Exploring Electroacoustic Music Analysis with Multimodal Large Language Models

Dr. Egor Polyakov

Jahrestagung der Gesellschaft für Musikforschung
Kollaborationen ∞ Wider den Methodenzwang
11. - 14. September 2024

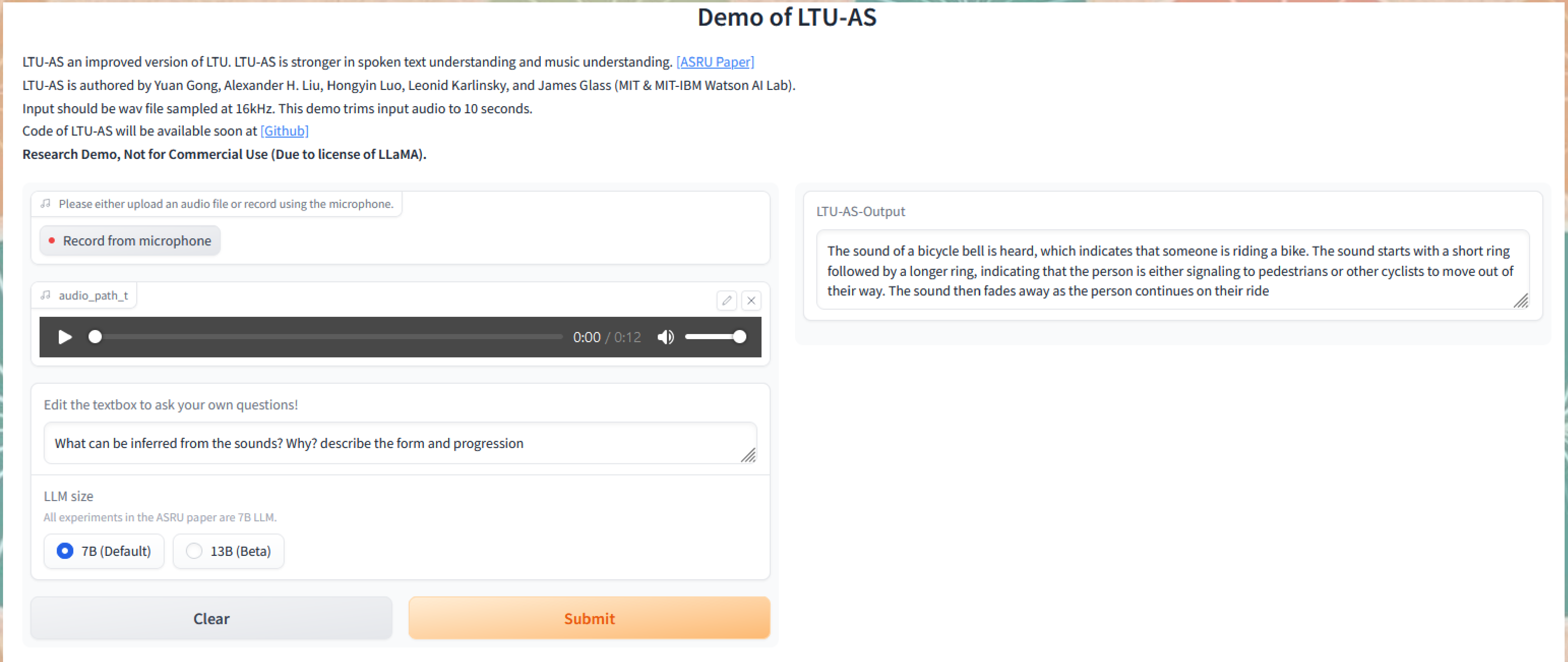
HOCHSCHULE
FÜR MUSIK UND THEATER
»FELIX MENDELSSOHN
BARTHOLDY«
LEIPZIG

Personal Background (in AI/ML)

- **Composer/Performer Background**
  - Strong foundation in algorithmic composition
  - Expertise in Max, Audiosculpt and OpenMusic
  - Focus on generative algorithms
- **CAMAT Project (2021-22)**
  - Python-based toolbox for symbolic music analysis
  - Conducted at HfM Weimar under the leadership of Martin Pfleiderer
  - Researchers: Egor Polyakov and Christon-Ragavan Nadar (Semantic Music Technologies Group at Fraunhofer-Institut für Digitale Medientechnologie)
  - AI was not a core component of the project, but interaction with Christon-Ragavan Nadar introduced me to machine learning topics, such as audio classification and dimensionality reduction
- **Shift in Research Focus towards machine learning applications in music**
  - ML-assisted audio synthesis
  - Audio and symbolic music generation

Analyzing Music with Large Language Models (LLMs)

- **Encounter with LLMs in Audio Analysis**
  - In early 2024, while preparing a talk on ML-based audio classification, I came across a series of papers by Gong et al.
  - Audio classification is a well-established task with a long research history and many standardized benchmarks
  - The papers by Gong et al. presented an intriguing approach that goes beyond traditional classification or feature extraction
- **Innovative Approach by Gong et al**.
  - Utilizes a Large Language Model (LLM) to describe audio content, rather than just classifying it
  - Enables reasoning and detailed questioning about what is happening within the analyzed audio file
  - Offers a new dimension to understanding and interpreting audio data
- **Implications for Music Analysis**
  - Let's explore the mechanics of how this approach works
  - Consider whether this new method could be beneficial for music analysis



https://huggingface.co/spaces/yuangongfdu/ltu-2

Examples

François Bayle - l'Aventure du cri (1969) 2:10 - 2:20

- The audio clip contains a series of short, sharp, and repetitive sounds that are likely produced by a mechanical device. The sounds have a consistent pattern and rhythm, suggesting they may be part of a larger system or process. They also appear to be relatively loud and intense, which could indicate that the machine is operating at full capacity or under heavy load.

François Bayle - l'Aventure du cri (1969) 0:22 - 0:32

- The audio clip contains a sound effect that is being created using a synthesizer. The sound starts with a low-pitched drone, followed by a series of rising and falling notes that create a melody. The sound gradually builds up in intensity before fading away.

Bernard Parmegiani - De Natura Sonorum - incidences / battements (1975) 0:00 – 0:10

- The sound of a bicycle bell is heard, which indicates that someone is riding a bike. The sound starts with a short ring followed by a longer ring, indicating that the person is either signaling to pedestrians or other cyclists to move out of their way. The sound then fades away as the person continues on their ride.

Spectromorphology?

- Introduced in 1997 by Dennis Smalley as an analytical theoretical framework for electroacoustic music
- Used to describe and interpret the ways in which sounds are shaped and interact within a musical context
- Particularly useful for analyzing complex, non-traditional musical forms, such as acousmatic and electroacoustic compositions, where traditional music theory may not fully apply
- Relies on a specialized **vocabulary to describe the characteristics** and behaviors of sound
- Focuses on the perceptual aspects of how sound evolves and interacts within a piece of music
- Key categories of descriptive terms in spectromorphology
    - Spectral Terms
    - Morphological Terms
    - Behavioral Terms
    - Source-Cause Terms

What are the capabilities of systems similar to LTU-AS in a spectromorphological context?

- **Spectral Terms**:
  - Brightness/Darkness: Referring to the high or low frequency content of a sound.
  - Density: The concentration of frequencies within the spectrum.
  - Texture: The layering or blending of different spectral elements.
- **Morphological Terms**:
  - Attack: The onset of a sound, which can be sharp, soft, gradual, etc.
  - Decay: The fading or termination of a sound.
  - Growth: The development or increase in a sound’s intensity or complexity over time.
  - Gesture: The dynamic movement of sound, often perceived as having a physical or directional component.
- **Behavioral Terms**:
  - Motion: The perceived movement of sound in space or its evolution over time.
  - Flow: How smoothly or abruptly a sound transitions or changes.
  - Grain: The perceived texture or granularity of a sound.
- **Source-Cause Terms**:
  - Mimesis: The imitation of real-world sounds.
  - Immateriality: Describing sounds that do not suggest a clear physical source.

What are multimodal LLMs?

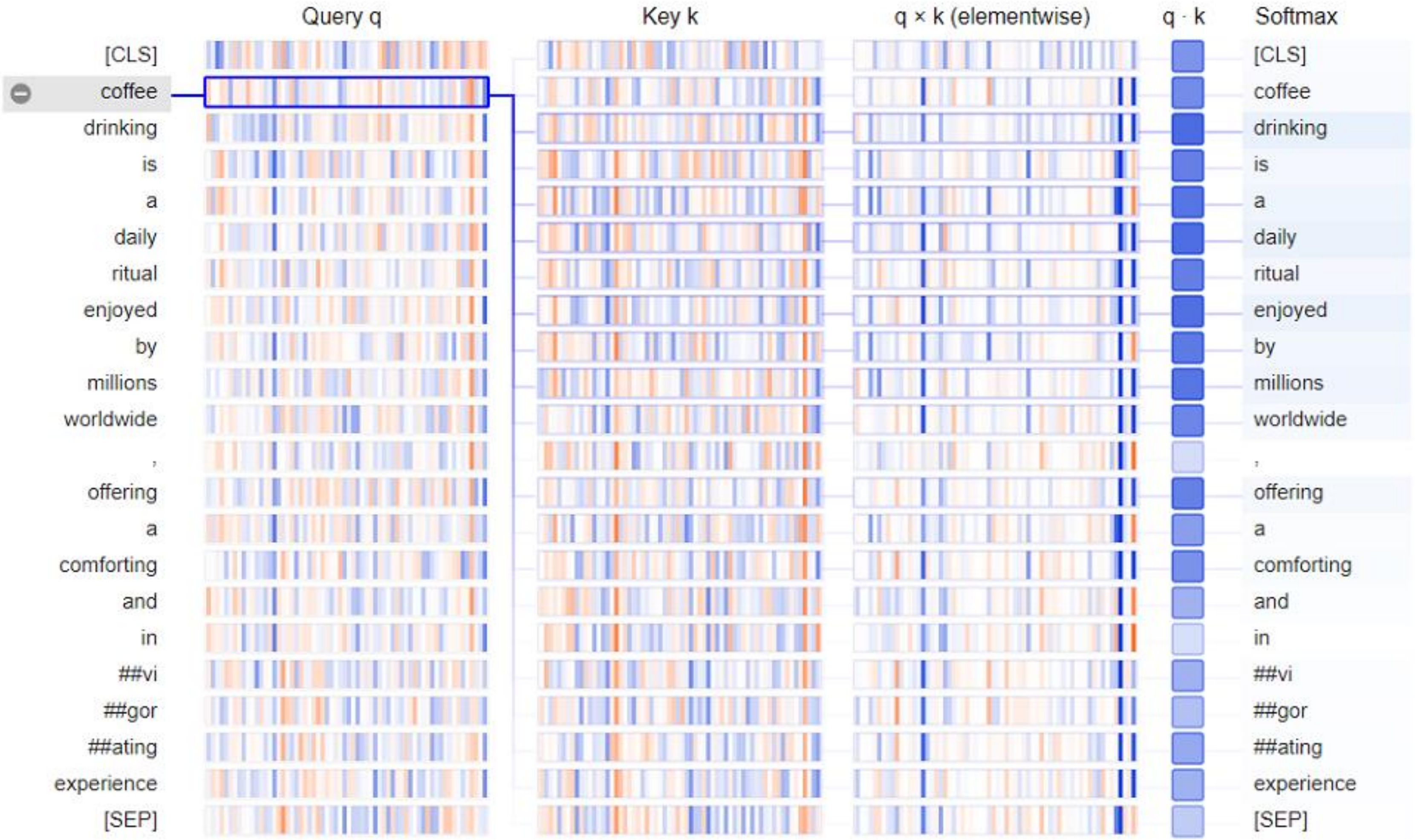
- „Classic“ LLMs are token predictors
- Token = flexible unit that can represent a subword, a word, or even a character
- Common words like "the" or "is" might be a single token, less common or compound words might be broken down into multiple tokens (e.g., "unhappiness" might be tokenized as "un", "happi", and "ness")

When I hear rain on my roof, I _______ in my kitchen.

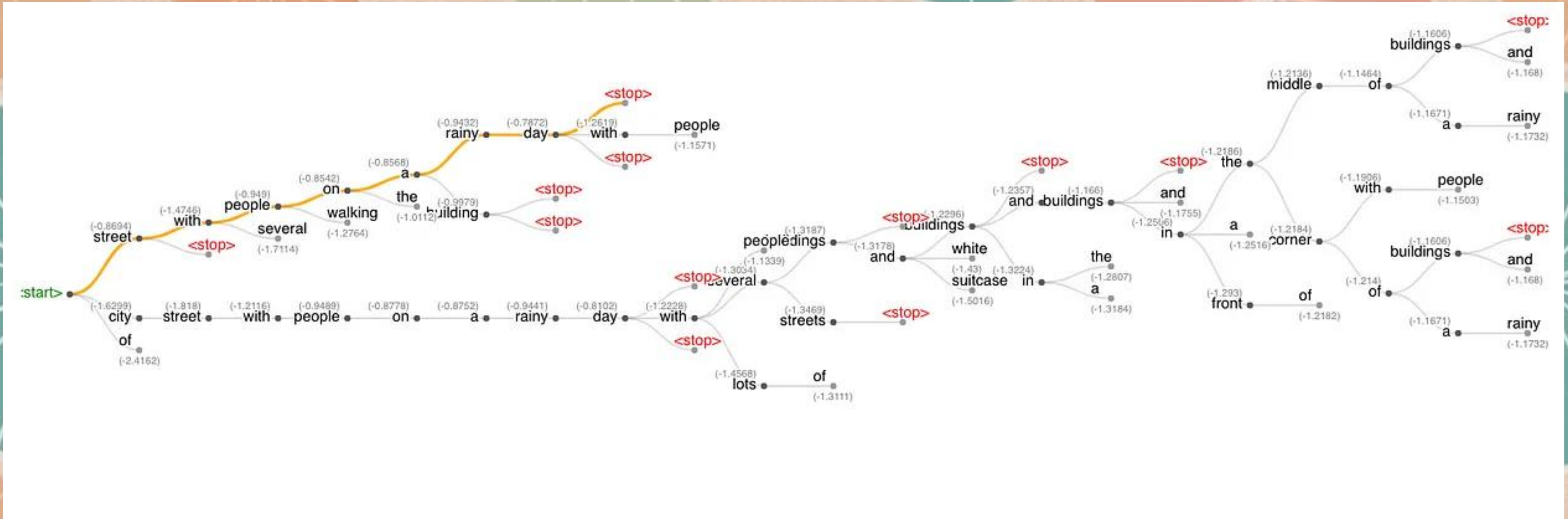
Probability	**Token(s)**
9.4%	cook soup
5.2%	warm up a kettle
3.6%	cower
2.5%	nap
2.2%	relax

- "How much does each other token of input affect the interpretation of this token?“

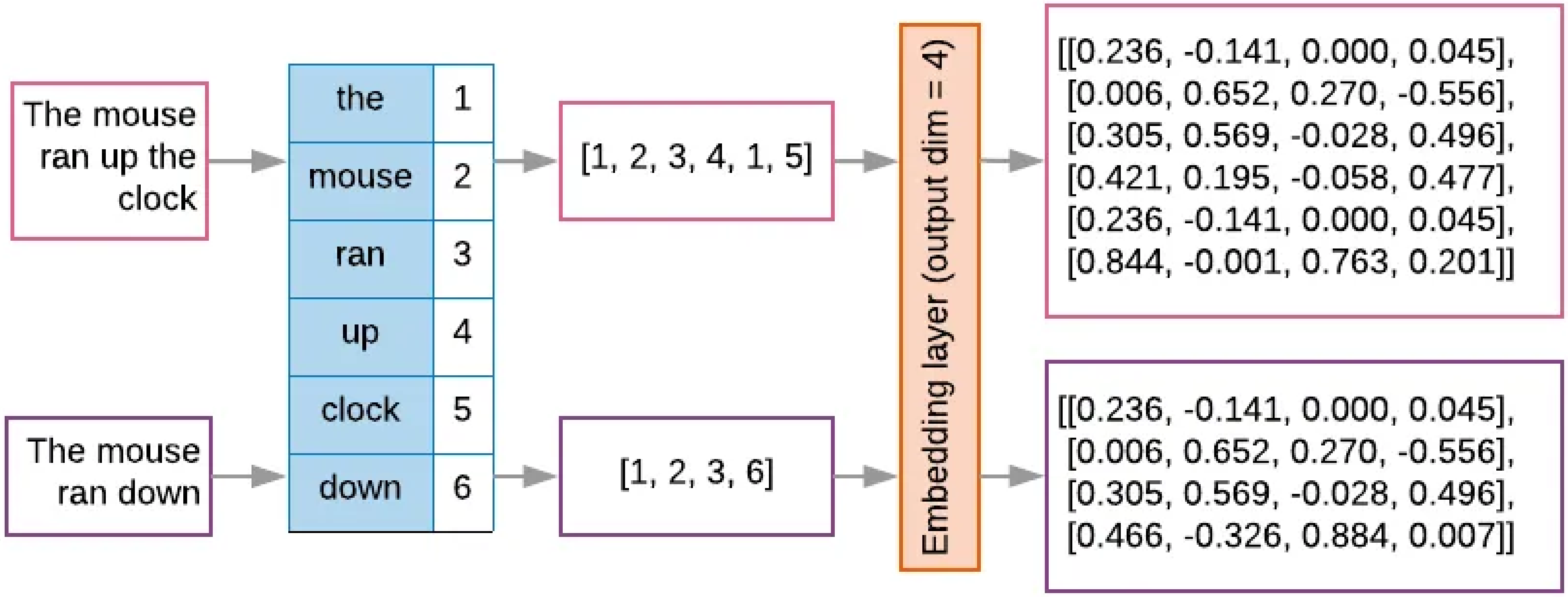
=> **Self-Attention**



https://projector.tensorflow.org/



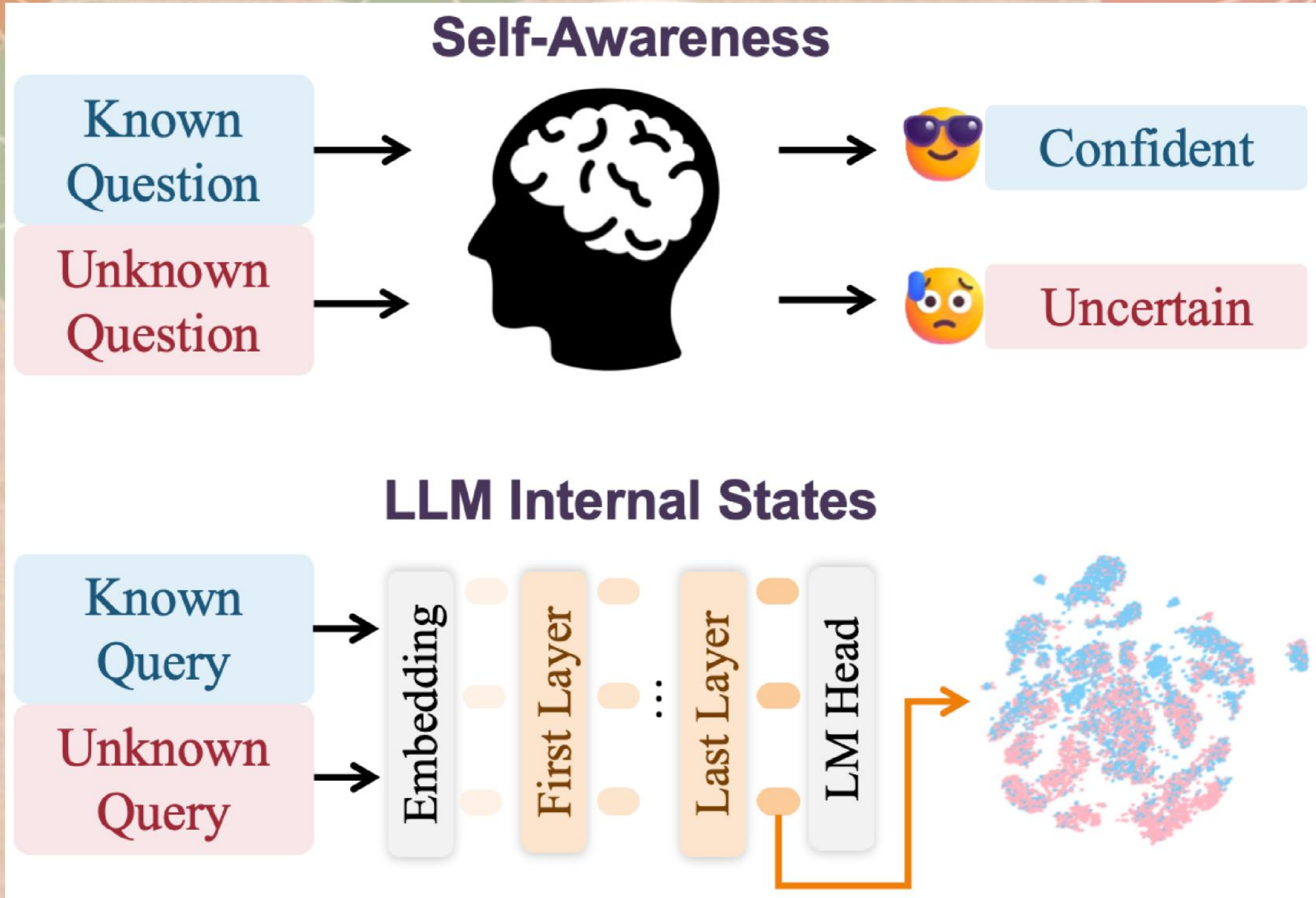
https://cohenori.medium.com/llm-token-economy-optimization-d1c3feea880b



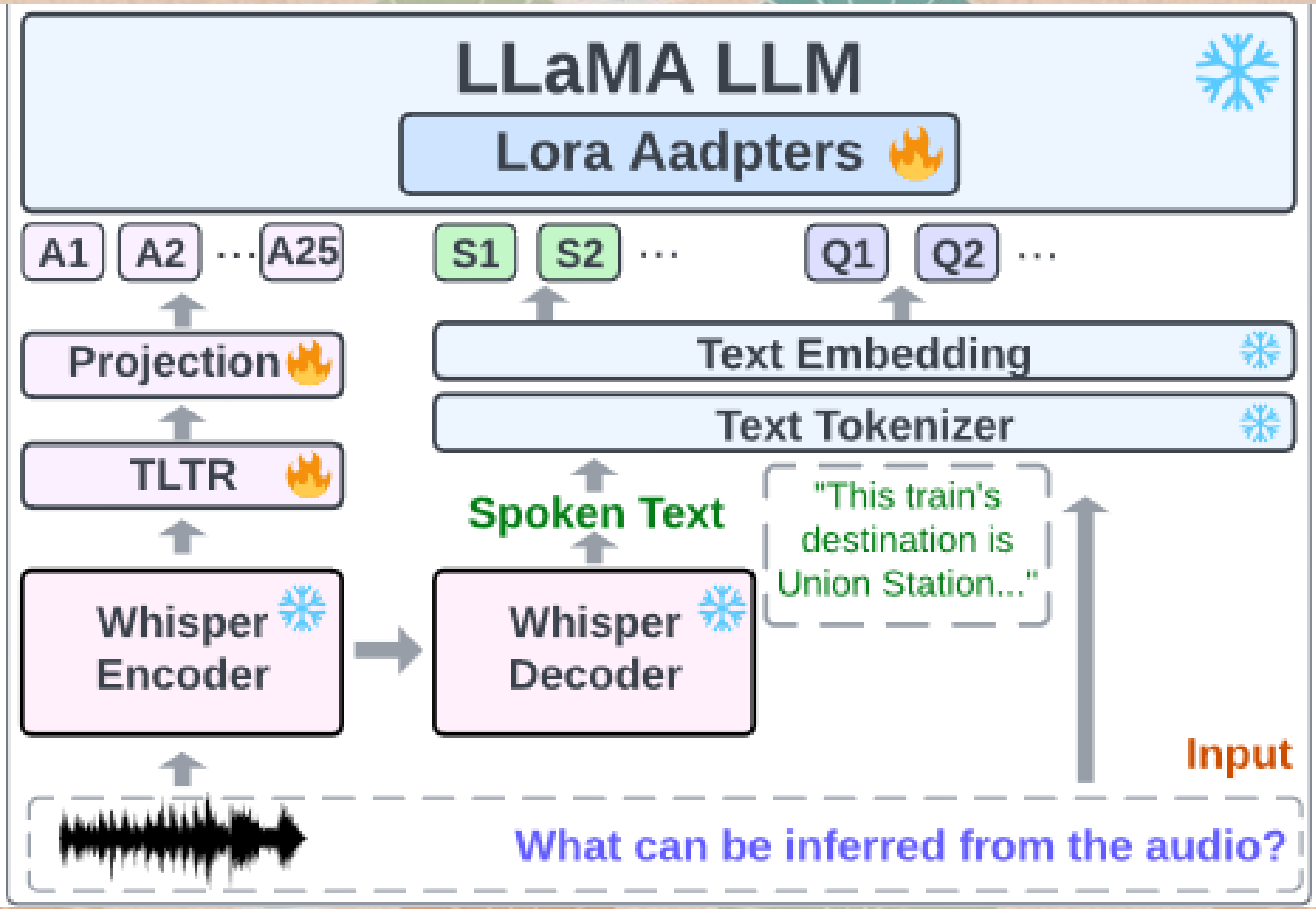
https://geoffrey-geofe.medium.com/

Hallucinations

- Phenomenon in which an LLM generates information or responses that are factually incorrect, nonsensical, or fabricated but may sound plausible and coherent



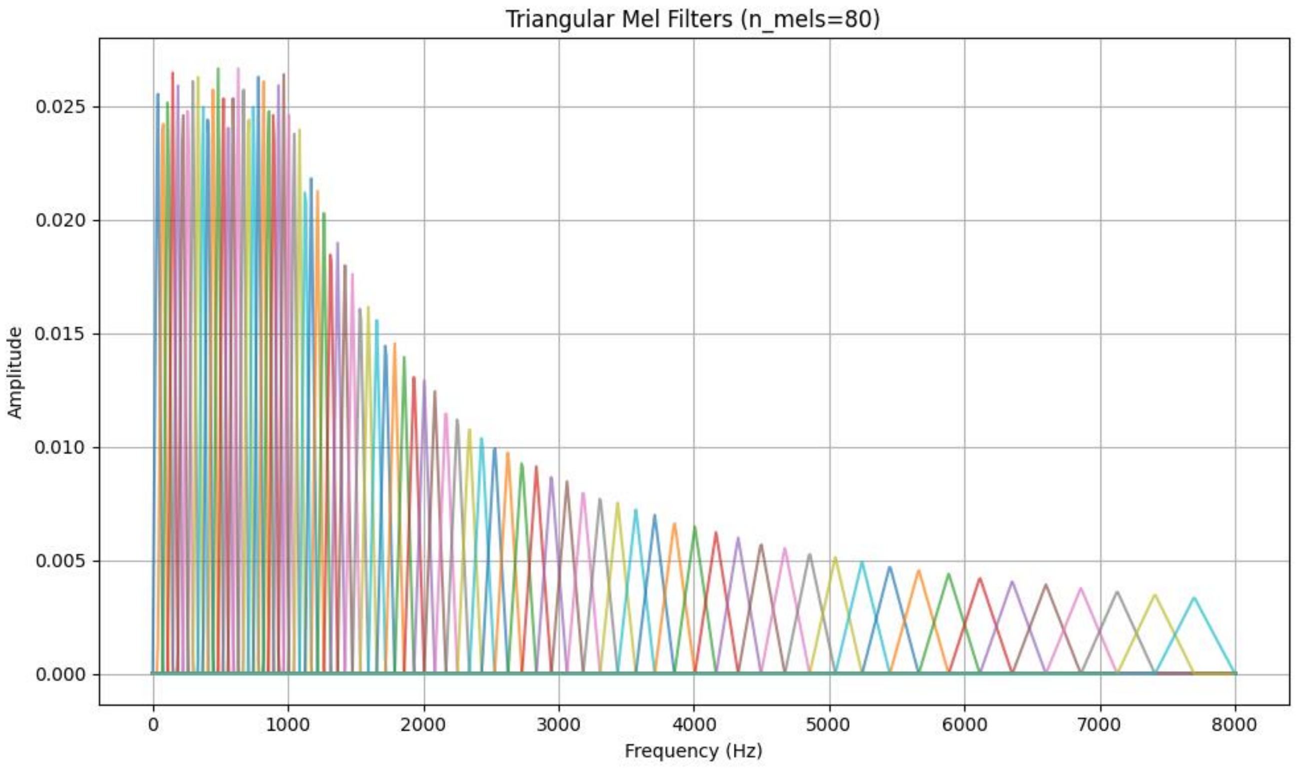
JI et al. 2024: https://arxiv.org/abs/2407.03282v1

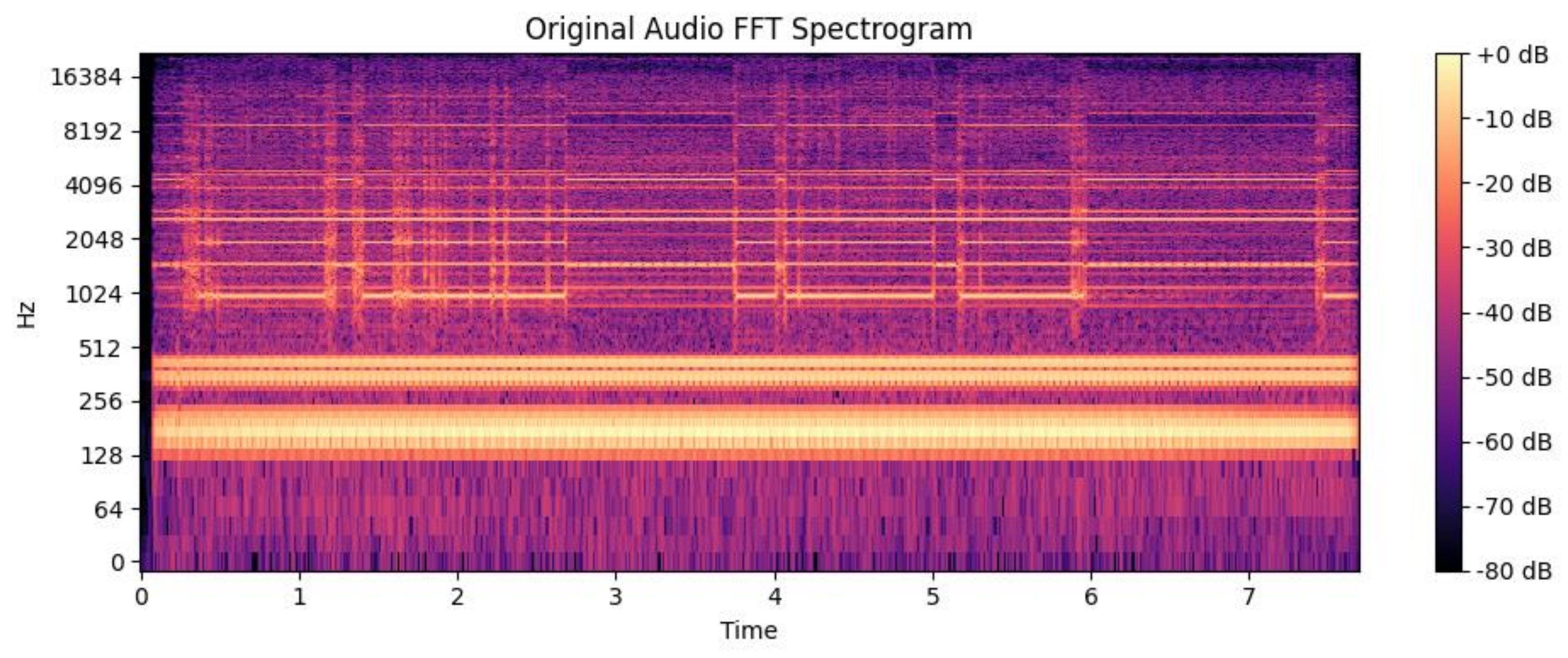


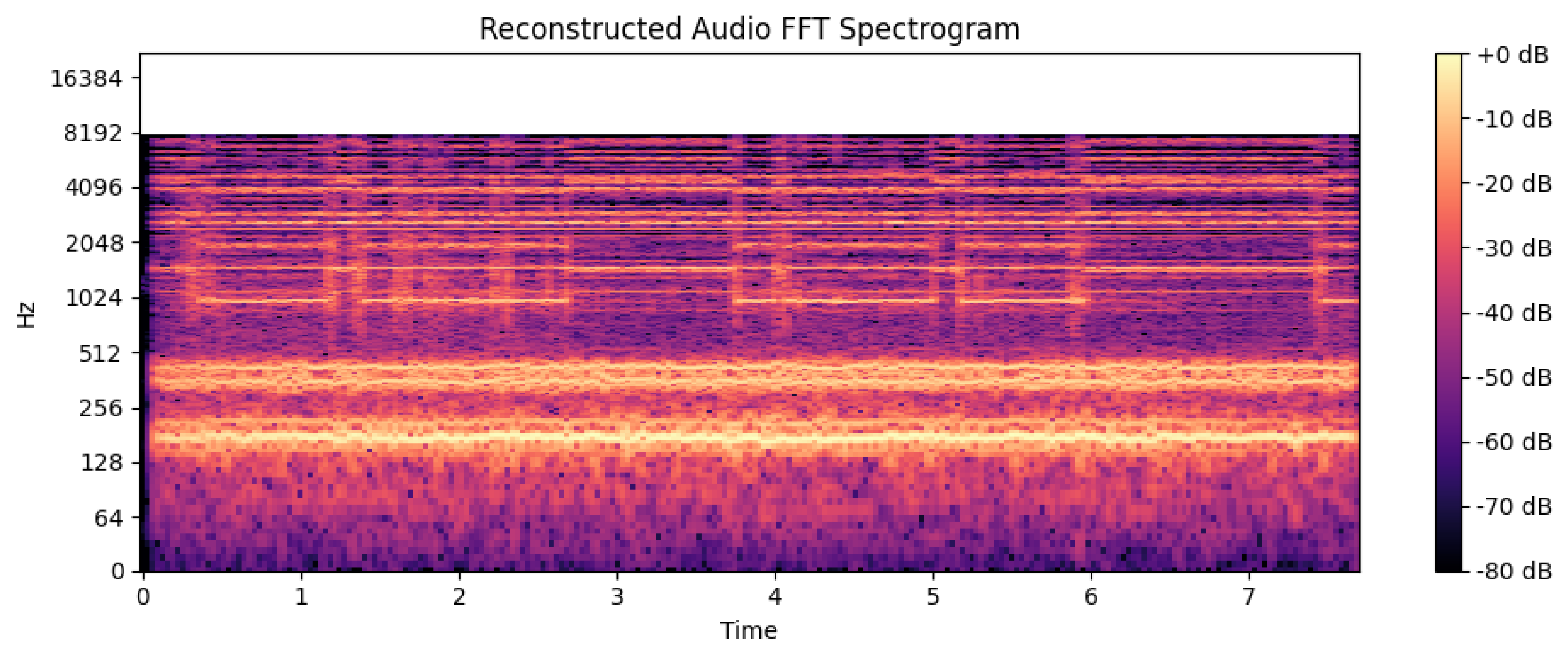
LTU-AS Structure, Gong et al. 2023: https://arxiv.org/pdf/2309.14405

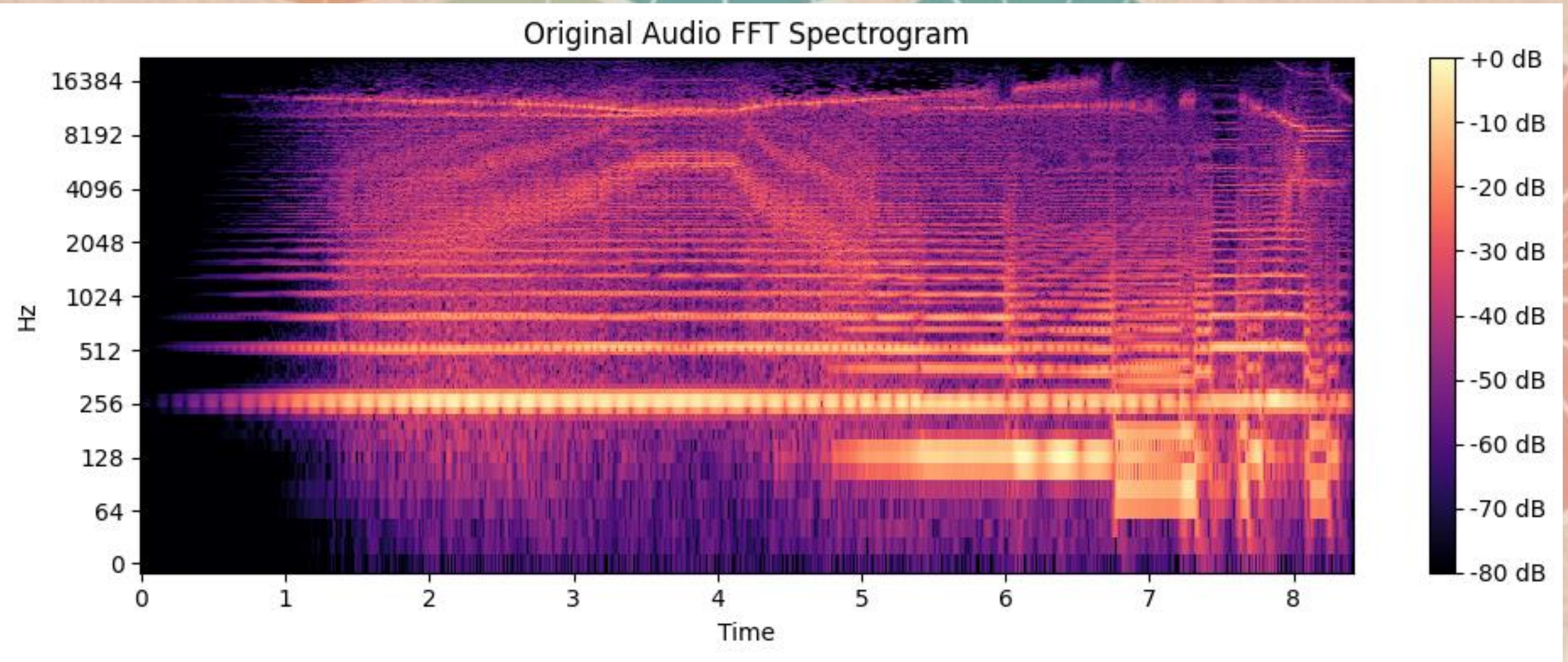
- What happens with audio within Encoder?
  - **HUGE quality loss**
  - Examples in *whisper_reconstruct.ipynb* notebook

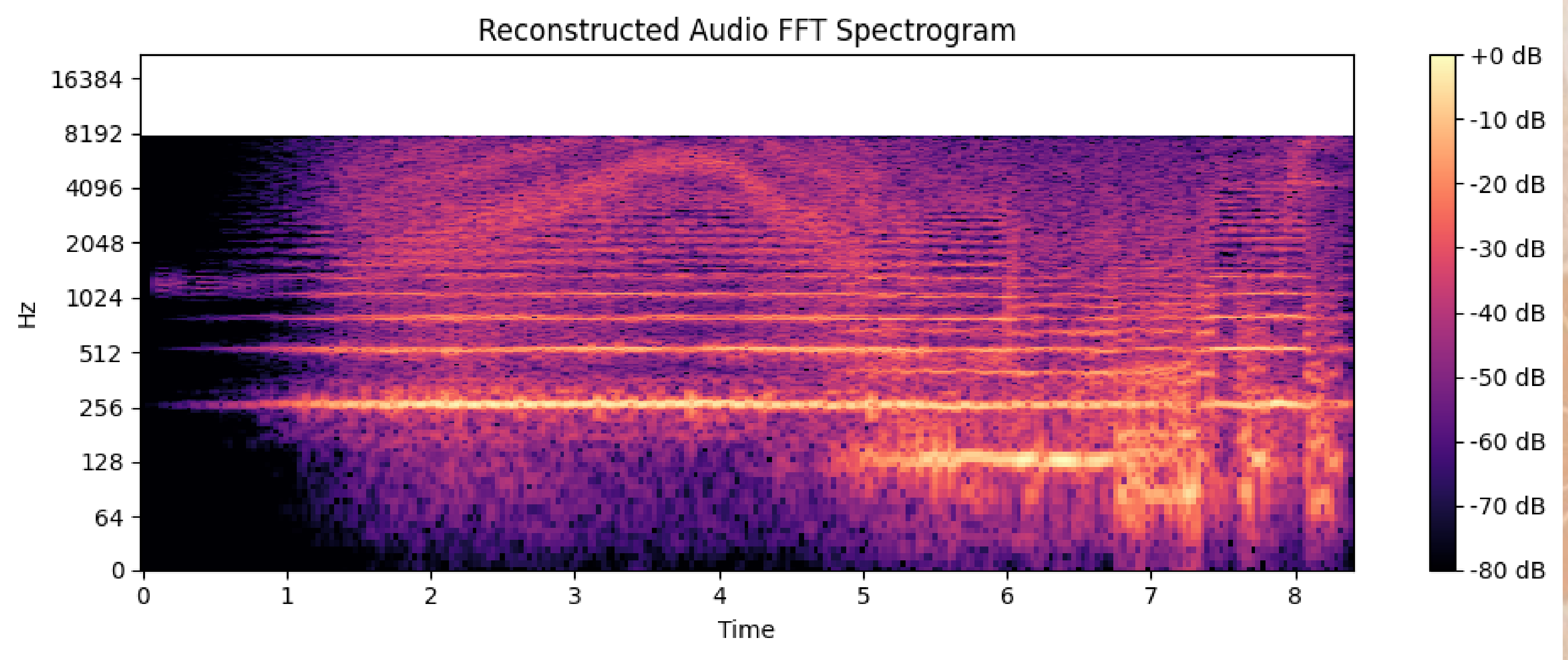
All audio is re-sampled to 16,000 Hz, and an 80-channel log-magnitude Mel spectrogram representation is computed on 25-millisecond windows with a stride of 10 milliseconds.

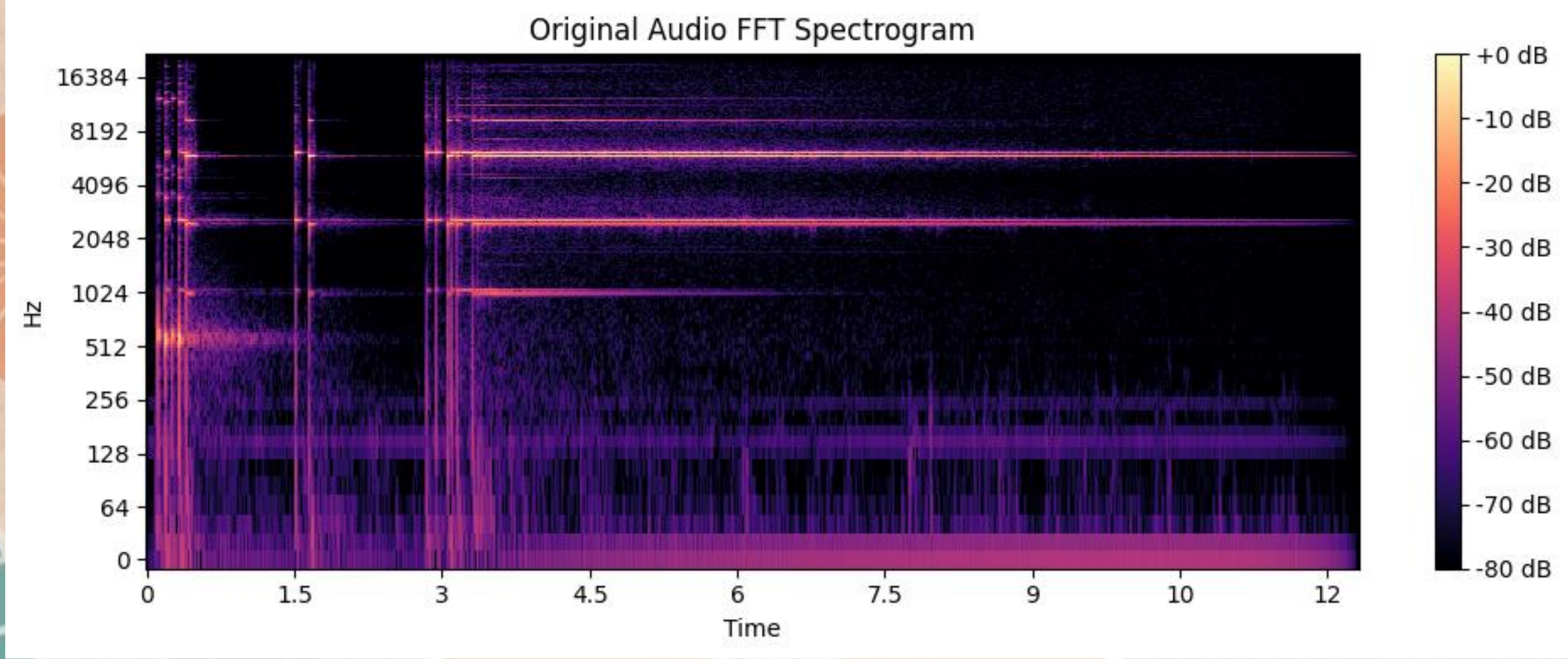


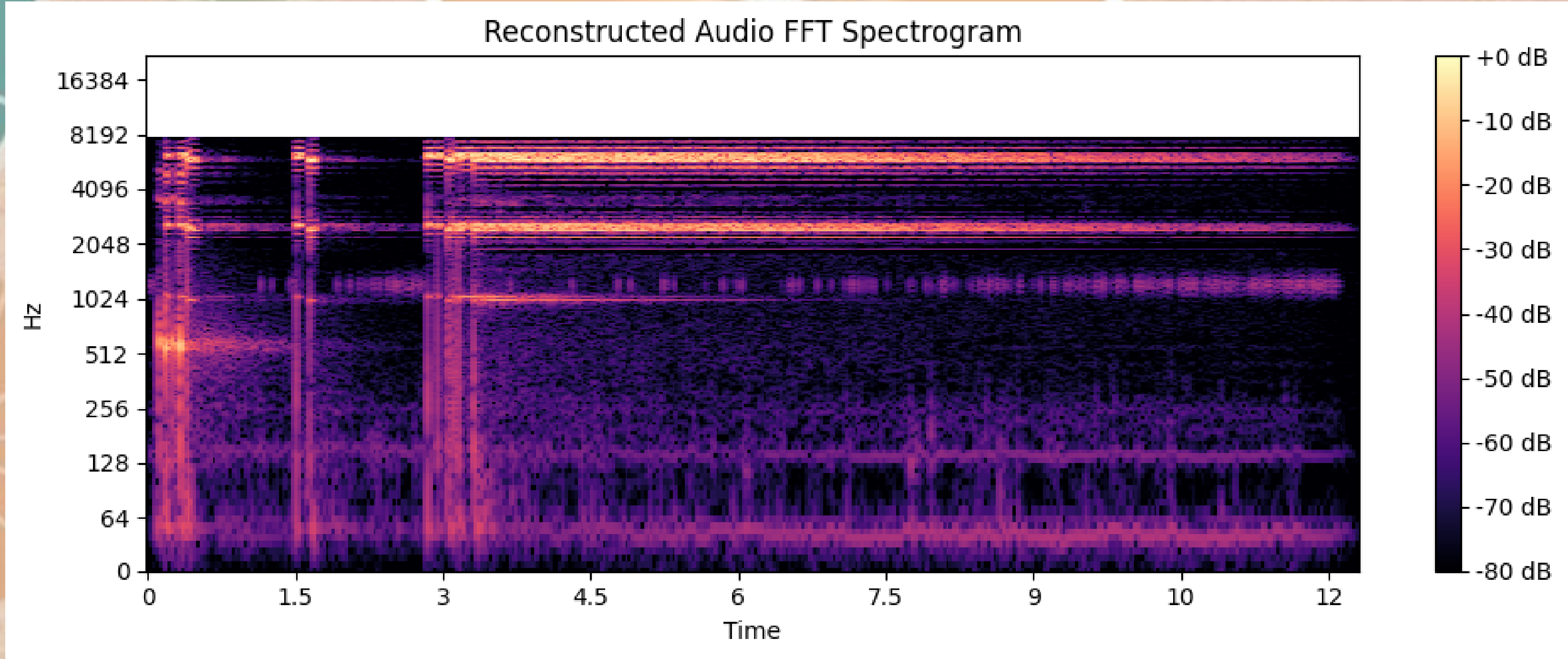

Original Audio FFT Spectrogram
Hz
16384
8192
4096
2048
1024
512
256
128
64
0
Time
0
1
2
3
4
5
6
7
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB


Reconstructed Audio FFT Spectrogram
Hz
16384
8192
4096
2048
1024
512
256
128
64
0
Time
0
1
2
3
4
5
6
7
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

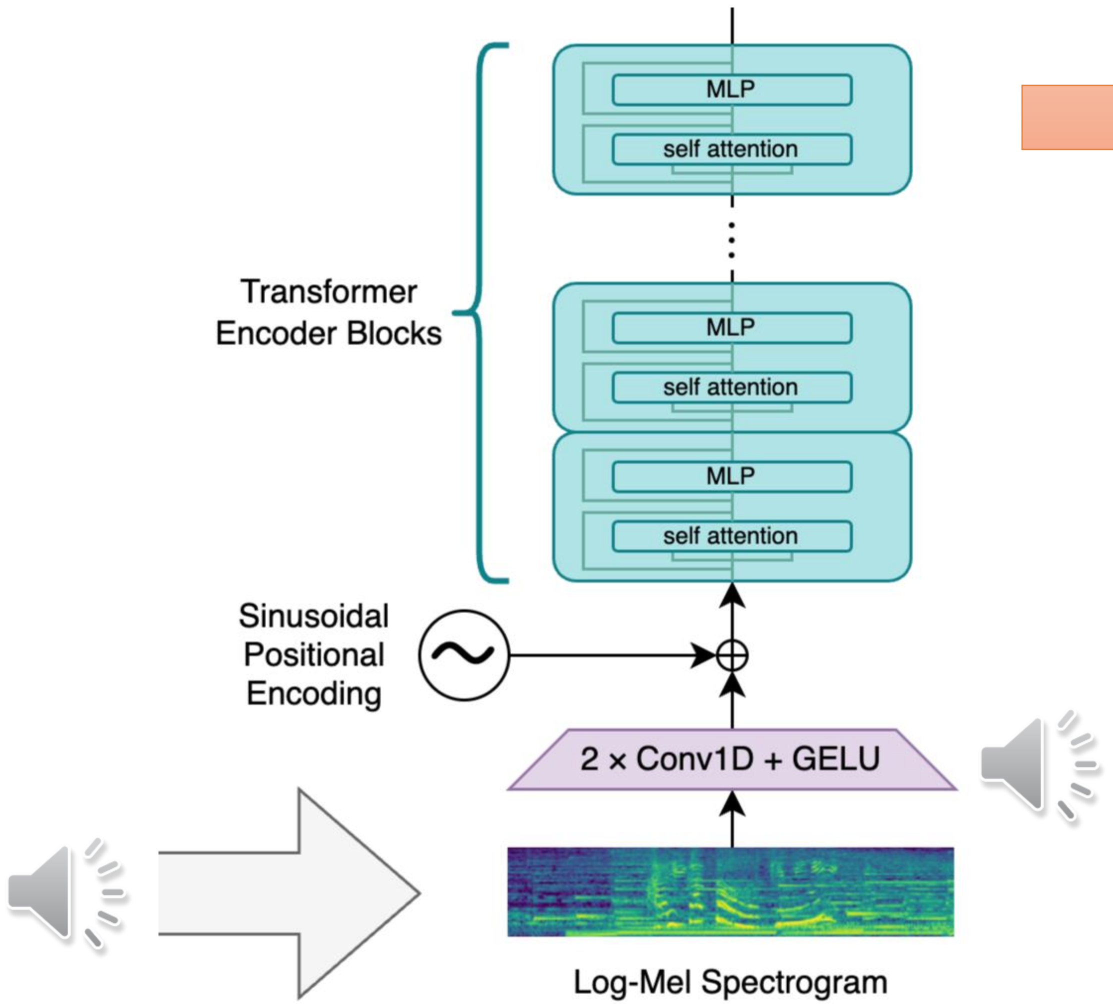

Original Audio FFT Spectrogram
Hz
16384
8192
4096
2048
1024
512
256
128
64
0
0
1
2
3
4
5
6
7
8
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB


Reconstructed Audio FFT Spectrogram
Hz
16384
8192
4096
2048
1024
512
256
128
64
0
0
1
2
3
4
5
6
7
8
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

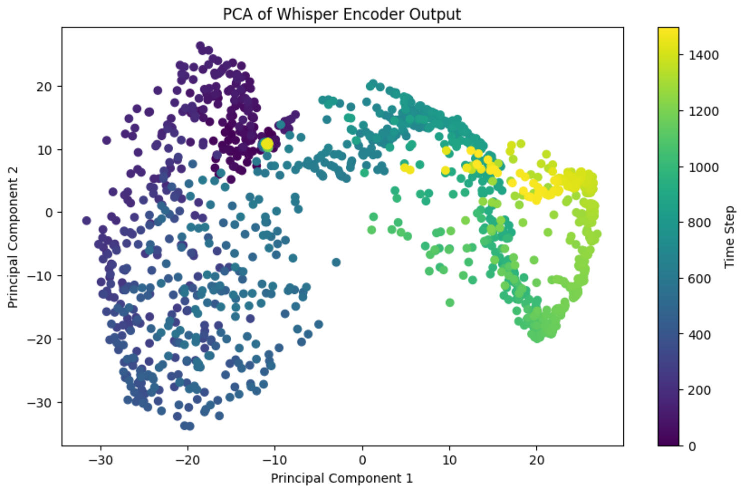

Original Audio FFT Spectrogram
16384
8192
4096
2048
1024
512
256
128
64
0
Hz
0
1.5
3
4.5
6
7.5
9
10
12
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB


Reconstructed Audio FFT Spectrogram
16384
8192
4096
2048
1024
512
256
128
64
0
Hz
0
1.5
3
4.5
6
7.5
9
10
12
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

- OpenAI Whisper encoder dataflow




```
torch.Size([1, 1500, 1280])
tensor([[[-5.9789e-01,  1.3652e-01, -8.5596e-01,  ..., -7.8826e-02,
          -1.2007e+00, -1.9048e+00],
         [ 5.7440e-01, -1.5710e-03, -4.3474e-01,  ..., -9.0926e-01,
          -1.3070e+00, -1.9070e+00],
         [ 2.2129e-02, -2.0025e-01, -1.2897e-01,  ..., -6.6805e-01,
          -1.3218e+00, -1.7689e+00],
         ...,
         [ 2.6864e-02, -5.8206e-01,  1.2360e-01,  ...,  7.3163e-01,
          -2.5886e-01,  2.2350e-01],
         [-5.1238e-02, -3.0040e-01, -3.1124e-02,  ...,  8.1508e-01,
          -2.9442e-01,  1.9190e-01],
         [-1.9422e-01,  1.0346e-01, -1.7967e-01,  ...,  9.2487e-01,
          -2.4711e-01,  1.6967e-02]]], device='cuda:0')
```



Datasets (over 1m examples total)

“A child shouts, and an adult male speaks, while an emergency vehicle siren sounds with the horn blowing”

https://research.google.com/audioset/index.html

„playing snare drum“

AQA LLM Training

- **Synthetic Dataset Creation**
  - ChatGPT uses reasoning ("hallucinates„) to generate hypothetical scenarios where specific audio cues might be present
  - The model leverages both reasoning and creativity to infer what could happen in a particular audio environment
- **Sources for Data Generation**
  - Audio descriptions from an existing dataset provide a foundational reference for the hallucinated scenarios
  - Whisper decoder output is used to extract detailed information from the audio, offering a more refined input for generating plausible descriptions
- **Training Objective**
  - The goal is to create a synthetic dataset that can train large language models (LLMs) to analyze and describe audio in a way that integrates reasoning and contextual awareness, going beyond basic feature extraction or classification
  - By simulating hypothetical settings with relevant audio cues, this approach aims to enhance the LLM's ability to understand and interpret complex audio scenes

```
{'time': {'start': 0, 'end': 2}, 'audio tags': [('Music', -0.4527301490306854), ('Speech', -0.888166606426239), ('Silence', -1.8608324527740479)]}
{'time': {'start': 2, 'end': 4}, 'audio tags': [('Music', 0.9356539845466614), ('Speech', -0.15668557584285736)]}
{'time': {'start': 4, 'end': 6}, 'audio tags': [('Music', 0.7164449095726013), ('Speech', 0.38899049162864685)]}
{'time': {'start': 6, 'end': 8}, 'audio tags': [('Music', 0.40905559062957764), ('Speech', -0.6937562227249146)]}
{'time': {'start': 8, 'end': 10}, 'audio tags': [('Music', 1.6711302995681763), ('Speech', 0.5111501812934875), ('Drum kit', -1.738829493522644), ('Drum', -1.938528180l223755)]}
{'time': {'start': 10, 'end': 12}, 'audio tags': [('Music', 1.3954704999923706), ('Speech', -0.2355932742357254)]}
{'time': {'start': 12, 'end': 14}, 'audio tags': [('Music', 0.7912232875823975), ('Speech', -1.9295458793640137)]}
{'time': {'start': 14, 'end': 16}, 'audio tags': [('Music', 1.0189353227615356), ('Speech', -1.862793207168579)]}
{'time': {'start': 16, 'end': 18}, 'audio tags': [('Music', 1.245056390762329), ('Rock and roll', -1.1084270477294922)]}
{'time': {'start': 18, 'end': 20}, 'audio tags': [('Music', 1.1750296354293823), ('Rock and roll', -1.69615638256073)]}
```


Whisper-AT audio tagging output

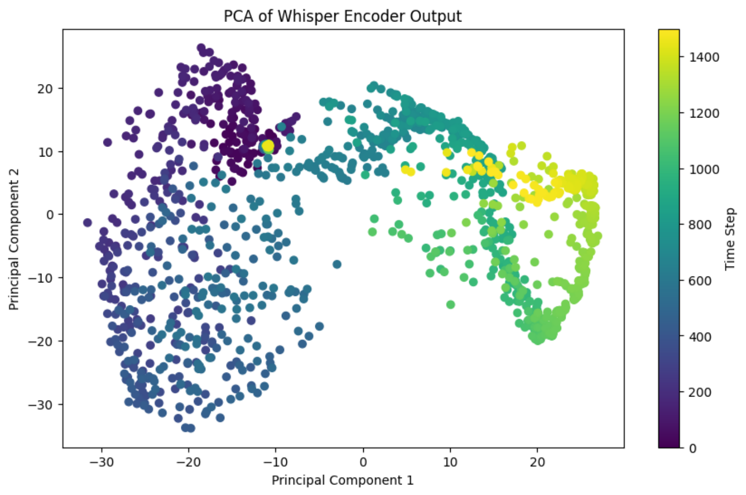
GPT-Input: Music genre: Folk; Music Lyrics: "The man of his dreams is all that he seems Flowers in sunshine..."
GPT-Output: Q: What mood does this music convey?
A: The music conveys a cheerful and relaxed mood because of the lively melody and lyrics about flowers and sunshine.

GPT-Input: In the recording, background sound of explosion and speech of "Flames engulfed the tank and began to heat the propane inside. Twenty volunteer firefighters..." is heard.
GPT-Output: Q: Why do the speech content and explosion sound create a tense atmosphere? A: The speech content describes a tragic event and the explosion sound adds intensity and suspense to the narrative.

AQA pairings from LTU-AS dataset

What exactly are we getting as the output?

Heavily distorted audio used as an analysis source (max. 10 sec)



Embedding space fine-tuned for audio classification based on annotated YouTube videos

Embedding space fine-tuned using ChatGPT-generated hypothetical scenarios where specific audio cues might be present

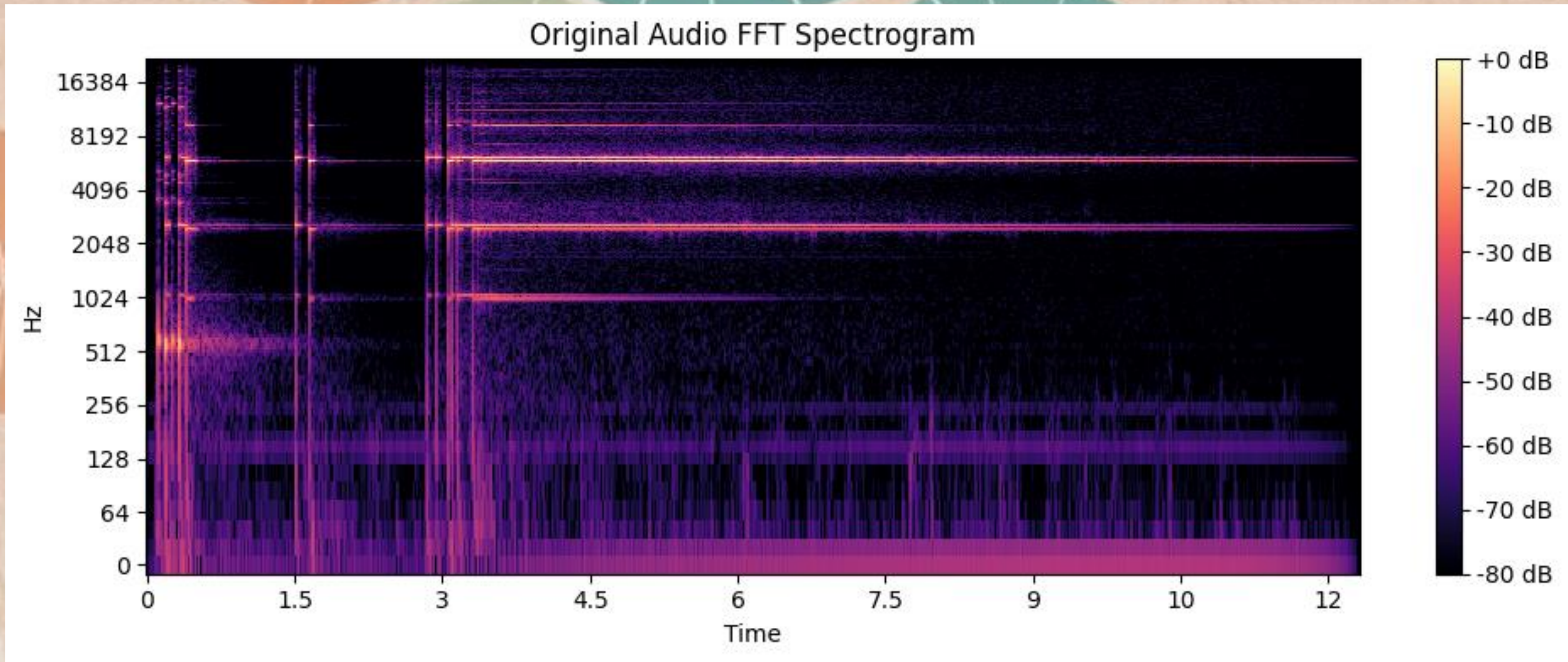
Another options?

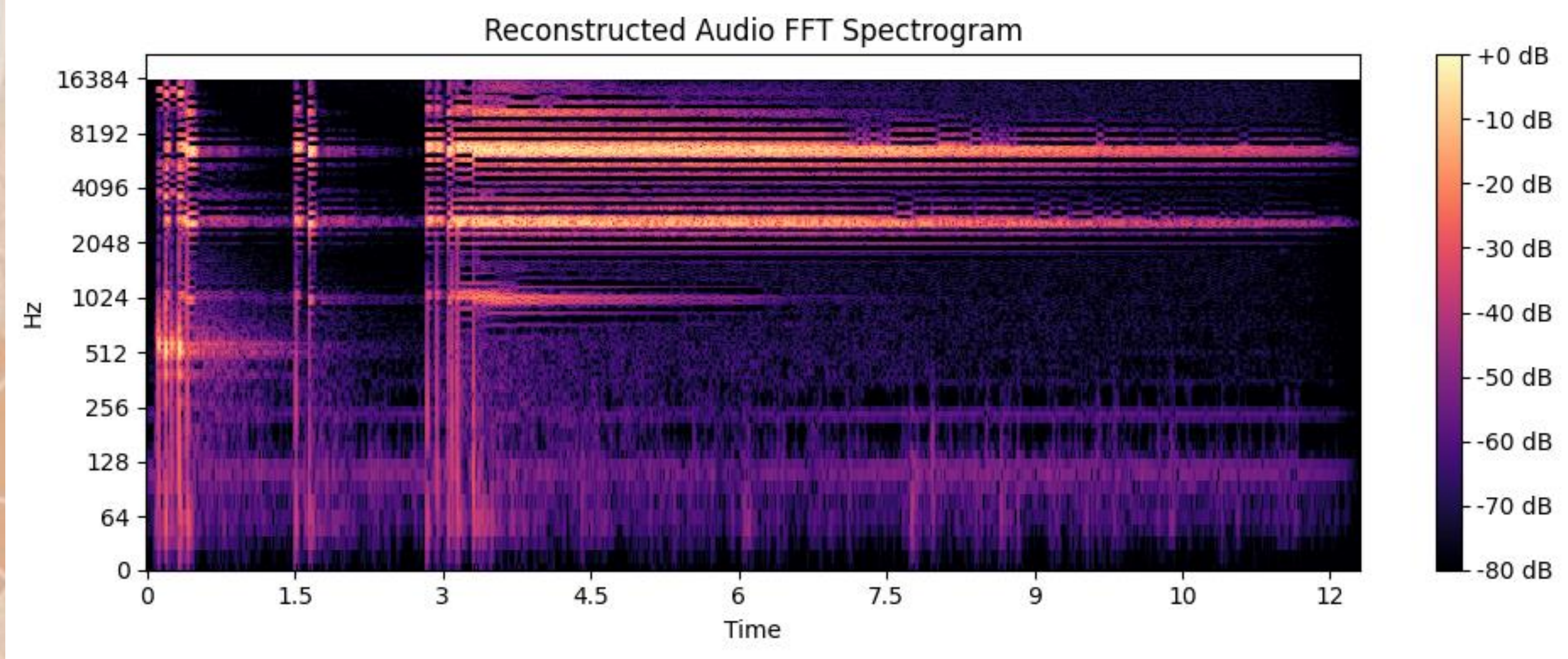
- **CLAP (Elizalde et al., 2022)**
  - Currently (one of the) best performing audio classifiers (98% on ESC-50)
  - Features higher signal bandwidth in the encoder compared to Whisper
  - Maximum sample length of 5 seconds

- **LLark (Gardner et al., 2024)**
  - Promising new model from Spotify, designed specifically for music understanding
  - Capable of analyzing audio segments up to 25 seconds in length
  - Analysis Approach: Waveform-based, based on OpenAI's Jukebox, with a resolution of 44.1 kHz

- **Common Limitations**
  - Neither model has online accessible inference
  - Requires server-grade hardware and a lot of python experience to run

CLAP audio encoder





Evaluation of multimodal LLMs for analysis?

- **Spectral Terms**:
  - Brightness/Darkness: Referring to the high or low frequency content of a sound

  => possible, but restricted by bandwidth
  - Density: The concentration of frequencies within the spectrum

  => possible, but restricted by bandwidth
  - Texture: The layering or blending of different spectral elements

  => possible (even favorable, esp. if elements can be easily classified as separate audio objects), but restricted by bandwidth, time resolution and context window (5-10 seconds are not enough)
- **Morphological Terms**:
  - Attack: The onset of a sound, which can be sharp, soft, gradual, etc.

  => possible, but restricted by bandwidth and time resolution
  - Decay: The fading or termination of a sound

  => possible, but restricted by bandwidth and time resolution
  - Growth: The development or increase in a sound's intensity or complexity over time

  => possible, but restricted by bandwidth and time resolution
  - Gesture: The dynamic movement of sound, often perceived as having a physical or directional component

  => Possible (even favorable, esp. if elements can be easily classified as separate audio objects), but restricted by bandwidth, time resolution and spatial understanding (most current models are monophonic by design)

- **Behavioral Terms**:
  - Motion: The perceived movement of sound in space or its evolution over time

  => possible, but restricted by bandwidth, time resolution and spatial awareness
  - Flow: How smoothly or abruptly a sound transitions or changes

  => possible, but restricted by bandwidth and time resolution
  - Grain: The perceived texture or granularity of a sound

  => missing training data, also restricted by bandwidth and time resolution
- **Source-Cause Terms**:
  - Mimesis: The imitation of real-world sounds

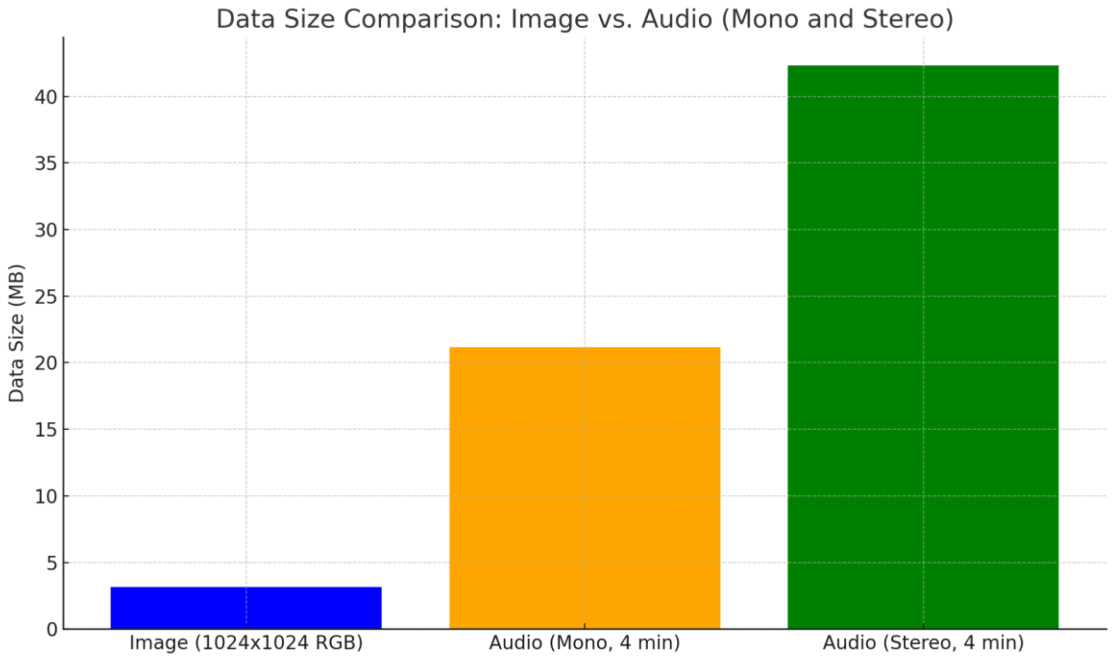
  => can work exceptionally well with current models
  - Immateriality: Describing sounds that do not suggest a clear physical source

  => can work exceptionally well with current models

Problems

- **High Computational Requirements**
  - Training or fine-tuning large-scale models, like LTU-AS, on audio data requires powerful hardware, often necessitating **cloud GPU services**, which can be costly

We consider music in the raw audio domain represented as a continuous waveform $\mathbf{x} \in [-1, 1]^T$, where the number of samples T is the product of the audio duration and the sampling rate typically ranging from 16 kHz to 48 kHz. For music, CD quality audio, 44.1 kHz samples stored in 16 bit precision, is typically enough to capture the range of frequencies perceptible to humans. As an example, a four-minute-long audio segment will have an input length of $\sim$10 million, where each position can have 16 bits of information. In comparison, a high-resolution RGB image with 1024×1024 pixels has an input length of $\sim$3 million, and each position has 24 bits of information. This makes learning a generative model for music extremely computationally demanding with increasingly longer durations; we have to capture a wide range of musical structures from timbre to global coherence while simultaneously modeling a large amount of diversity.



Dhariwal et al. 2020

- **Technical Expertise**
  - Many of the current state-of-the-art audio models are implemented in Python and require a deep understanding of both Python programming and machine learning libraries like PyTorch or TensorFlow
  - This high barrier of technical expertise excludes many potential users, particularly in fields like musicology, who may not have sufficient programming experience
- **Lack of dedicated Datasets for Electroacoustic Music Analysis**
  - There is a lack of dedicated datasets for genres like electroacoustic music, which has its own unique sonic characteristics
  - Existing general-purpose datasets may not capture the full range of sounds in electroacoustic compositions, leading to suboptimal performance in genre-specific tasks
- **Copyright Issues with Original Recordings**
  - All essential works in electroacoustic music are still under copyright protection, raising challenges for using original recordings in research or model training
  - The limited access to such recordings hampers the creation of large, high-quality datasets for this genre
- **Suboptimal Audio Features (MFCC):**
  - Many current models still rely on **Mel Frequency Cepstral Coefficients (MFCC)**, a low-bandwidth encoding often used in speech recognition, but **suboptimal** for capturing the full richness of musical timbre, especially in electroacoustic music

Solutions?

- **Streamlined Research in Humanities:**
  - Multimodal LLM research requires a capable local CUDA setup or funds for cloud services. Both training and inference incur significant costs:
  - LTU-AS training/fine-tuning costs **several thousand dollars** on cloud computing, based on current server prices.
  - Inference costs: **$25–50 per day** for smaller models; hundreds per day for larger models.
  - Without accessible research funding, especially in humanities, LLM integration/refinement becomes inefficient.
  - Training smaller/downscaled models on platforms like Paperspace Gradient offers affordable options ($8–39 GPU flat-rate subscriptions), but takes significantly more time due to restrictions and limited resources.
- **Dedicated Datasets:**
  - Creating datasets for electroacoustic music is constrained by copyright issues.
  - Use synthetic datasets to capture basic timbre, motion, and spatial features (leveraging available plugins).
  - Collaborate with composers willing to provide works for AI training, ensuring clear legal agreements to avoid potential disputes.

Thank you!

egor.polyakov@hmt-Leipzig.de

https://egorpol.github.io/

https://github.com/egorpol/Electroacoustic-LLM-Analysis