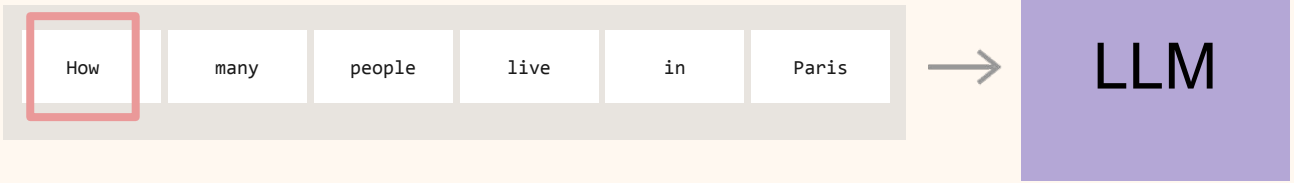
Request:



Confidential ©2026 Mistral AI. All rights reserved.

How LLMs currently work

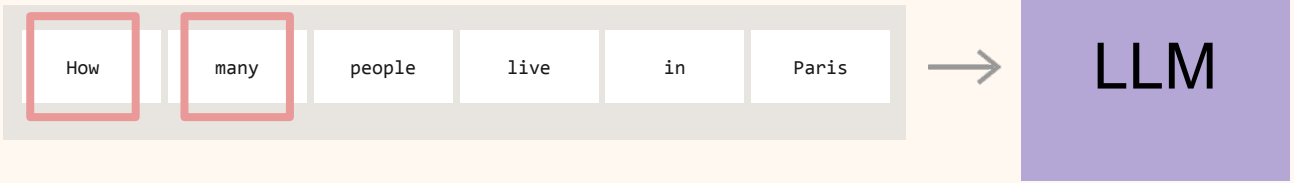
Request:



Confidential ©2026 Mistral AI. All rights reserved.

How LLMs currently work

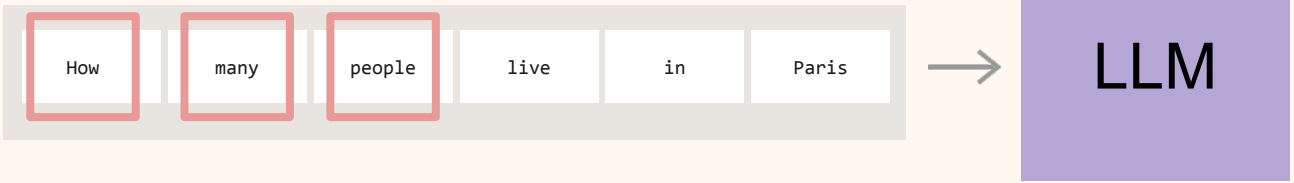
Request:



Confidential ©2026 Mistral AI. All rights reserved.

How LLMs currently work

Request:



Confidential ©2026 Mistral AI. All rights reserved.

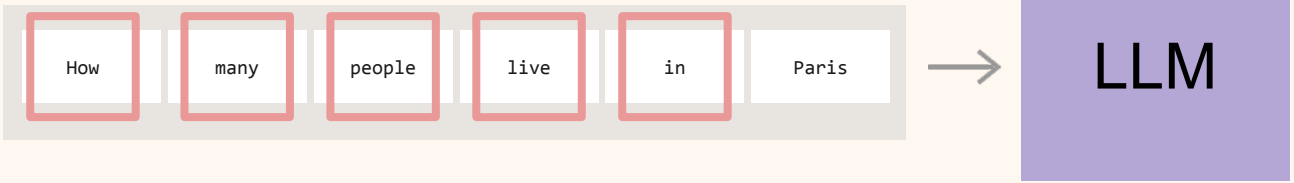
How LLMs currently work

Request:

Confidential ©2026 Mistral AI. All rights reserved.

How LLMs currently work

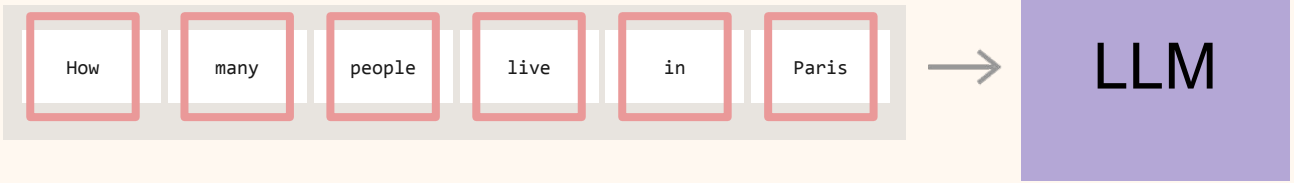
Request:



Confidential ©2026 Mistral AI. All rights reserved.

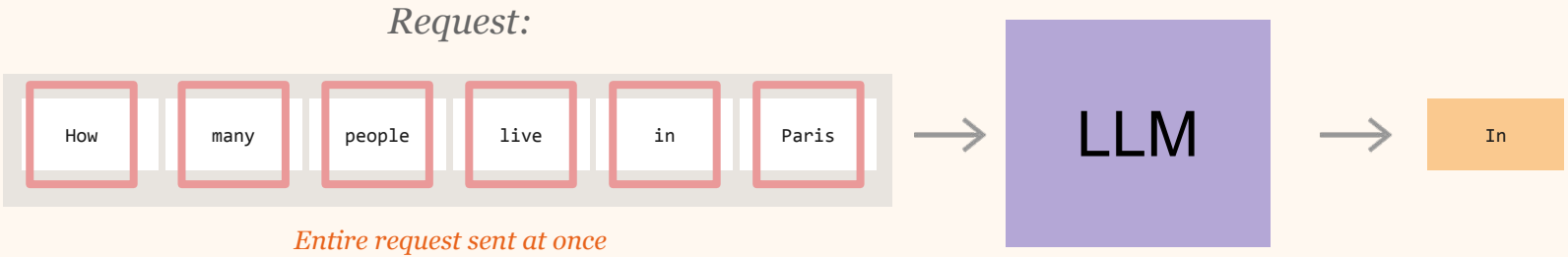
How LLMs currently work

Request:



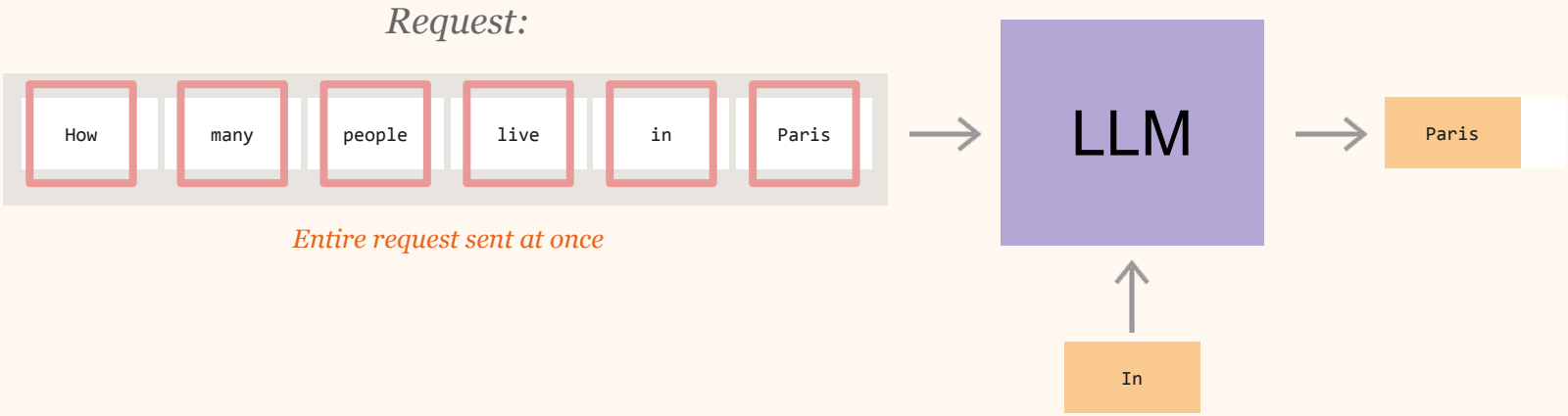
Confidential ©2026 Mistral AI. All rights reserved.

How LLMs currently work



Confidential ©2026 Mistral AI. All rights reserved.

How LLMs currently work

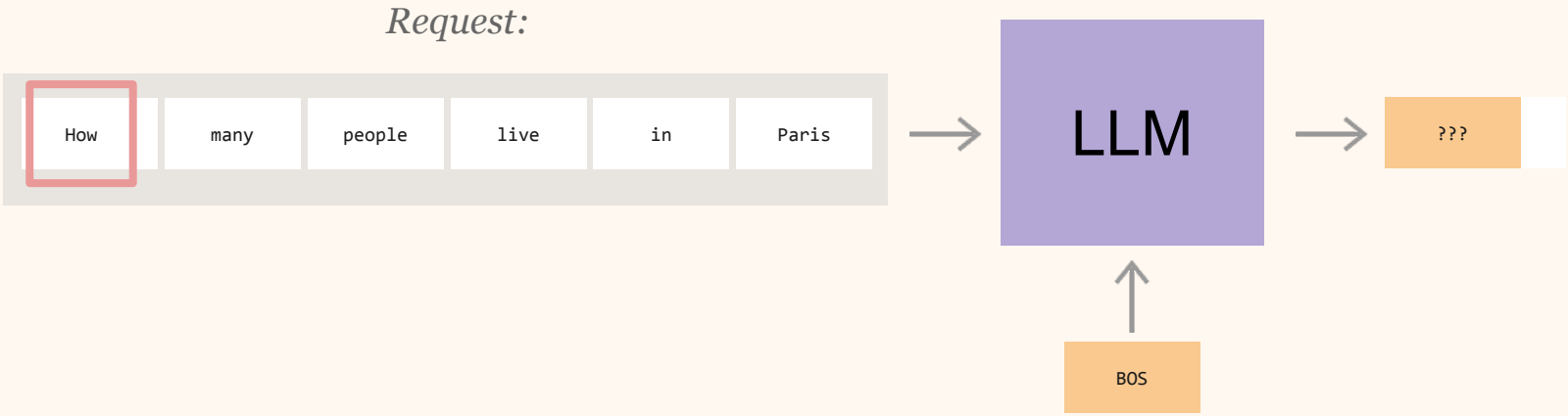


Confidential ©2026 Mistral AI. All rights reserved.

Can we stream LLMs?

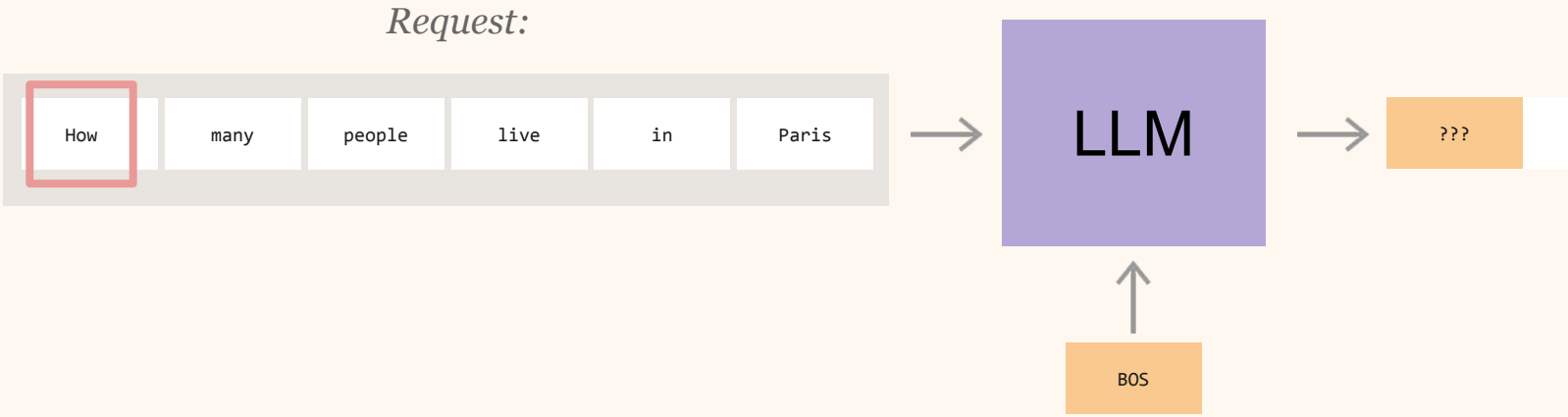
Confidential ©2026 Mistral AI. All rights reserved.

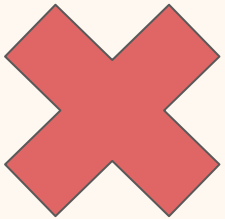
Can we stream LLMs?



Confidential ©2026 Mistral AI. All rights reserved.

Can we stream LLMs?



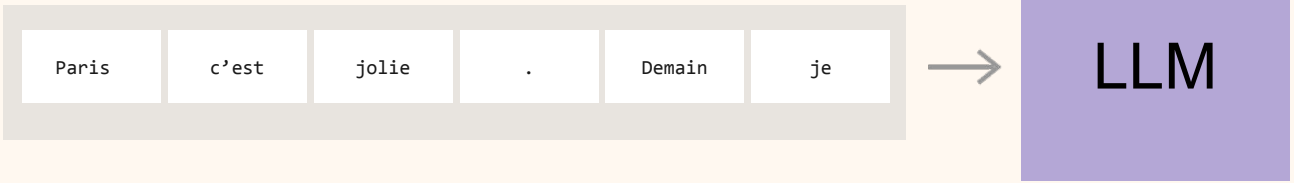


What should we predict from just “How”?

Confidential ©2026 Mistral AI. All rights reserved.

Live Translation

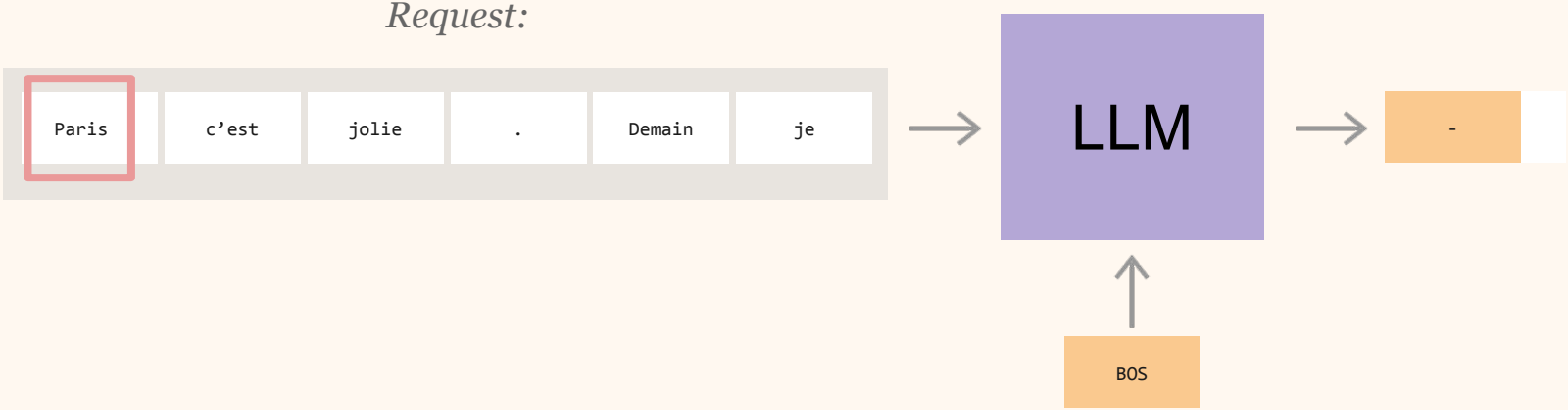
Request:



Confidential ©2026 Mistral AI. All rights reserved.

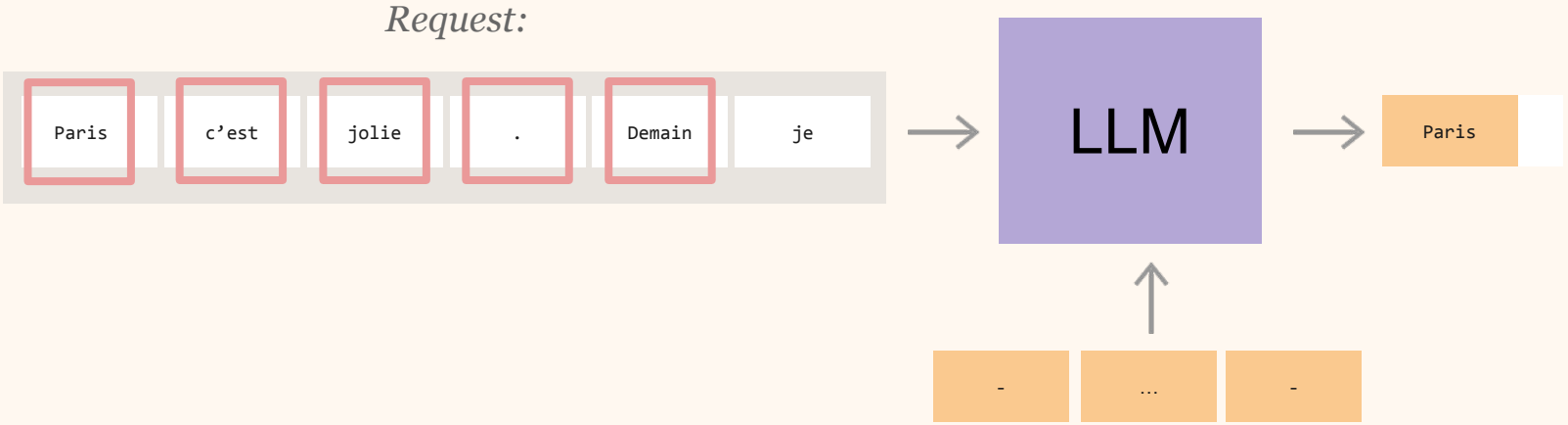
Live Translation

Request:



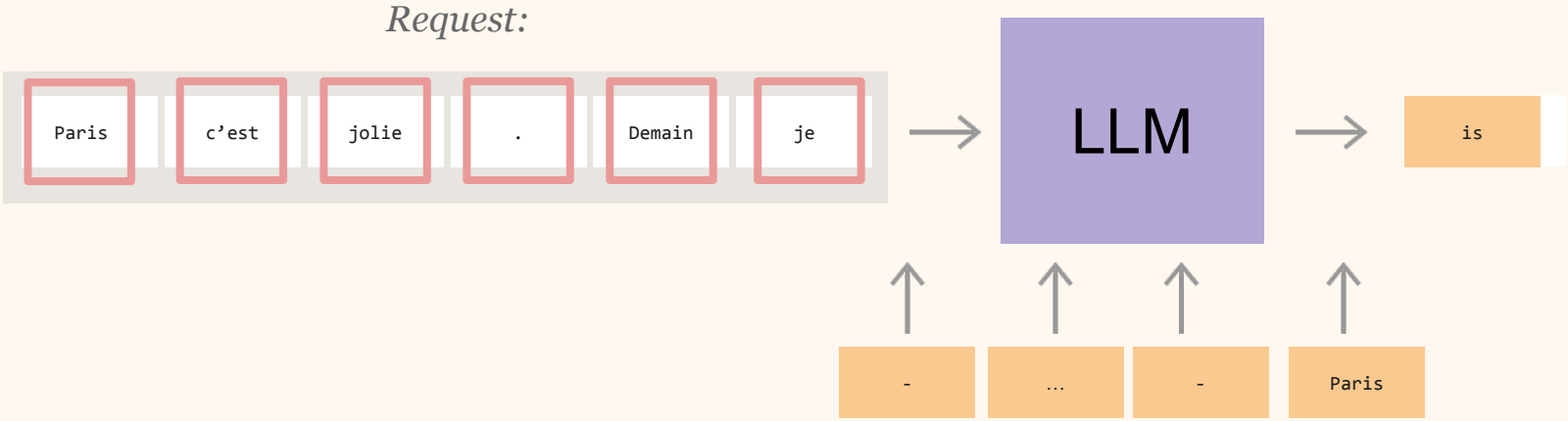
Confidential ©2026 Mistral AI. All rights reserved.

Live Translation



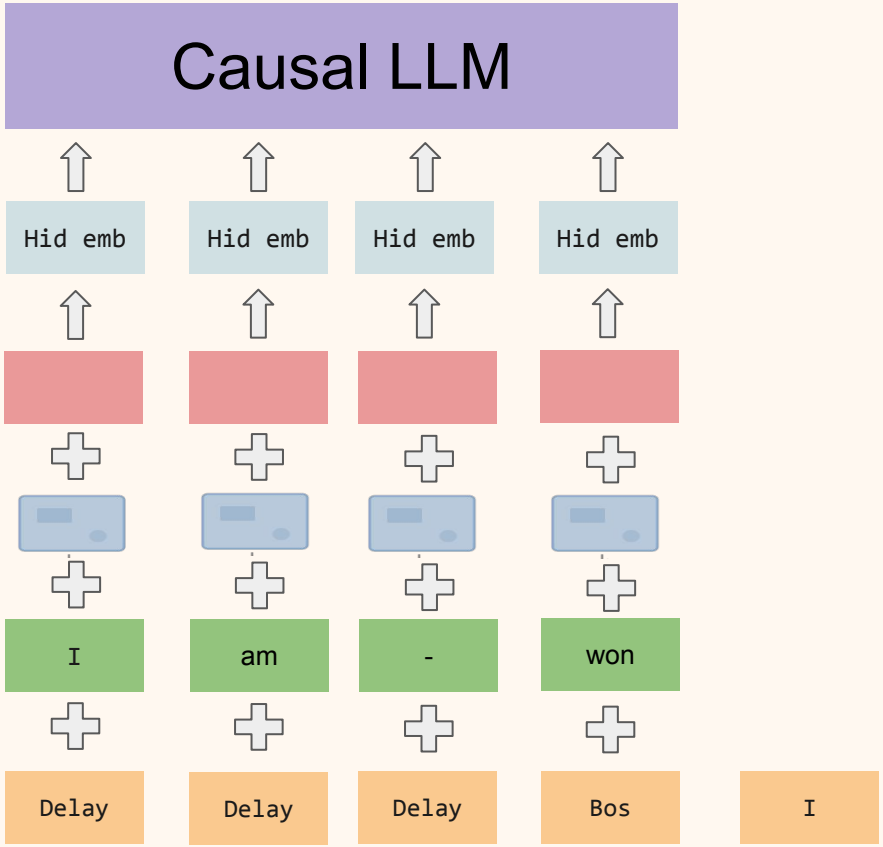
Confidential ©2026 Mistral AI. All rights reserved.

Live Translation



Confidential ©2026 Mistral AI. All rights reserved.

Speech & Video & Text - Realtime



In Word emb

Video emb

Audio emb

Out Word emb

Confidential ©2026 Mistral AI. All rights reserved.

Practical considerations

Confidential ©2026 Mistral AI. All rights reserved.

Practical considerations

Difficult in practice:

- How to pretrain? Need input & output pair
- Input & output pair is not necessarily aligned (log parsing)
- ...

Confidential ©2026 Mistral AI. All rights reserved.

Practical considerations

Difficult in practice:

- How to pretrain? Need input & output pair
- Input & output pair is not necessarily aligned (log parsing)
- ...

BUT:

- Promising approach to combine modalities
- Enables early stopping / termination
- Many industry task need “realtime” systems
- Pretrained LLMs are adaptable

Confidential ©2026 Mistral AI. All rights reserved.

Streaming Input for current LLMs still?

Confidential ©2026 Mistral AI. All rights reserved.

Streaming Input for current LLMs still?

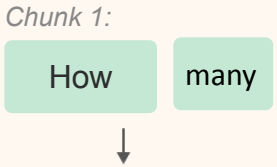
Yes, Faster prefill!

Long input has to be chunked anyways, let's start early!

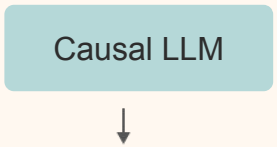
Confidential ©2026 Mistral AI. All rights reserved.

Streaming Input for faster prefilling

Input stream:



Model:



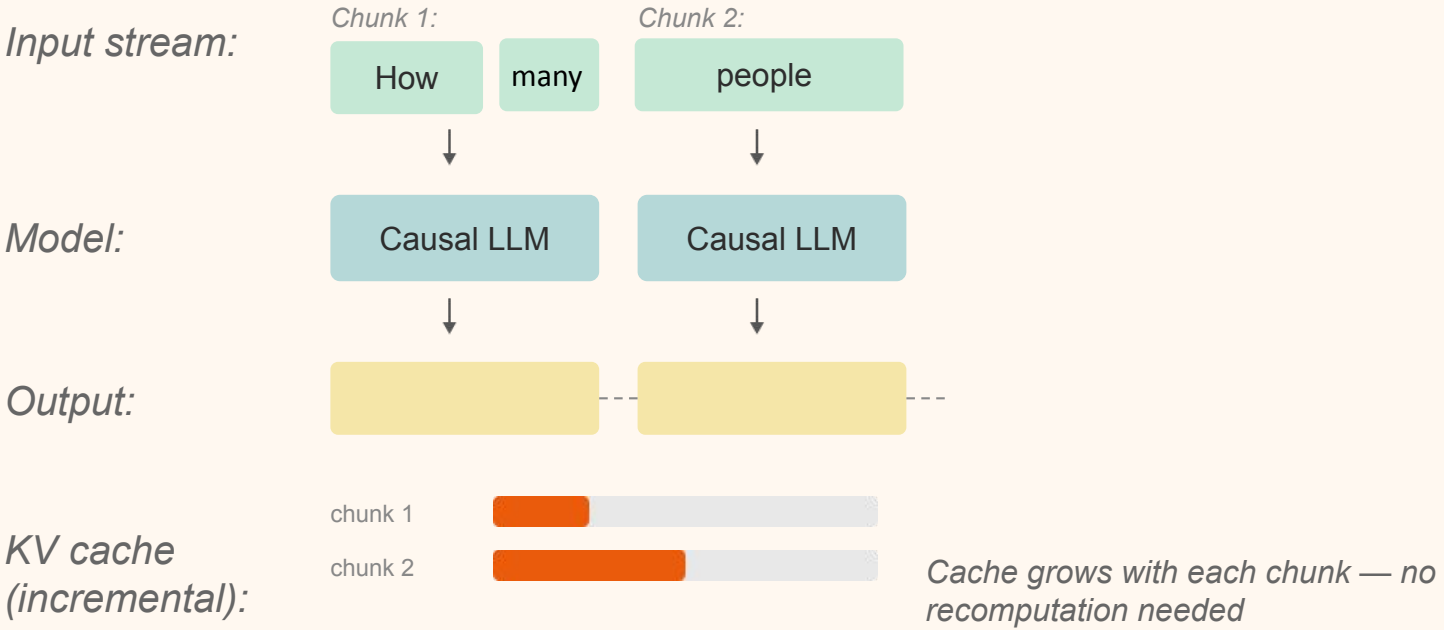
Output:

KV cache (incremental):

Cache grows with each chunk — no recomputation needed

Confidential ©2026 Mistral AI. All rights reserved.

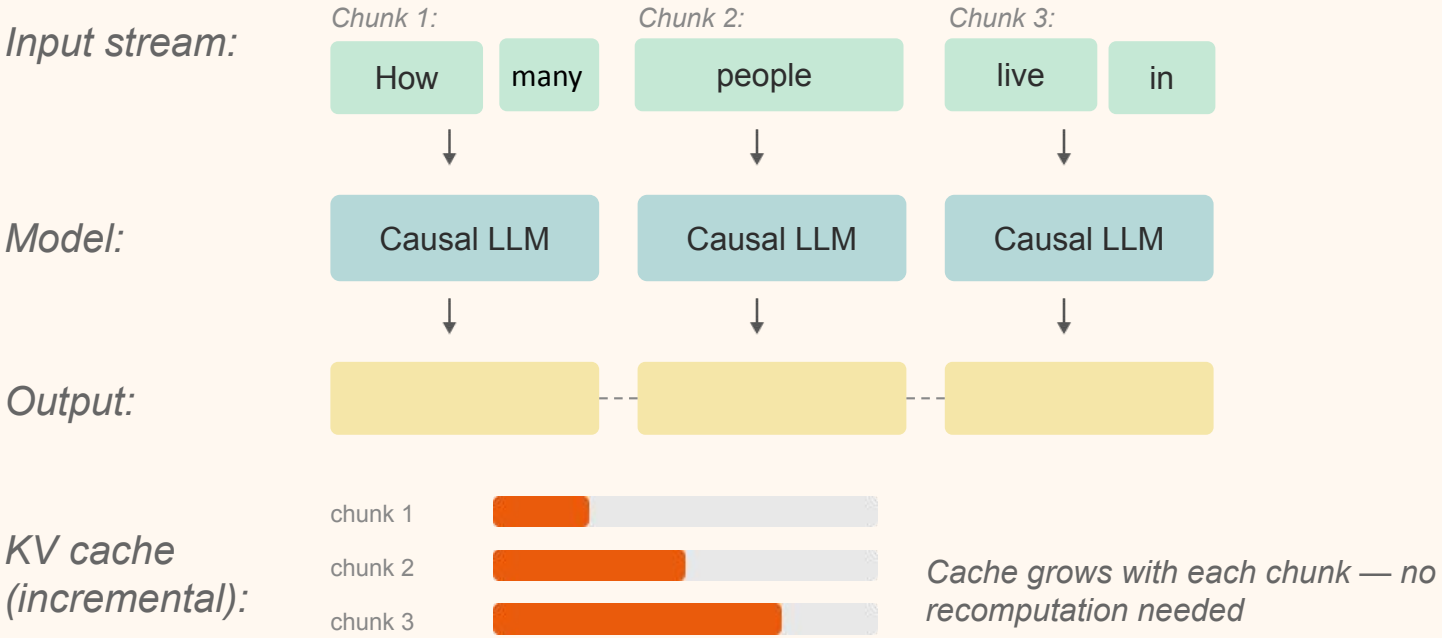
Streaming Input for faster prefilling



Cache grows with each chunk — no recomputation needed

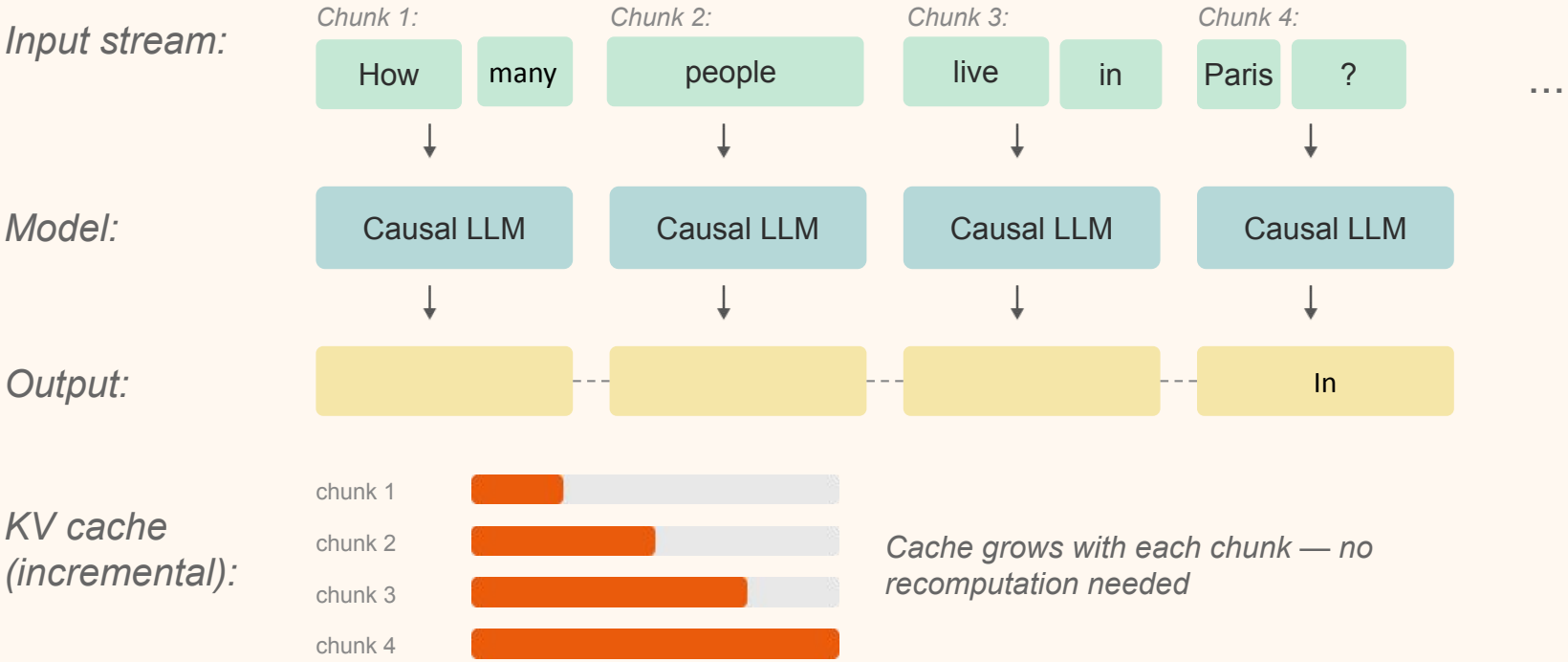
Confidential ©2026 Mistral AI. All rights reserved.

Streaming Input for faster prefilling



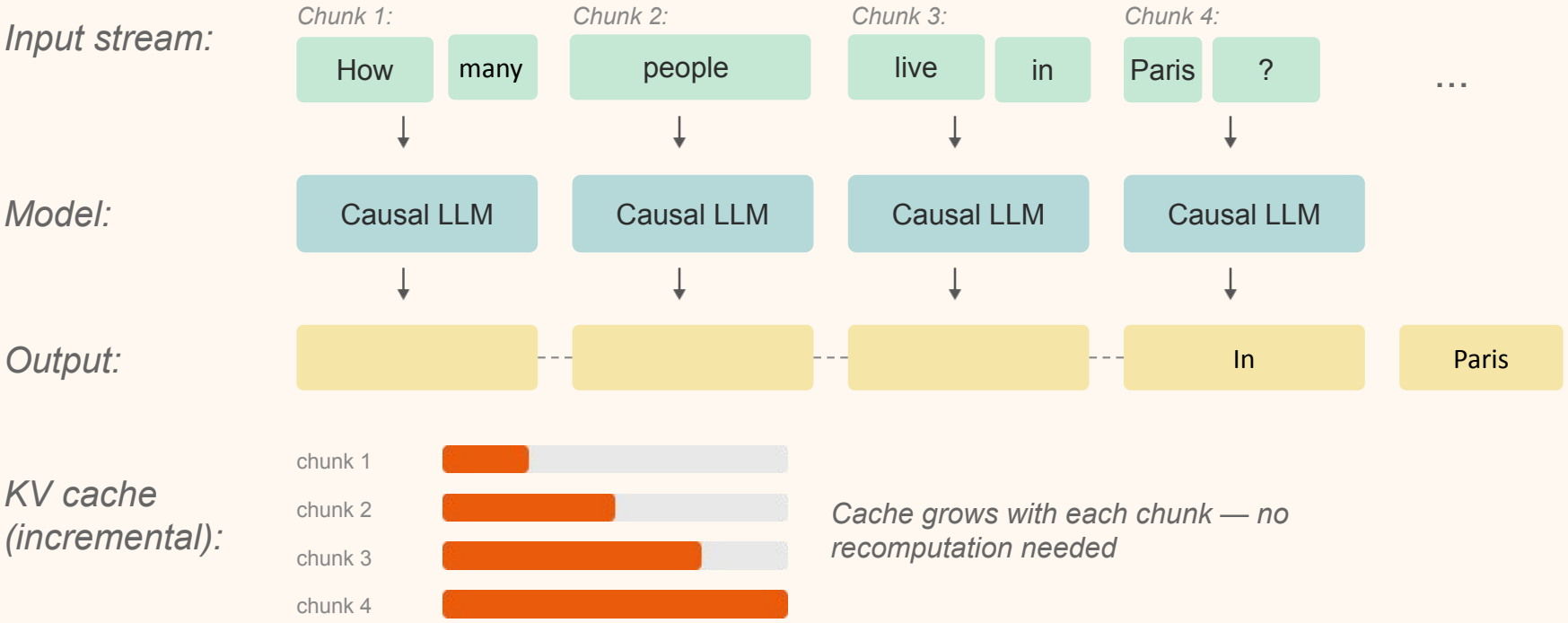
Confidential ©2026 Mistral AI. All rights reserved.

Streaming Input for faster prefilling



Confidential ©2026 Mistral AI. All rights reserved.

Streaming Input for faster prefilling



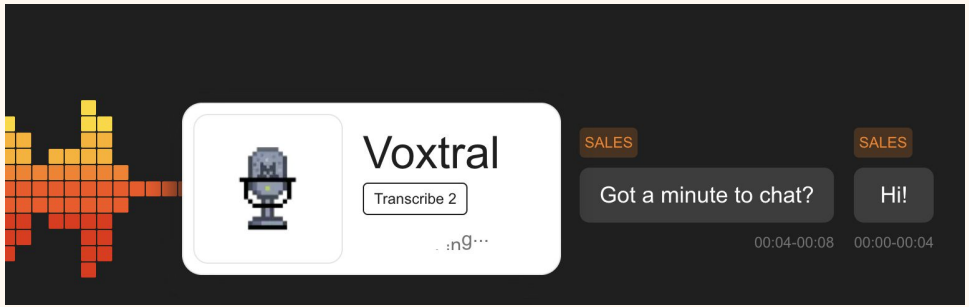
Confidential ©2026 Mistral AI. All rights reserved.

Support for input streaming in vLLM

- Blog: https://vllm-project-github-a2gx9w6hp-simon-mos-projects.vercel.app/2026/01/31/streaming-realtime.html
- Docs: https://docs.vllm.ai/en/latest/serving/openai_compatible_server/?h=realtime#realtime-api
- Voxtral: https://mistral.ai/news/voxtral-transcribe-2

```python
from dataclasses import dataclass
from vllm.inputs import PromptType
from vllm.sampling_params import SamplingParams

@dataclass
class StreamingInput:
    prompt: PromptType
    sampling_params: SamplingParams | None = None
```



Confidential ©2026 Mistral AI. All rights reserved.