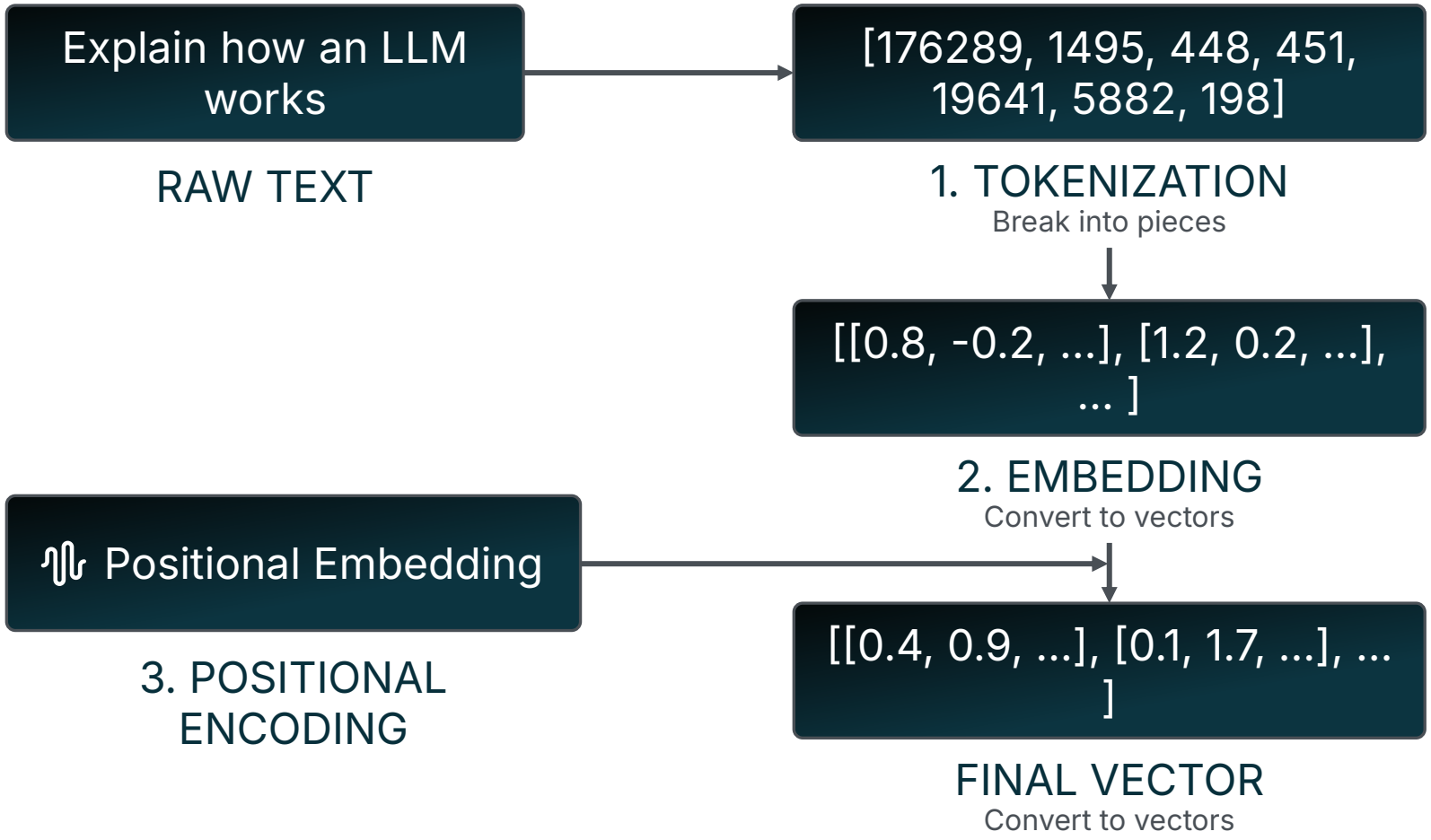
bitbasti

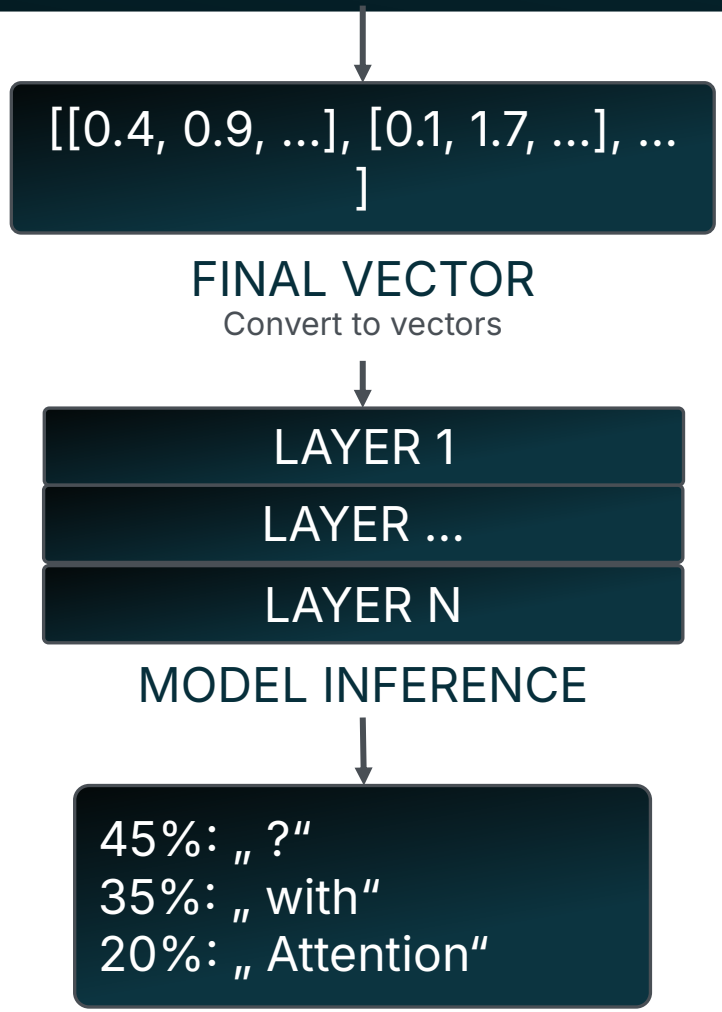
From Llama to Llasa to Chatterbox

How to Make LLMs Talk

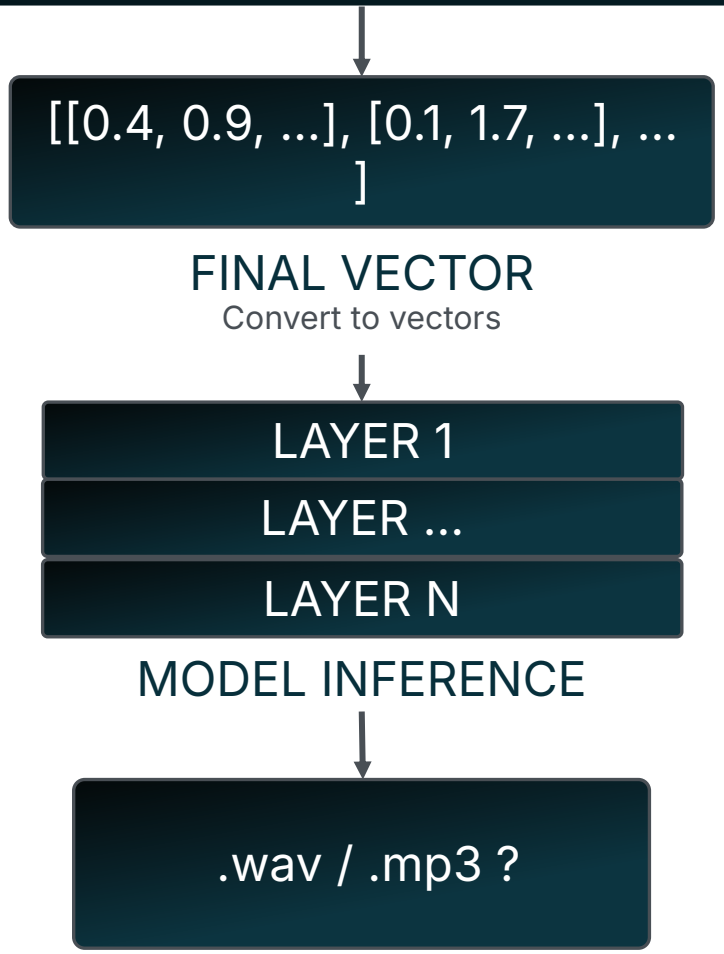
How does an LLM work?



How does an LLM work?



How to generate Audio?



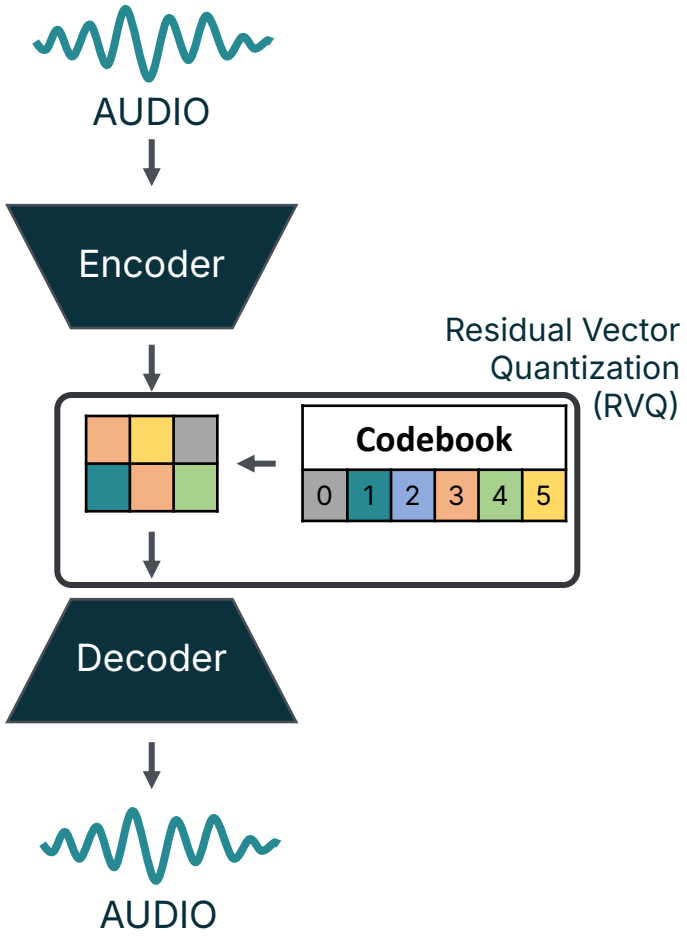
The Problem: Raw Audio is Too Big

- A 10-second WAV file = 441 000 raw samples.
- As a token sequence, this is too long for an LLM to handle.
- Result: Extremely slow performance and very short audio generation limits.

The Solution: Neural Encoder/Decoder

- The Neural Encoder acts as a smart "audio compressor.“
- It converts the long raw audio stream into a short, efficient sequence of tokens.
- The LLM processes these tokens, not the raw audio.
- A Neural Decoder then reconstructs the final audio from the LLM's token output.

How to generate Audio?



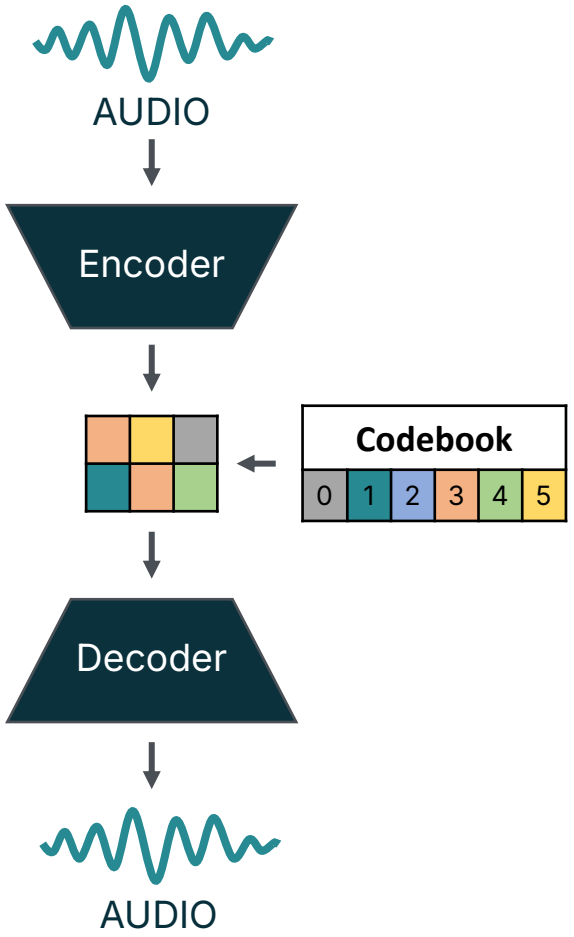
The Solution: Neural Encoder/Decoder

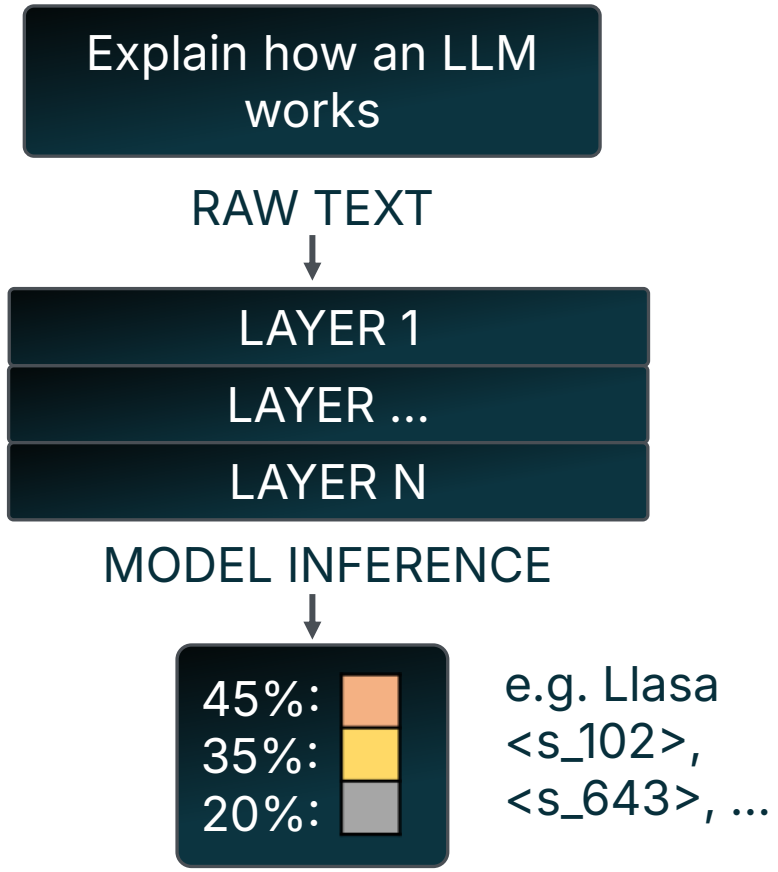
- The Neural Encoder acts as a smart "audio compressor.“
- It converts the long raw audio stream into a short, efficient sequence of tokens.
- The LLM processes these tokens, not the raw audio.
- A Neural Decoder then reconstructs the final audio from the LLM's token output.

Examples:

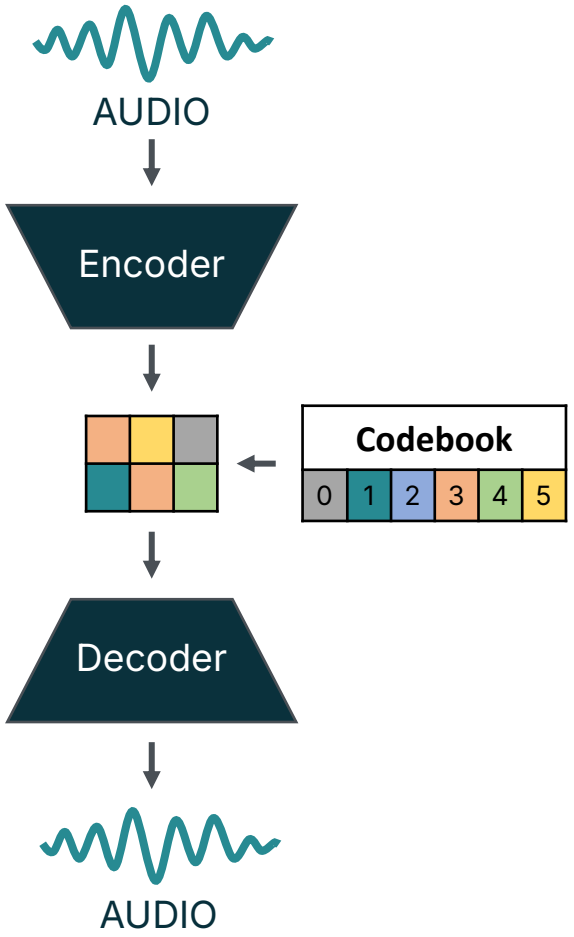
- SNAC → Orpheus
- XCodec-2 → Llasa
- S3 (Cosyvoice) → Chatterbox

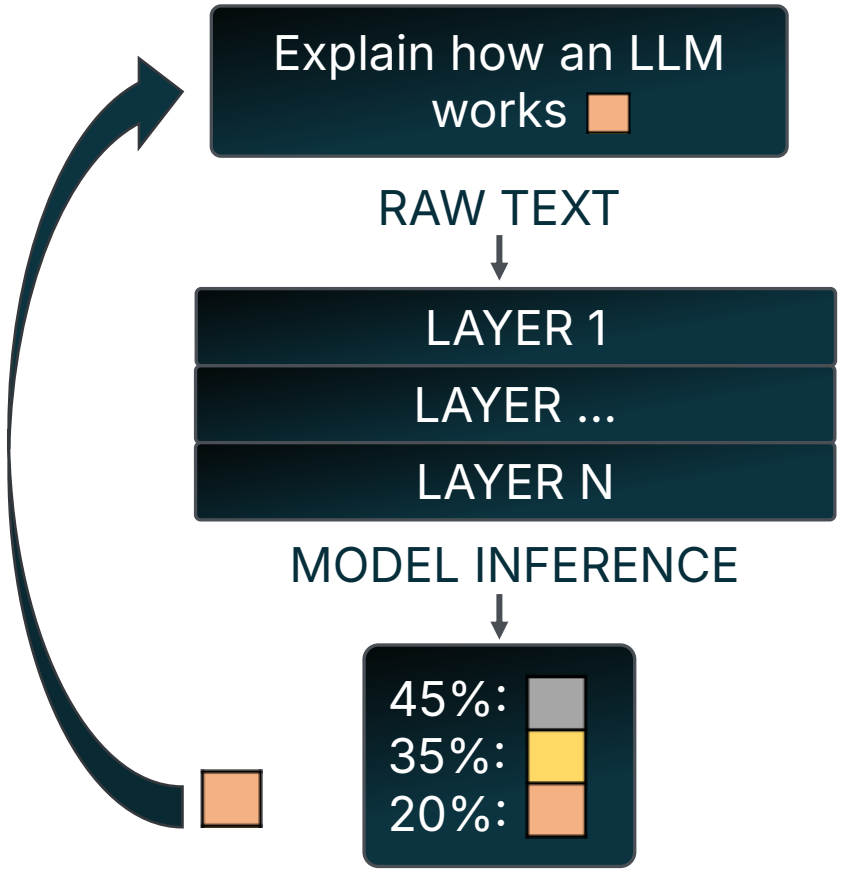
How to generate Audio?



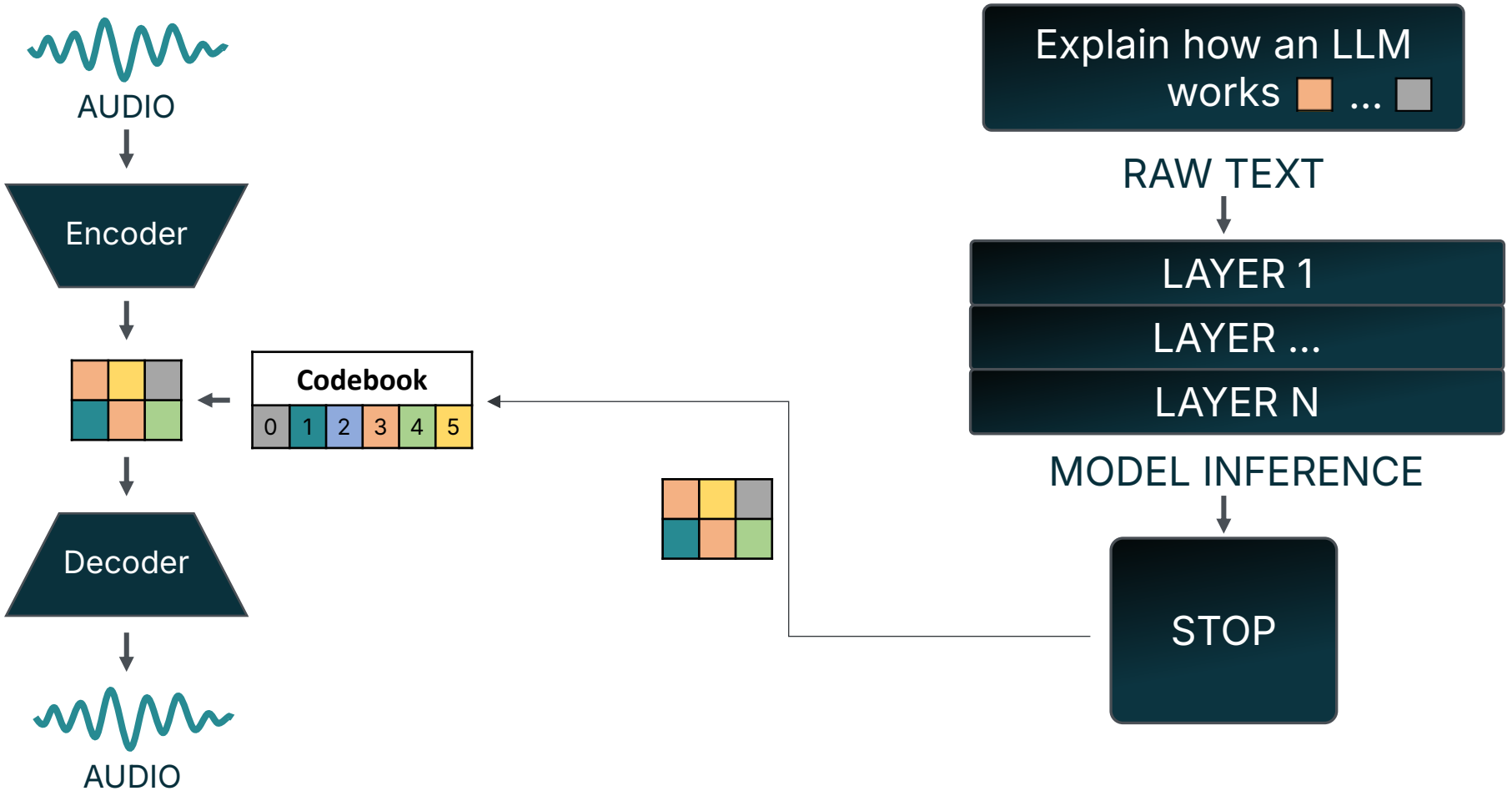


How to generate Audio?





How to generate Audio?



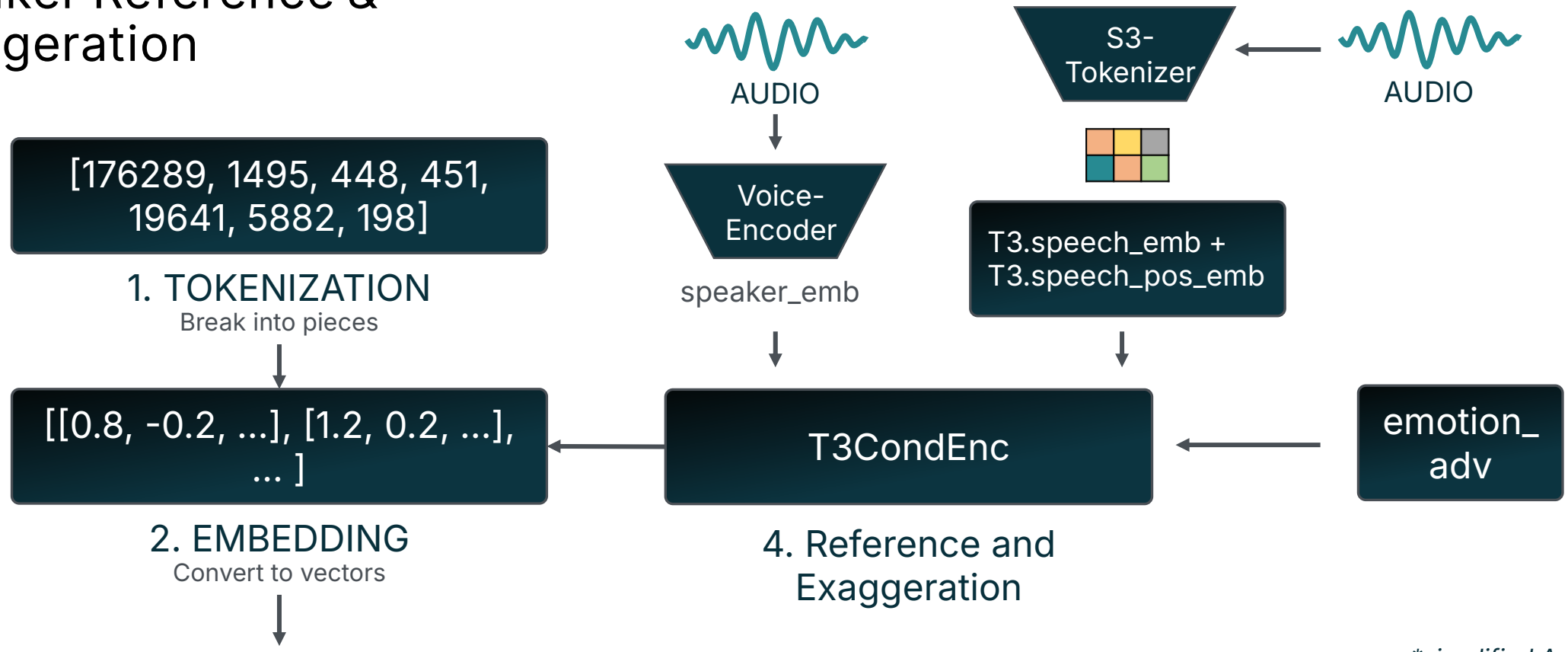
Why use a LLM?

+ Optimized inference libraries → vLLM, SGLang
+ Optimized training libraries → Liger Kernel, Axolotl, ...
+ Push towards multimodality in LLMs

- „Slower“ Inference → autoregressive generation
- Currently mostly more parameters used
- Neural Encoder/Decoder necessary
- Orpheus, Llasa and partially Chatterbox suffer from repeating token

What is special about Chatterbox?

Speaker Reference & Exaggeration



**simplified Architecture*

bitbasti

Find me at

Try it out

https://huggingface.co/SebastianBodza/Kartoffelbox-v0.1
https://huggingface.co/spaces/SebastianBodza/Kartoffelbox