Advanced NLP/ASR

What does GPT in ChatGPT stand for?

TIL-AI
TODAY I LEARNED-AI

GPT

Generative

- Can develop coherent and contextually relevant text based on a given prompt

Pre-Trained

- Model has been trained on and learnt from a large amount of data during a pre-training phase

Transformer

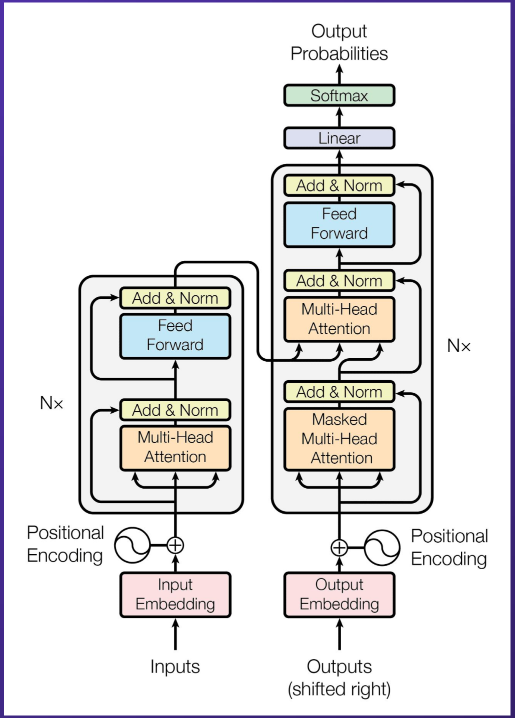
- Based on the transformer architecture

Introduction to Transformers

TIL-AI
TODAY I LEARNED-AI

Introduction to Transformer

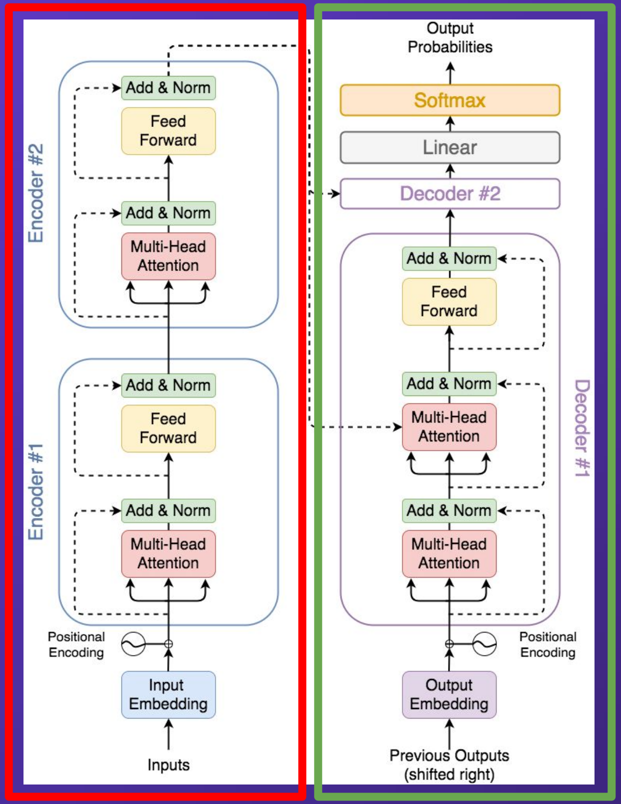
- Introduced in the 2017 paper "Attention is All You Need"
- Key innovation: the self-attention mechanism
- Self-attention allows the model to weigh the importance of different words in a sentence simultaneously
- No need for sequential processing, enabling massive parallelization



TIL-AI
TODAY I LEARNED-AI

Encoder-Decoder Transformer

Encoder

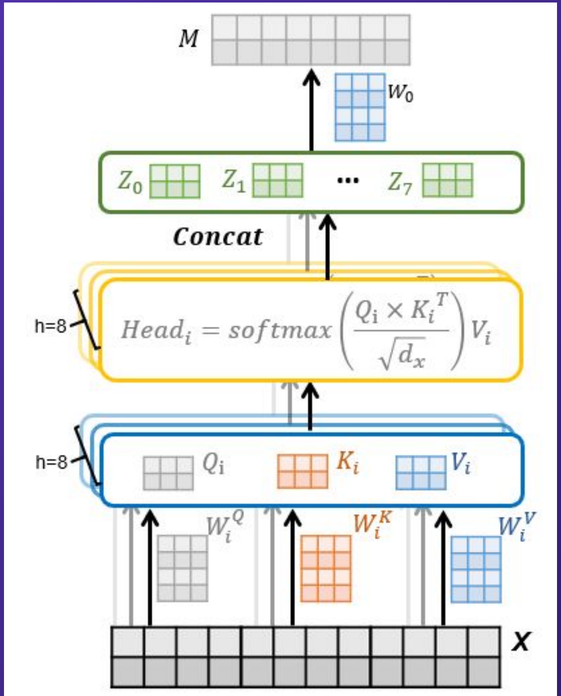


Decoder

TIL-AI
TODAY I LEARNED-AI

Multi-Head Self-Attention

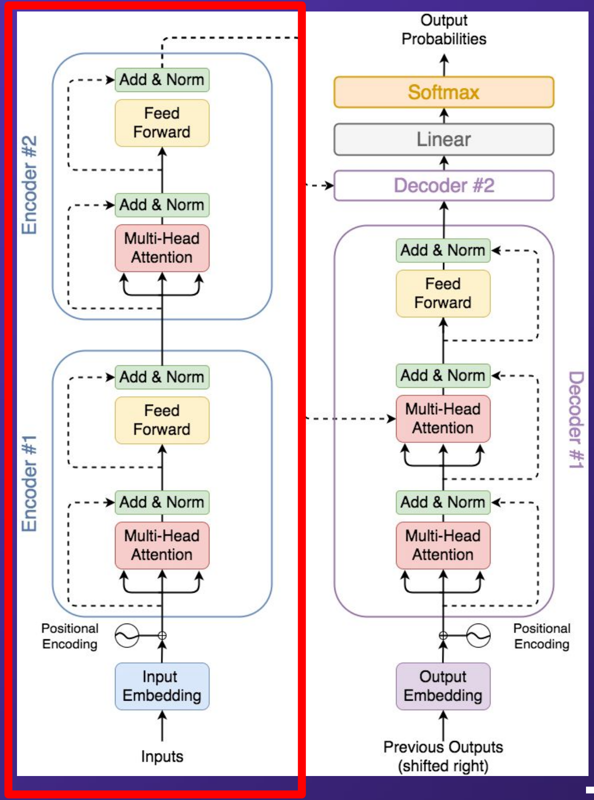
- The heart of the transformer
- Each word in the sentence attends to all other words, determining relationships and importance
- Multi-head means multiple sets of attention calculations run in parallel, each focusing on different aspects of the input



TIL-AI
TODAY I LEARNED-AI

Encoder Blocks

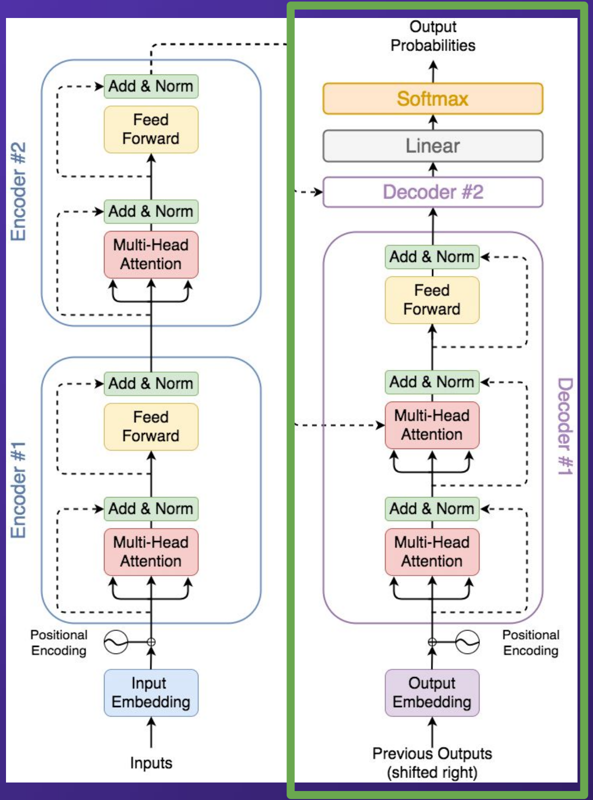
- The encoder consists of:
  - Input Embeddings
  - Positional Encoding
  - Attention blocks that include:
    - Multi-head self-attention layer
    - Feed-forward network for further processing
    - Add & normalize layers for stabilizing training



Decoder Blocks

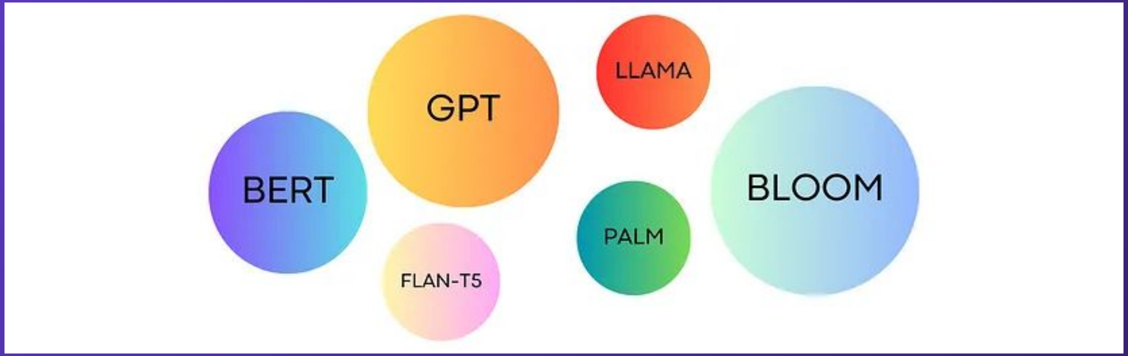
- The decoder has a similar structure to the encoder but with an extra element

- Masked multi-head attention: Ensures the model only looks at past words when generating text



Large Language Models (LLMs)

- Large Language Models (LLMs) like GPT-3 and T5 form the backbone of zero/few-shot learning
- Pre-trained on massive text corpora, LLMs acquire rich linguistic representations and implicit knowledge



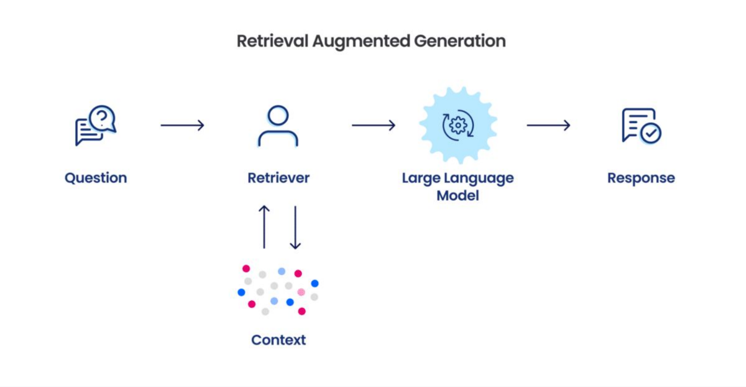
Few-Shot & Zero-Shot

- Few-Shot Learning:
  - Aims to learn from a very small number of labeled examples
  - Leverages prior knowledge from pre-training on broader tasks
  - Better at adapting to new tasks with minimal data

- Zero-Shot Learning:
  - Adapts to a new task without any explicit training examples
  - Relies heavily on pre-trained models' (ex: LLMs) ability to understand language and extract information from prompts and instructions

Zero/Few-Shot QA Techniques

- Prompt Engineering: Carefully designing prompts that best leverage the LLM's capabilities
- Answer Verification: Adding mechanisms to assess the quality and reliability of generated answers
- Retrieval Augmented Generation: Incorporating external knowledge bases or structured data to enhance LLM reasoning



Hands-on activity

TIL-AI
TODAY I LEARNED-AI

Advanced Audio Processing

TIL-AI
TODAY I LEARNED-AI

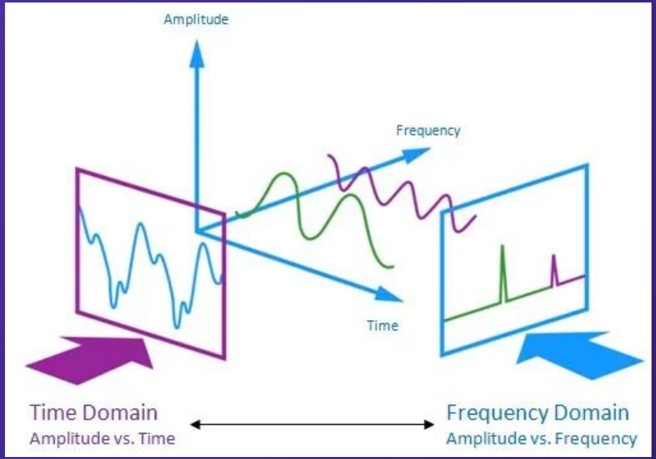
Importance of Feature Extraction

- Raw audio signals are complex and contain a mix of relevant and irrelevant information

- Feature extraction transforms raw audio into compact, meaningful representations that highlight speech patterns

- Effective features make it easier for ASR models to distinguish between different sounds and words

- The quality of feature extraction directly impacts the overall accuracy of an ASR system

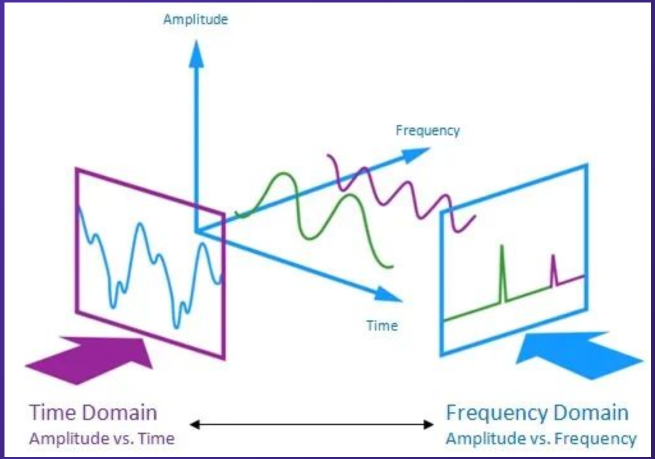


TIL-AI
TODAY I LEARNED-AI

From Time to Frequency Domain

- ASR systems rely heavily on transforming speech signals from time domain to frequency domain to reveal the underlying frequencies that make up the speech

- Time domain: Speech is represented as a waveform where the amplitude (intensity) of the signal is plotted over time

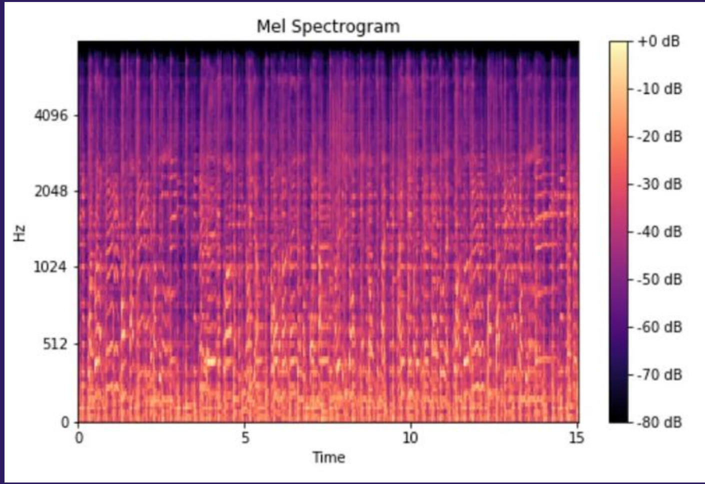


- Discrete Fourier Transform (DFT) is used to decompose a time-domain signal into individual frequencies and amplitudes

- The output of the DFT tells us how much energy is present at each frequency in the original speech signal

Mel-Spectrograms

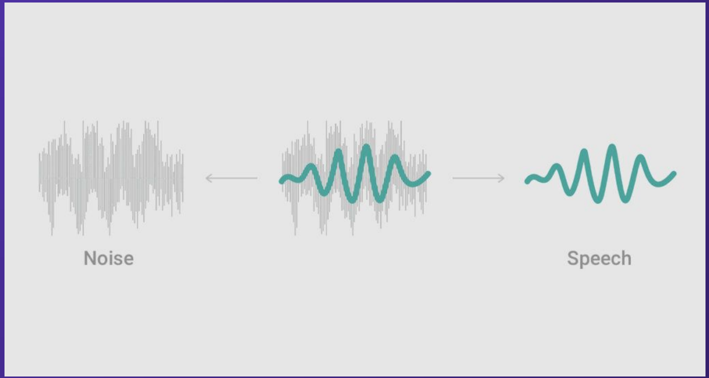
- Type of spectrogram where the frequency scale is converted to the Mel scale
- Mel scale more closely approximates human auditory system's response than the linear frequency scale; making it more effective for audio-related tasks in human speech and music



TIL-AI
TODAY I LEARNED-AI

Noise Reduction

- Noise can come from various sources: background conversations, traffic, machinery, wind, etc
- Noise masks important speech features and disrupts the acoustic signal
- This noise makes it harder for ASR systems to identify words and phrases correctly
- Noise reduction helps minimize background noise without distorting speech



Noise Reduction Techniques

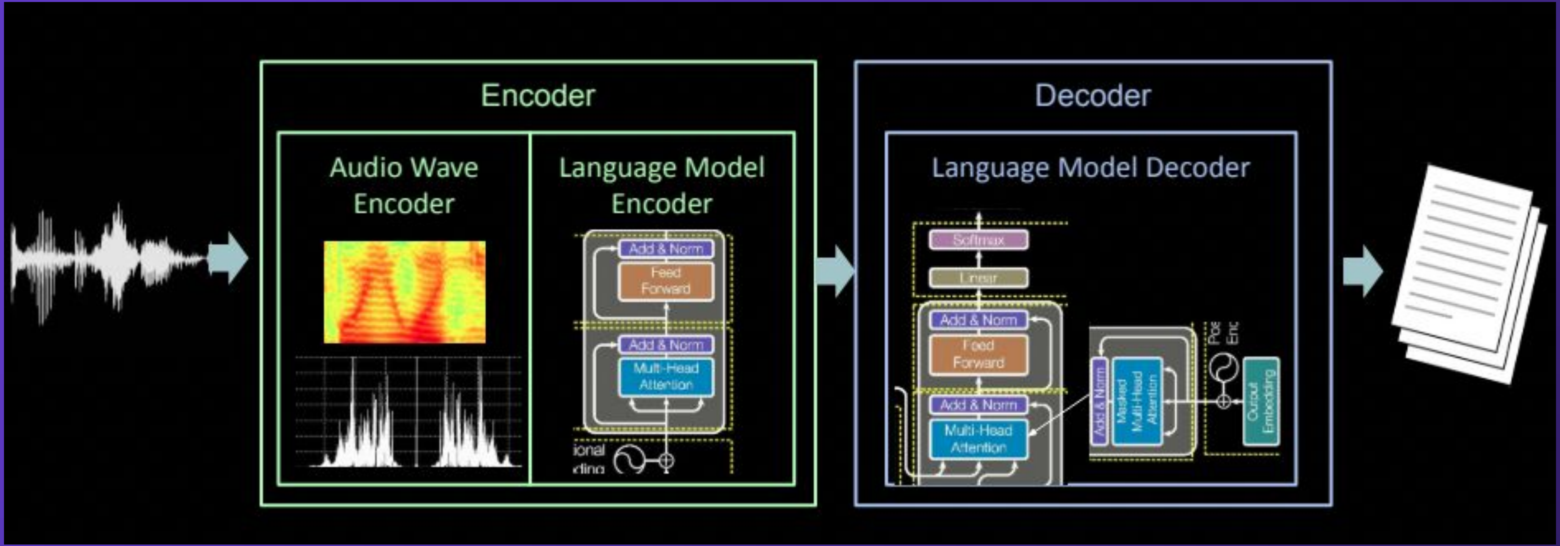
- Spectral Subtraction
  - Estimates the noise spectrum during non-speech segments (silent periods)
  - Subtracts the estimated noise spectrum from the speech spectrum
  - Effective for broadband noise, but it can introduce artifacts if the noise is not stationary or if it overlaps with speech frequencies

Noise Reduction Techniques

- Deep Neural Networks-Based
    - Neural Networks can be trained to distinguish between speech and noise, enhancing speech recognition accuracy
    - Can learn complex relationships between speech and noise, allowing for highly adaptive noise reduction in various scenarios
    - Example: Deep Denoising Convolutional Neural Network (DnCNN)

Transformers in ASR

How Transformers Process Speech



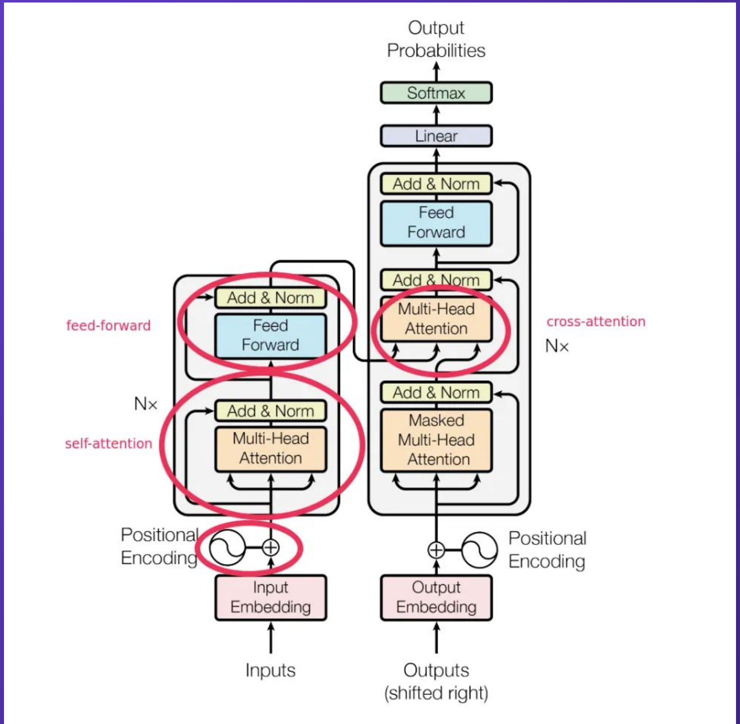
How Transformers Process Speech

- Speech as input is first converted into a spectrogram or a sequence of feature vectors, which represent the audio signal's power at various frequencies over time

- Encoder:
  - The encoder processes the speech input using multi-headed self-attention, allowing the model to focus on different parts of the speech input concurrently
  - This helps recognize patterns like phonemes, syllables, and words by analyzing features in parallel

- Decoder:
  - The decoder receives acoustic information from the encoder via cross-attention and combines it with a causal self-attention mechanism to predict the final transcript.

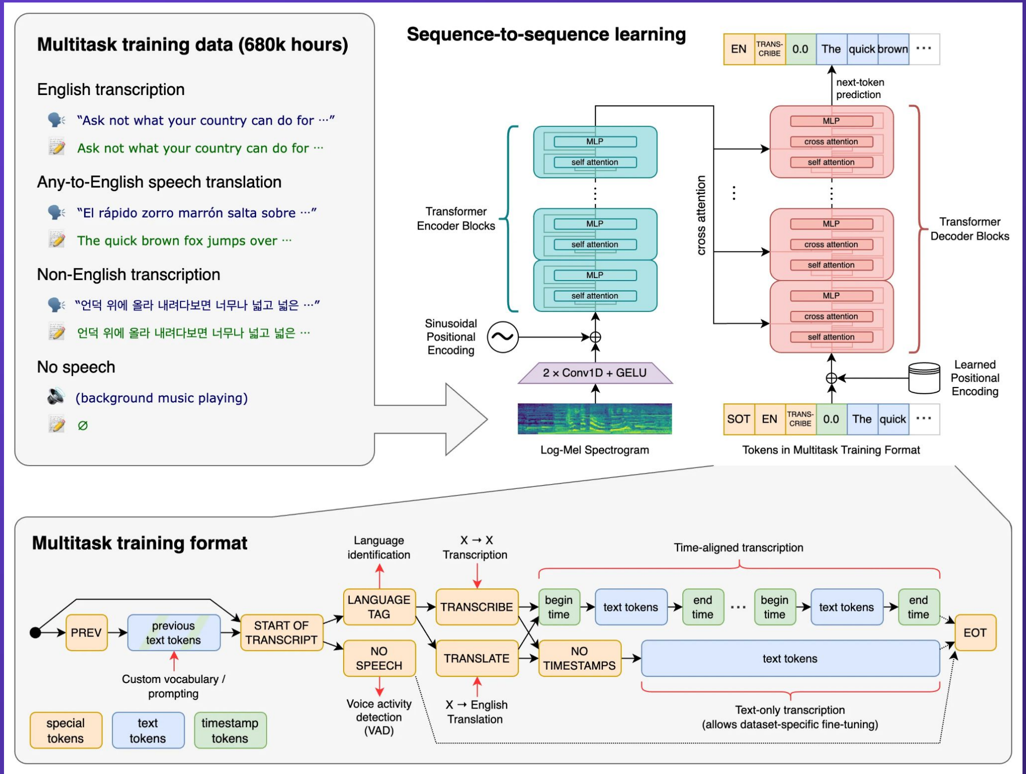
How Transformers Process Speech



Whisper Architecture

- Core Concept: Encoder-Decoder Transformer
- Key Components:
  - Audio Preprocessing
  - Encoder
  - Decoder
  - Special Tokens and Multitasking

Whisper Architecture



Audio Preprocessing

- Divides the raw input audio into 30-second chunks for efficient processing
- Each chunk is then converted into a log-Mel spectrogram
- Whisper can translate non-English audio directly into English text, having learned bilingual mappings during training on multilingual datasets

Hands-on activity