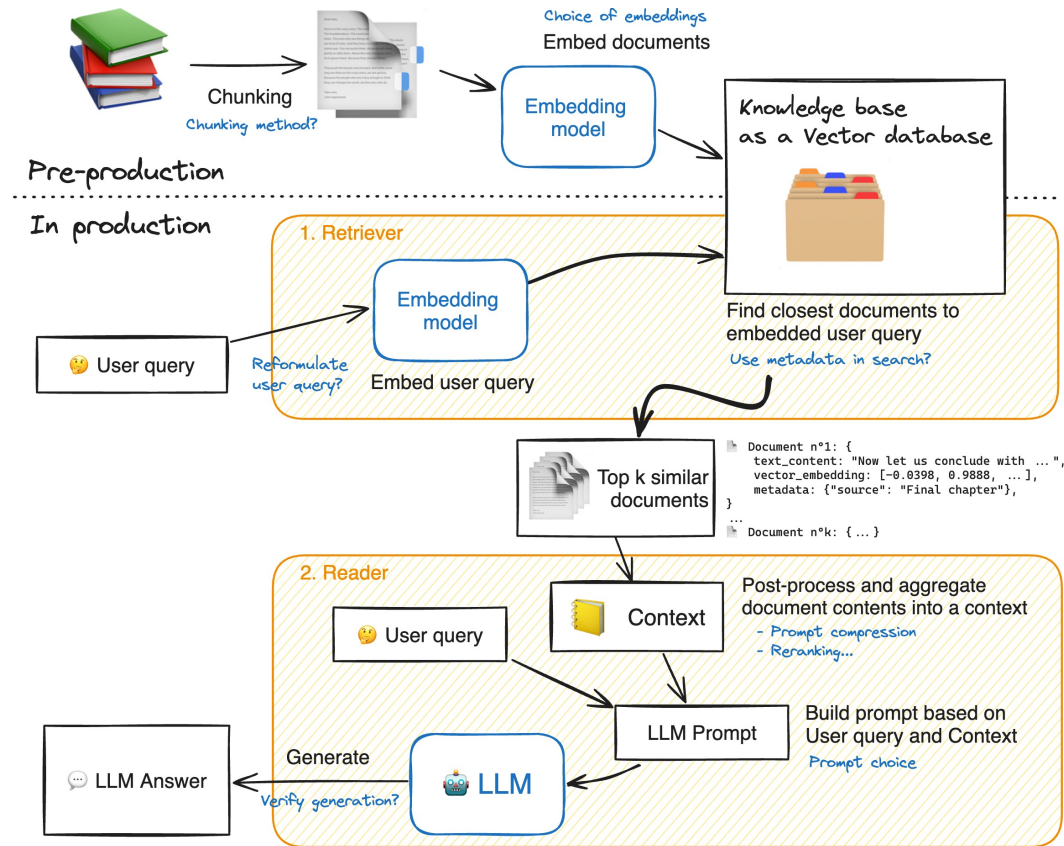


Offline Voice Based Retrieval Augmented Generation

- Retrieval Augmented Generation (RAG)
- Offline Voice Based RAG Architecture
 - Wake word Classifier
 - Automatic Speech Recognition
 - Vector database
 - Large Language Model
 - Text to Speech

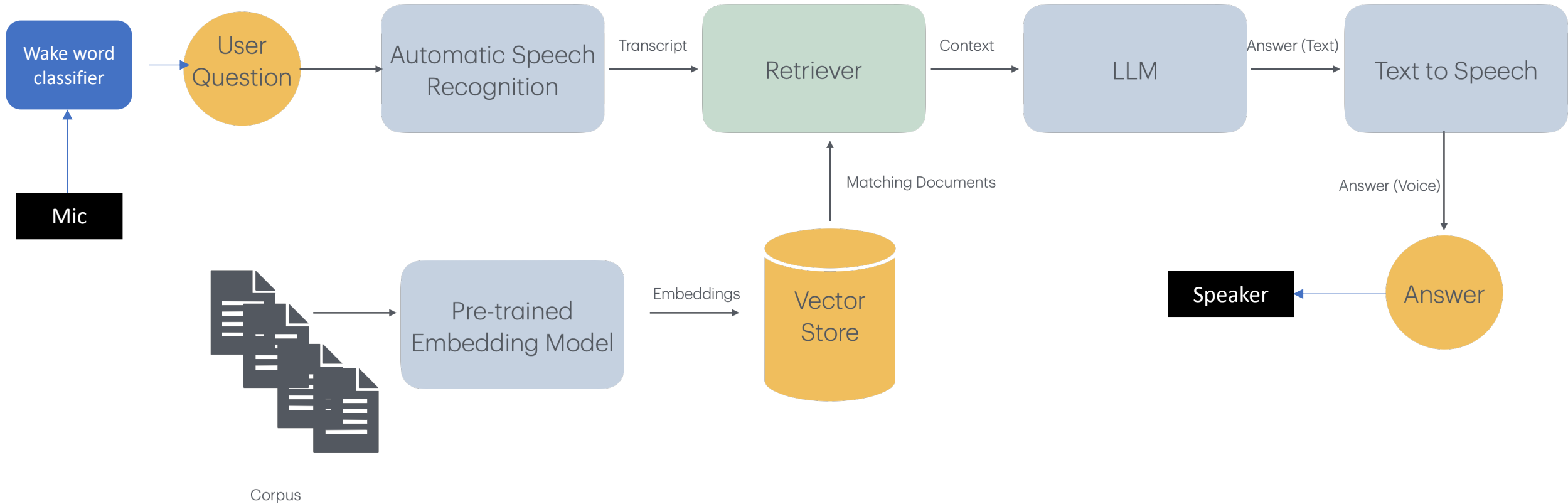
Retrieval Augmented Generation

- The concept of RAG was originally proposed as a way to provide additional context information to Large Language Models to generate more specific and accurate responses.



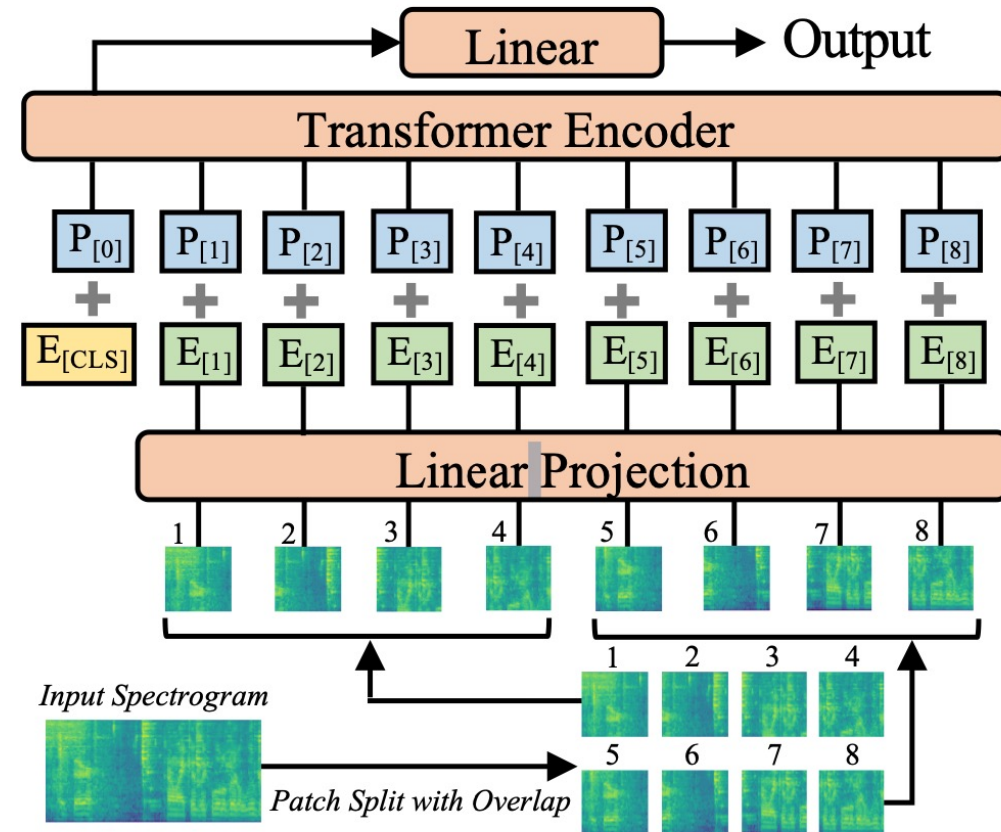
Source: [2]

Offline Voice Based RAG



Audio Classification

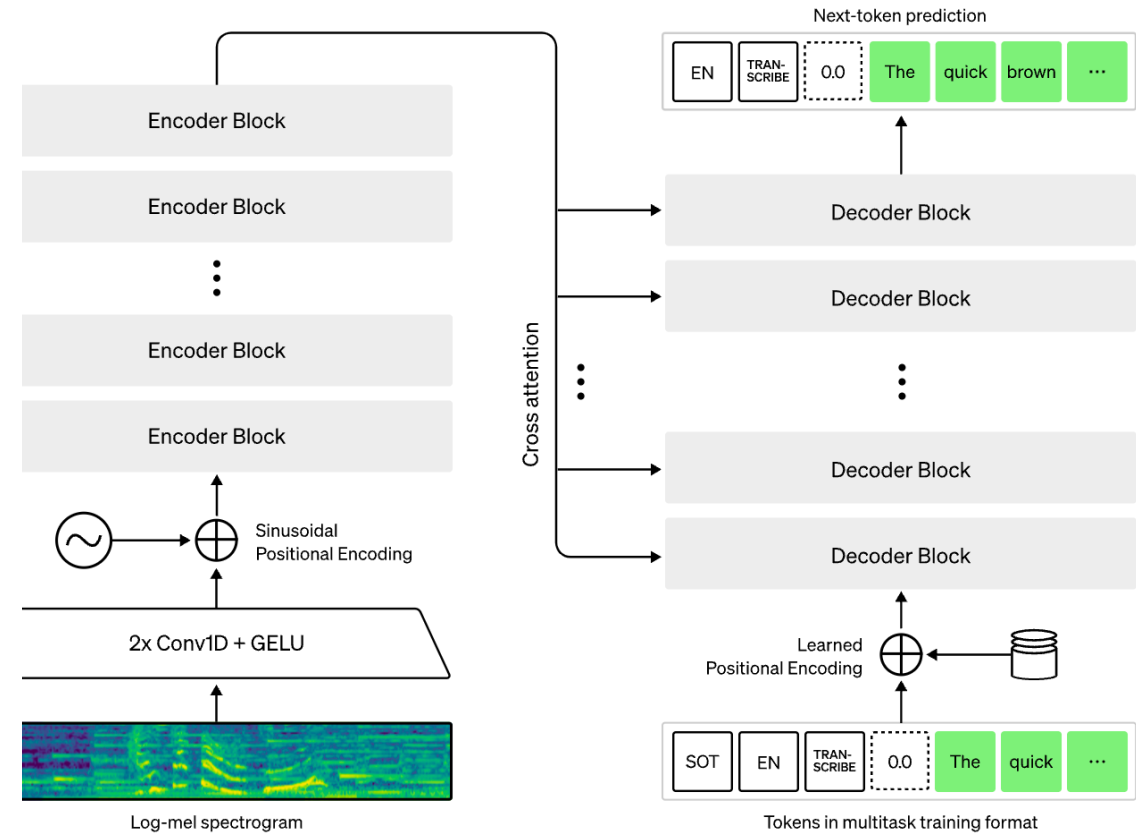
- Audio classification converts audio into a class label
- Audio Spectrogram Transformer (AST) model finetuned on speech commands dataset was used



Source: [3]

Automatic Speech Recognition (ASR)

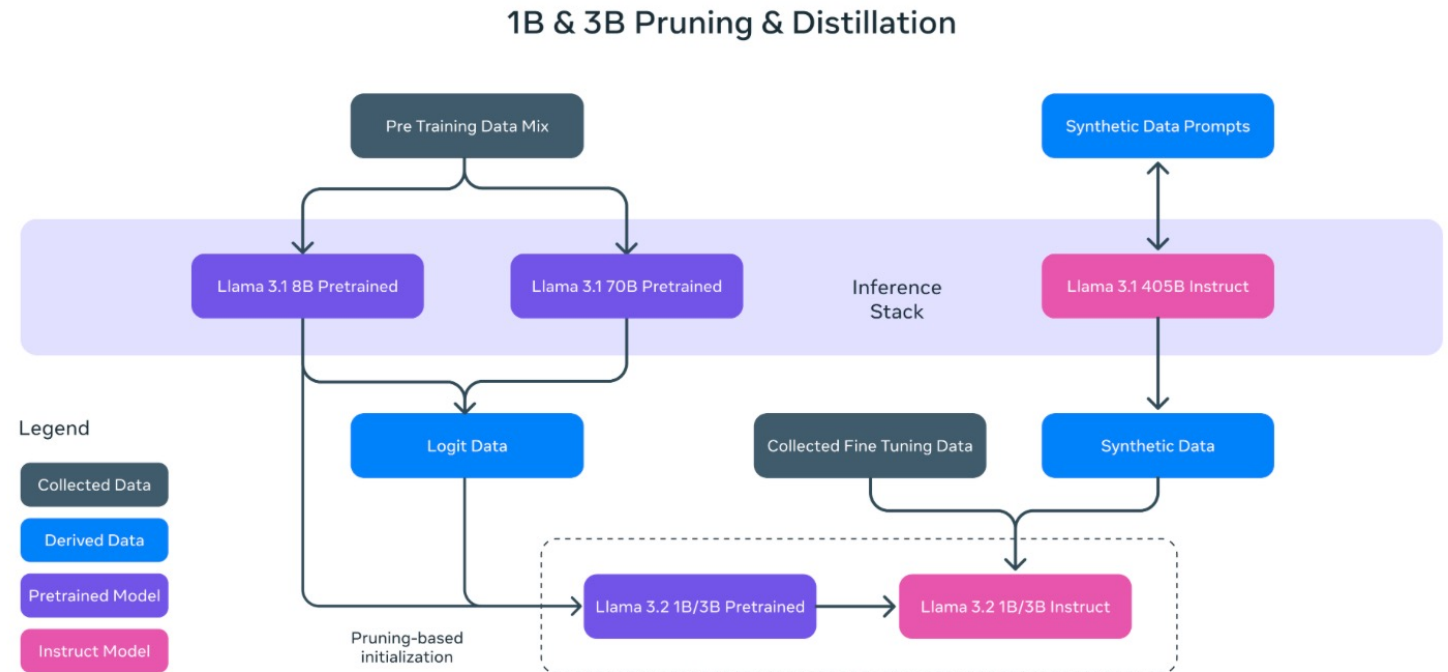
- ASR models convert speech into text
- Whisper model was used in this project



Source: [4]

Vector Database & Large Language Model

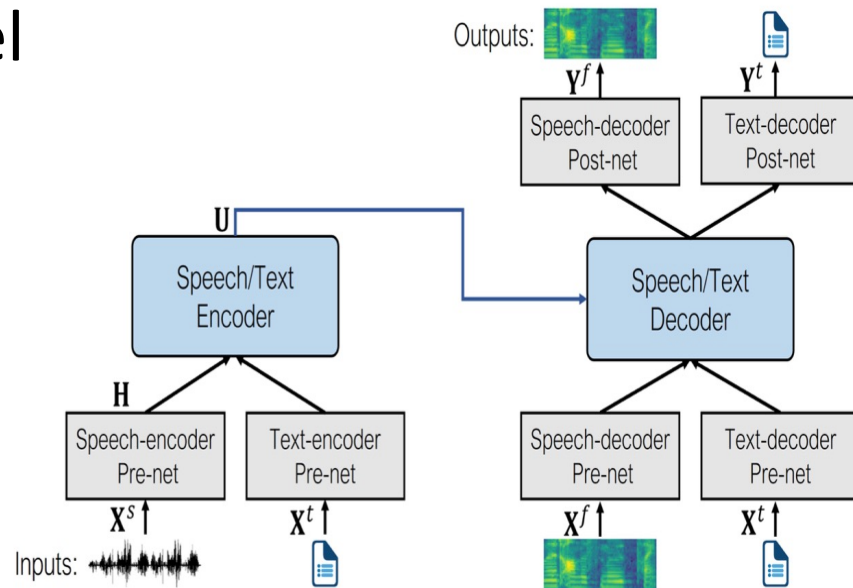
- chromadb is used as the vector database
- Llama 3.2 1B Instruct and 3B Instruct models by Meta are used



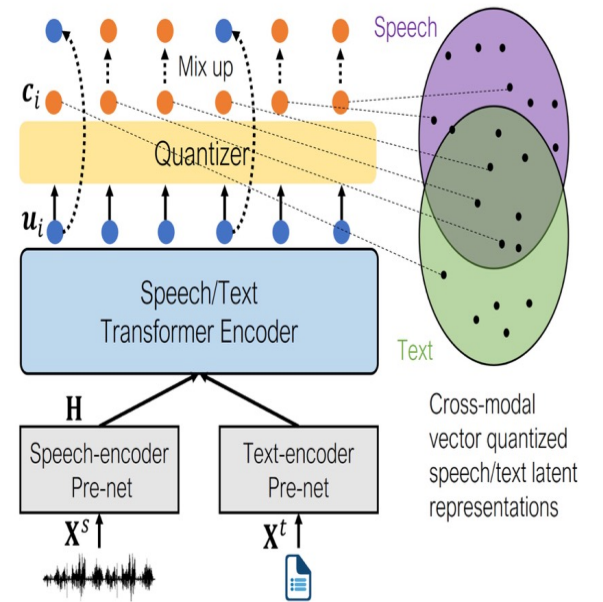
Source: [5]

Text to Speech

- SpeechT5 by Microsoft is used as text to speech model



(a) The model architecture of SpeechT5



(b) The joint pre-training approach

Source: [6]

Demonstration User Interface

- Loading Llama model
- Llama loaded
- Loading ASR model
- ASR model is loaded
- Loading TTS model
- TTS model is loaded
- Loading wake word model
- Wake word detection model loaded

 RUNNING... Stop Deploy

Voice RAG



What is meant by transformers?



Transformers are non-recurrent networks based on multi-head attention, a kind of self-attention. They are a type of neural network architecture that allows for parallel processing of input vectors, making them particularly well-suited for tasks such as language modeling and machine translation.



What is self-attention?



Self-attention is a mechanism in the transformer that weighs and combines the representations from appropriate other tokens in the context from the previous layer to build the representation for tokens in the current layer.



References

- [1] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Apr. 12, 2021, *arXiv*: arXiv:2005.11401. Accessed: Feb. 19, 2024. [Online]. Available: <http://arxiv.org/abs/2005.11401>
- [2] ‘Advanced RAG on Hugging Face documentation using LangChain - Hugging Face Open-Source AI Cookbook’. Available: https://huggingface.co/learn/cookbook/en/advanced_rag. [Accessed: Dec. 03, 2024]
- [3] Y. Gong, Y.-A. Chung, and J. Glass, ‘AST: Audio Spectrogram Transformer’. *arXiv*, Jul. 08, 2021. doi: 10.48550/arXiv.2104.01778. Available: <http://arxiv.org/abs/2104.01778>. [Accessed: Dec. 03, 2024]
- [4] Open AI., “Introducing Whisper”. Sep. 21, 2023, OpenAI Website Accessed Sep. 28. 2024.
[Online]. Available <https://openai.com/index/whisper/>
- [5] “Llama 3.2: Revolutionizing edge AI and vision with open, customizable models,” Meta, Llama 3.2: Revolutionizing edge AI and vision with open, customizable models (accessed Dec. 3, 2024).
- [6] J. Ao et al., ‘SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing’. *arXiv*, May 24, 2022. doi: 10.48550/arXiv.2110.07205. Available: <http://arxiv.org/abs/2110.07205>. [Accessed: Dec. 03, 2024]
- [7] D. Jurafsky and J. H. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition with Language Models*, 3. ed. Online manuscript released August 20, 2024. <https://web.stanford.edu/~jurafsky/slp3>.
- [8] ‘Hugging Face – The AI community building the future.’, Nov. 29, 2024. Available: <https://huggingface.co/>. [Accessed: Dec. 03, 2024]