

# Generating Complementary Instrument Tracks with a Transformer Model

---

Mikko Murhu

# Introduction

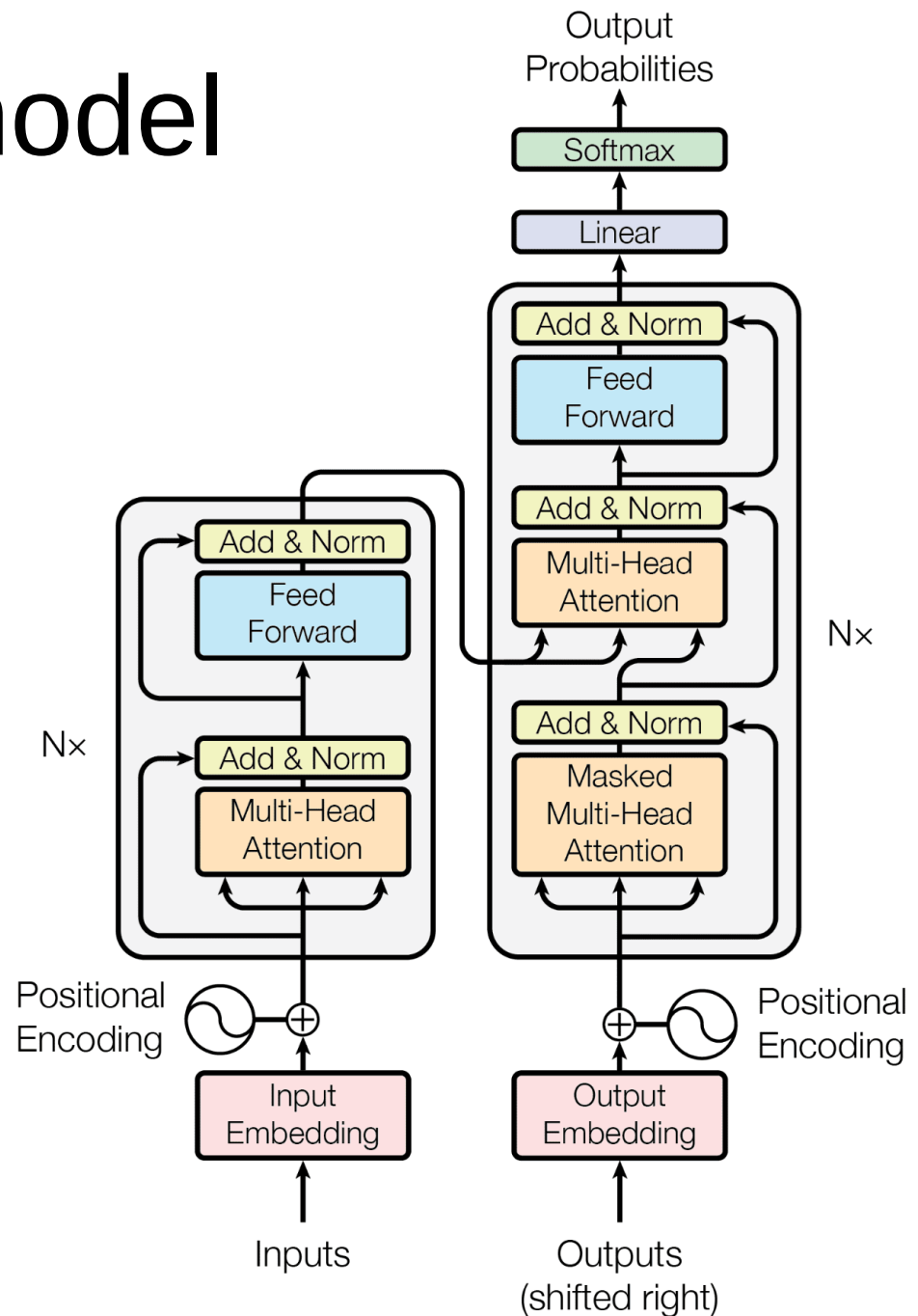
The objective: explore the capabilities of an encoder-decoder transformer model in generating complementary instrument tracks to existing instrument tracks

# Introduction – why transformers?

- Transformers are used in natural language processing for many purposes, including translation
- Translation converts instrument tracks to a set of accompanying instrument tracks
- Previously decoder-only transformers have been used for less conditioned music generation

# Transformer model

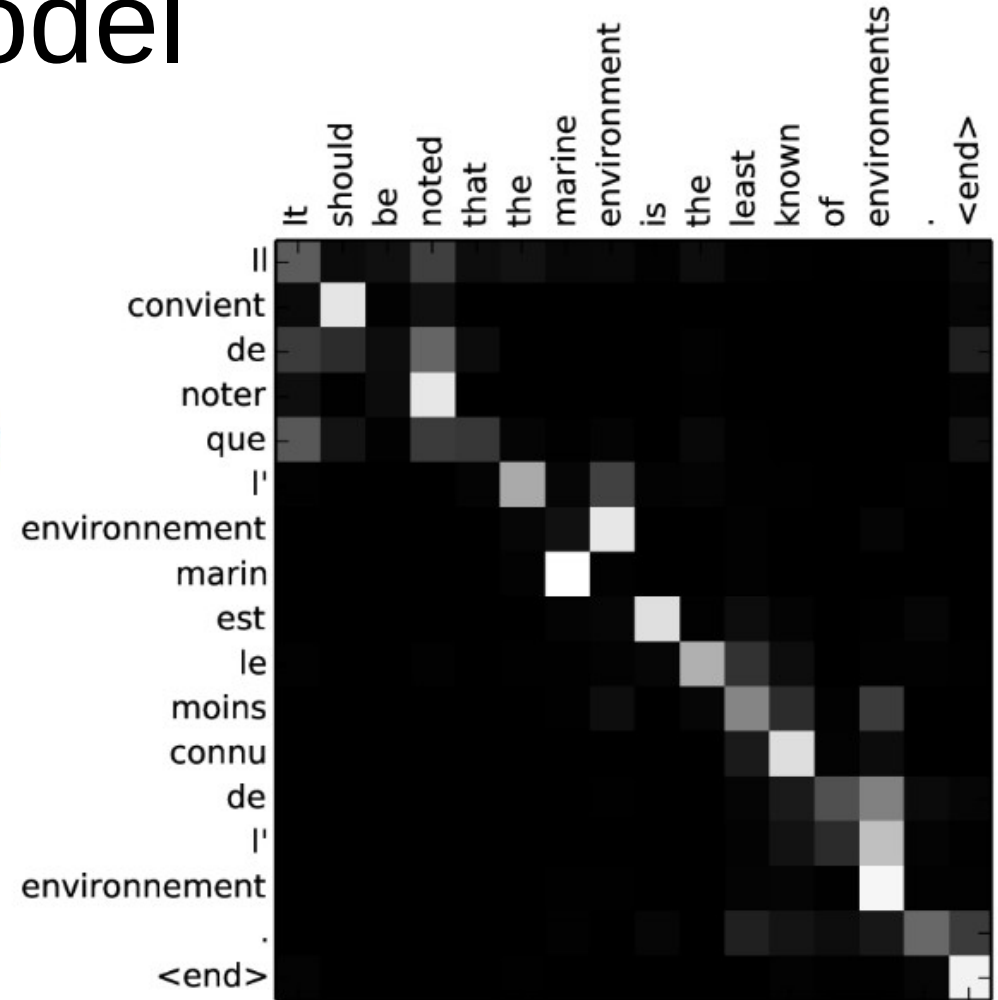
- A deep neural network
- Processes sequential data efficiently
- Replaces recurrency with attention mechanism and positional encoding



# Transformer model

## Attention mechanism

- Defined by
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$
- Calculates strengths of relationships of elements in sequences
- Three different contexts in an encoder-decoder transformer:
  - Self-attention in encoder
  - Self-attention in decoder
  - Cross-attention in decoder



An attention score heatmap

# Transformer model

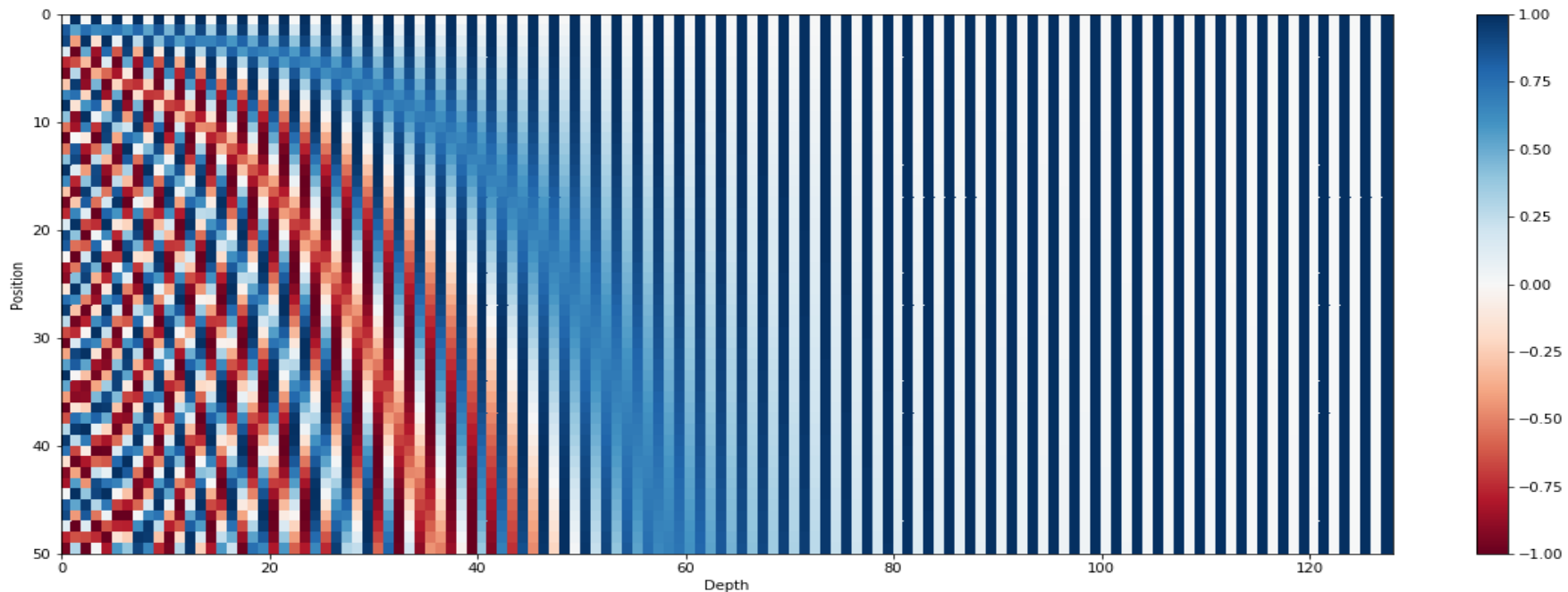
## Positional encoding

- A unique vector for each element in sequence

$$PE(pos, 2i) = \sin(pos/n^{(2i/d)})$$

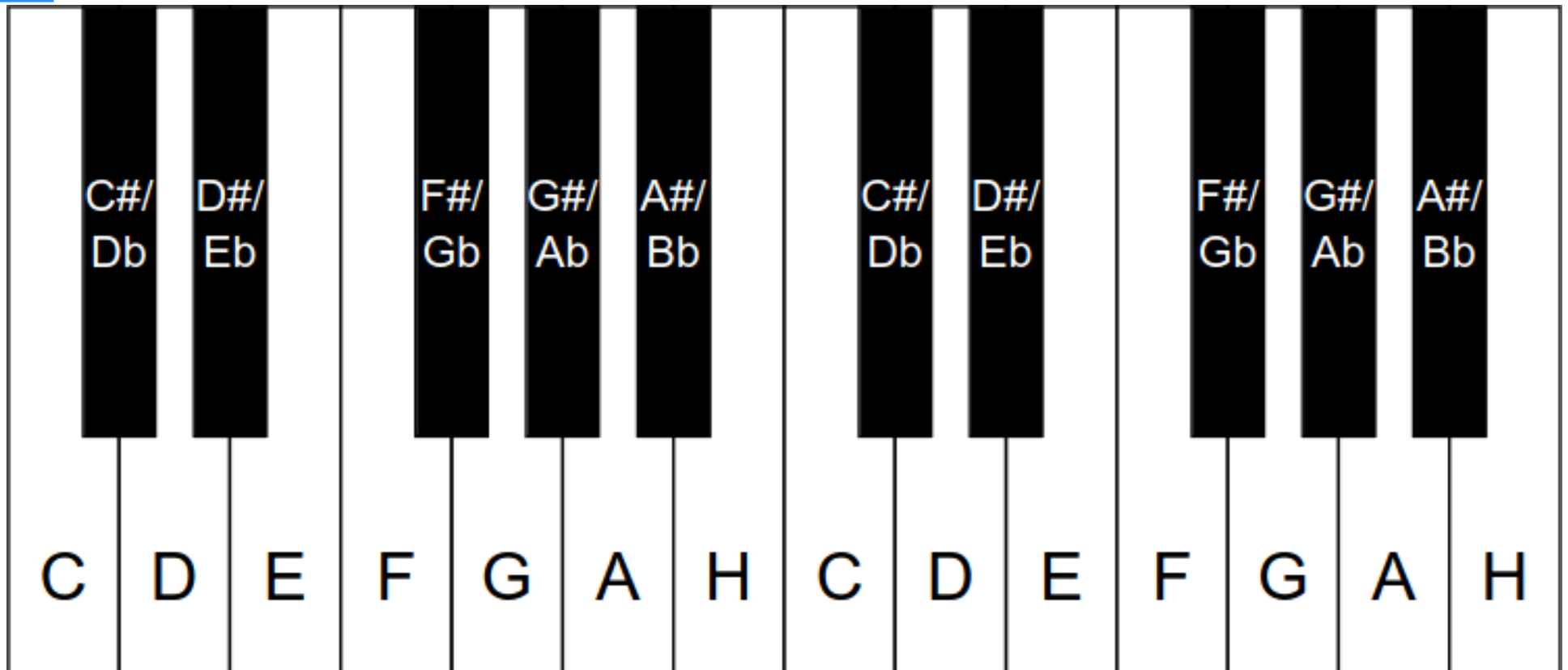
$$PE(pos, 2i + 1) = \cos(pos/n^{(2i/d)})$$

- Added to embedding vectors
- Injects positional information to the embedding



Positional encodings from positions 0 to 50 with model dimension 128.

# Elements of music



- 7 named pitches, including raised/lowered pitches create 12 unique pitches
- Harmony: key, scales, chords
- Rhythm: pulse, beat

# Elements of music

The diagram illustrates various elements of music notation across two systems of staves. The first system shows a piano score with a treble and bass clef, a 4/4 time signature, and a C-major scale. It includes a whole note, a half note with a C sharp accidental, a quarter note, a quarter rest, and eighth/sixteenth notes. The second system shows a key signature of two sharps (D major) and four major triads: D-major, G-major, A-major, and D-major. Labels include Treble clef, Bass clef, Time signature, C-major scale, Whole note, Half note, Quarter note, Quarter rest, Eighth/sixteenth notes/rests, Barline, 3rd bar, Accidentals, A flat, and A.

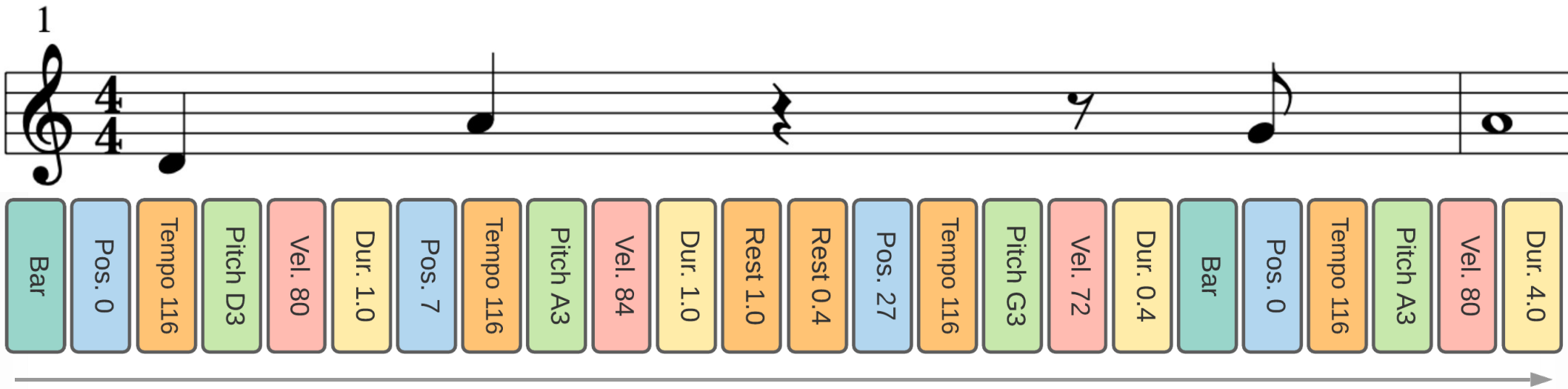
Music notation – sheet music



# Dataset

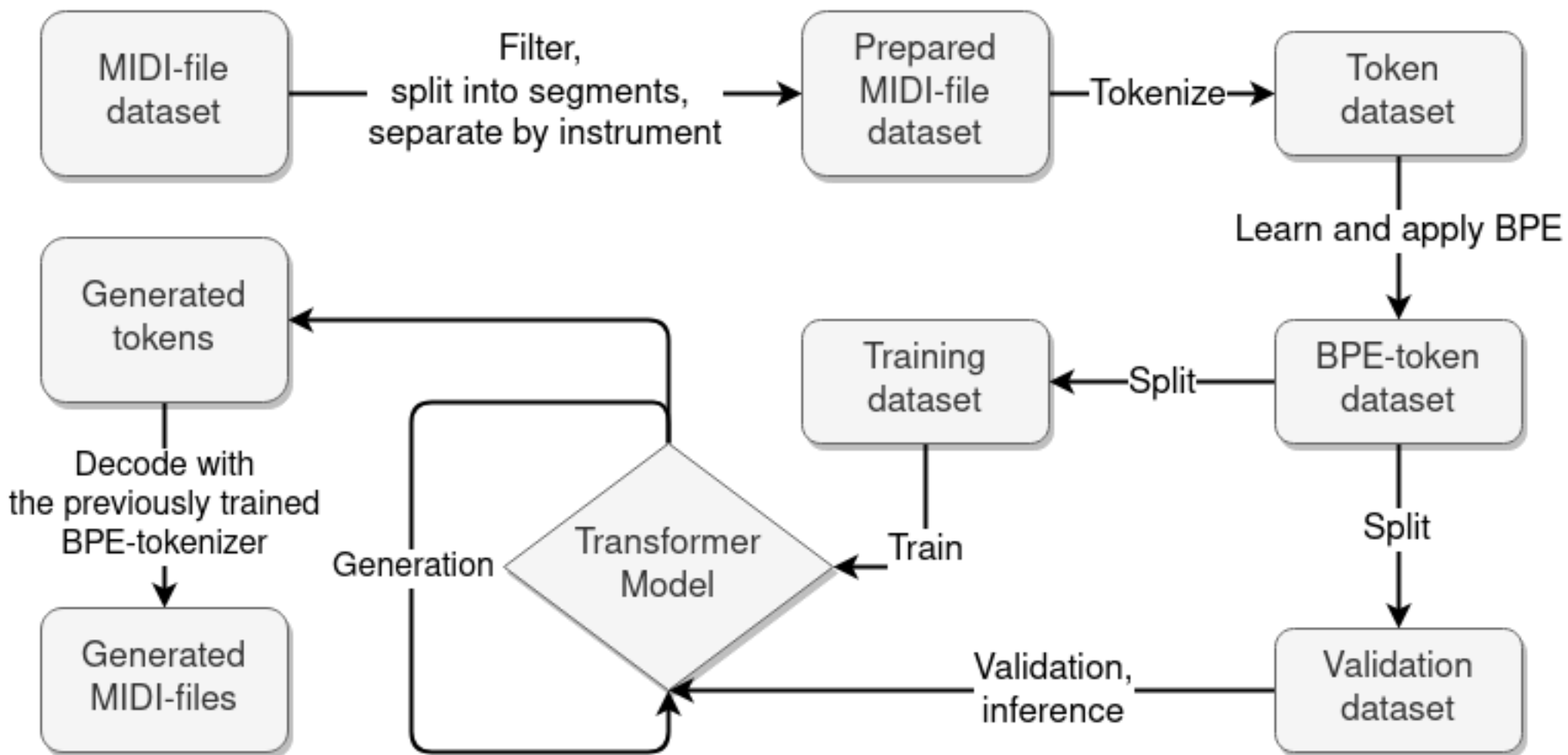
- MIDI files characterize music with discrete time-based events
- Lakh MIDI dataset has 176581 unique midi-files
- The thesis uses the "clean subset" of the Lakh dataset, which has 17233 midi-files

# Tokenization



- REMI-tokenization scheme
- Byte pair encoding (BPE)

# Steps in the pipeline



# Quantitative results

## Setup

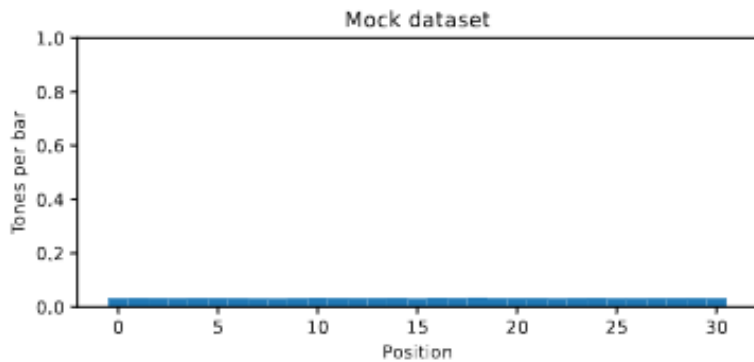
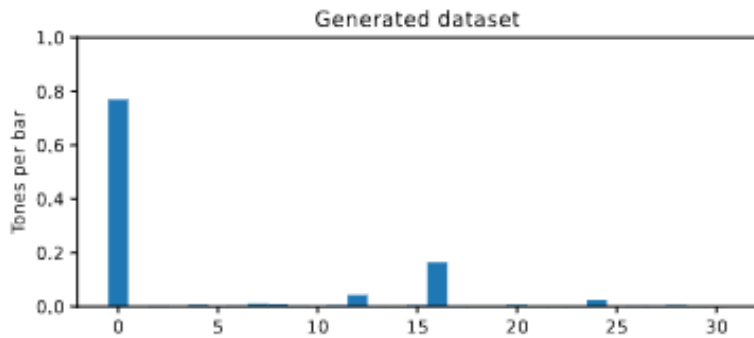
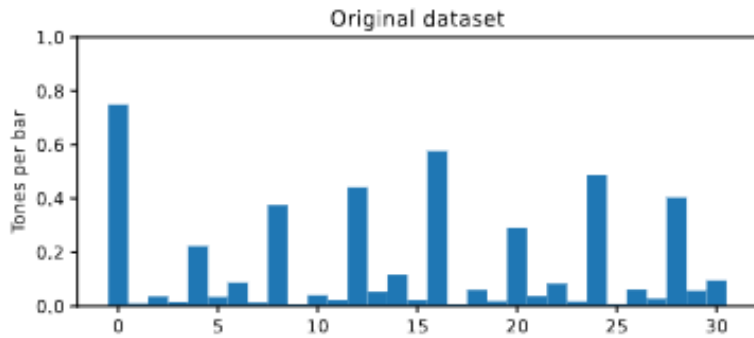
- 300 attempted generated samples, which resulted in 299 piano and 294 bass samples
- Rhythm analysis
- Harmony analysis

# Quantitative results

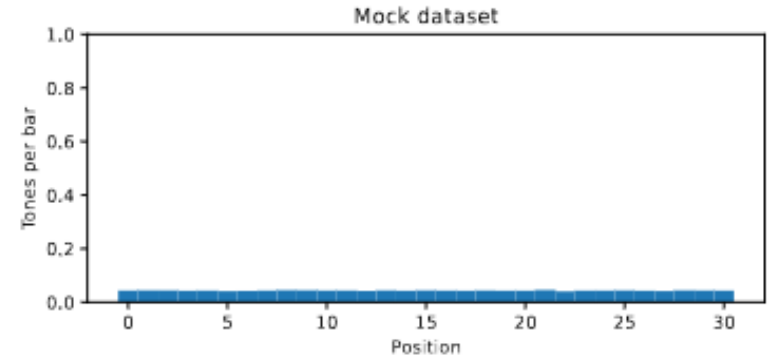
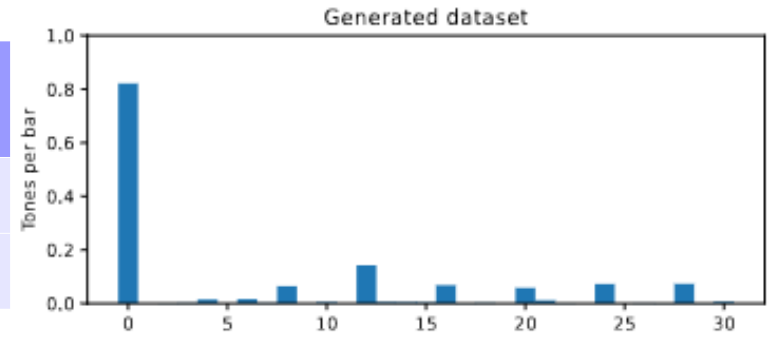
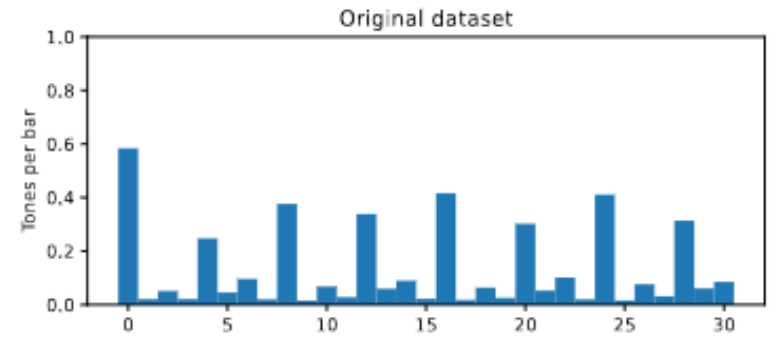
## Rhythm analysis

- Distributions of notes per bar
- Separate bars into discrete rhythm representation strings ("101...01")
- Sample Levenshtein distances with string pairs between generated sample vs. validation sample
- Compare with mock sample vs. validation sample

# Quantitative results



**Bass note distributions**



**Piano note distributions**

	Bass val.	Piano val.
Gen	4.104	4.103
Mock	4.973	4.800

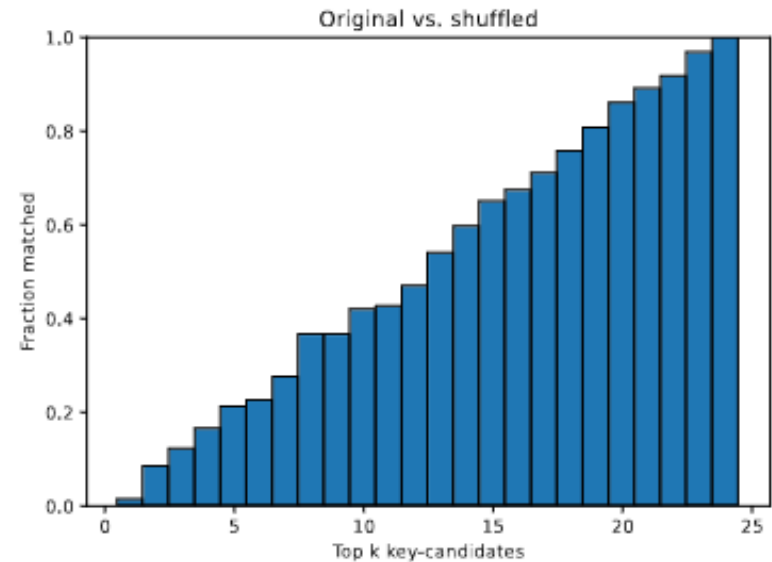
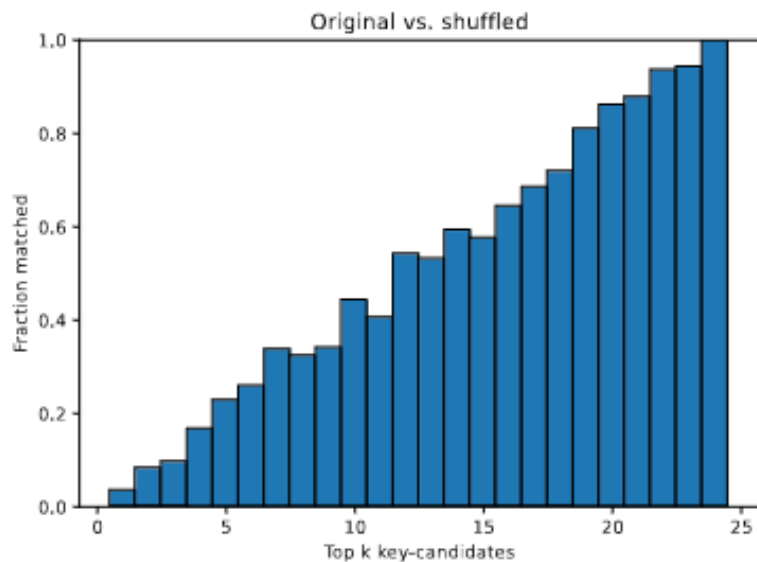
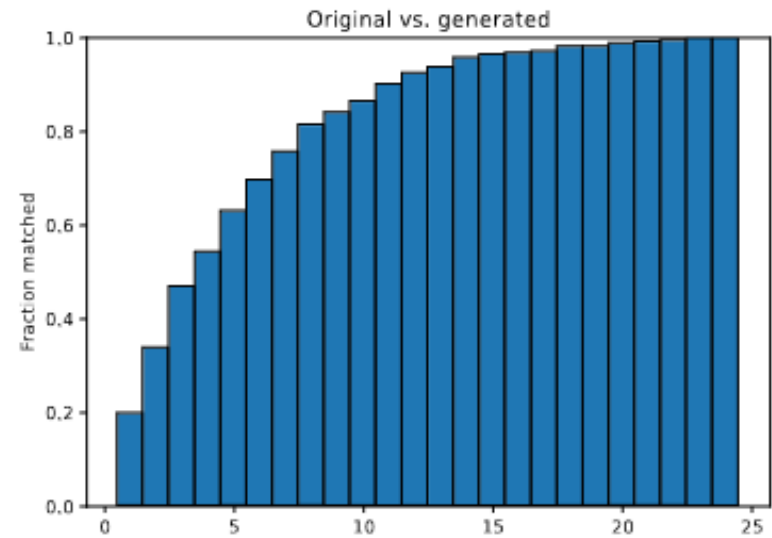
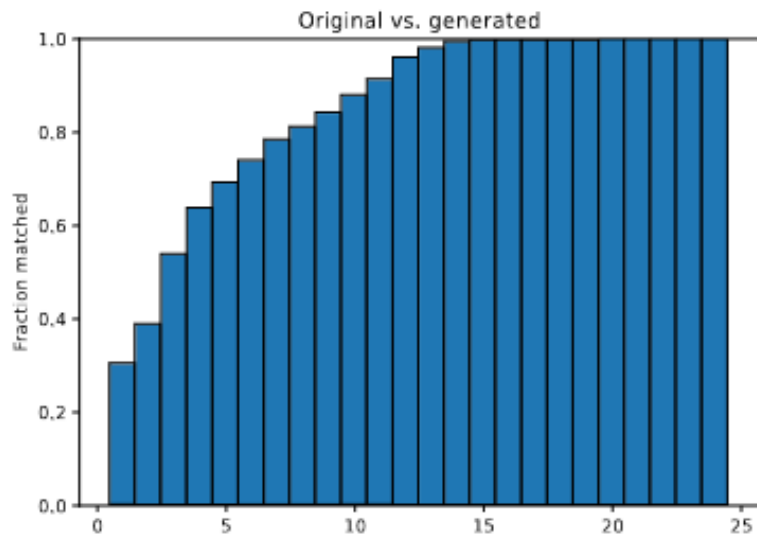
**Avg. Levenshtein distances**

# Quantitative results

## Harmony analysis

- Determine what *key* the samples are in with the Krumhansl-Schmuckler key-finding algorithm
- Compare the top-k key candidates given by the algorithm with the key of the reference

# Quantitative results



**Bass key comparisons**

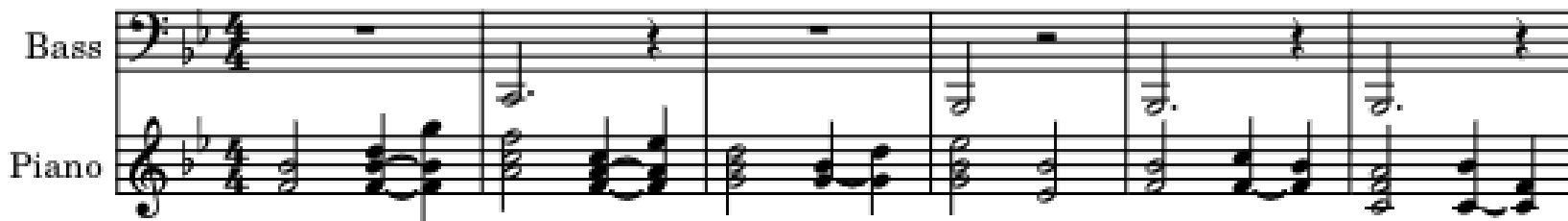
**Piano key comparisons**



# Qualitative results

Bass

Piano

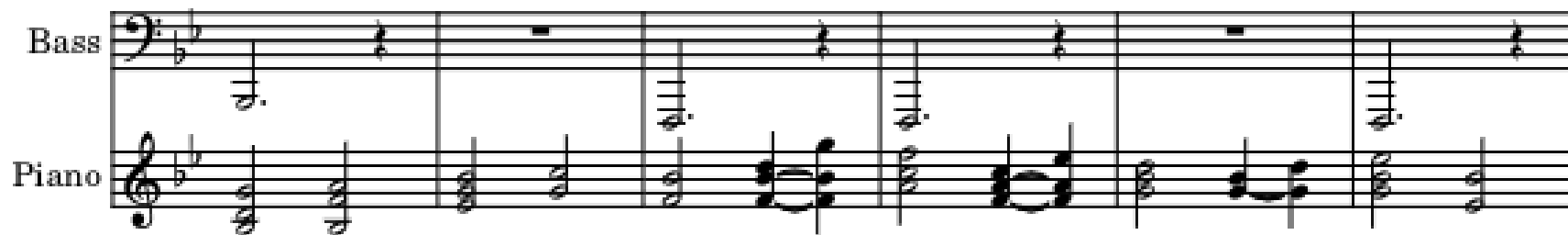


7

This system shows the first six measures of a musical piece. The Bass line is mostly silent, with a few notes in the final measures. The Piano line features a sequence of chords and moving lines, starting with a G major chord and moving through various intervals and chord structures.

Bass

Piano

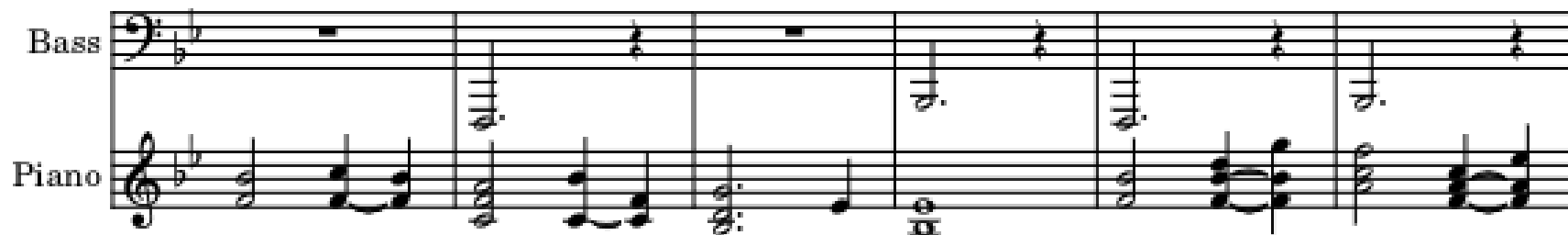


13

This system covers measures 7 through 12. The Bass line becomes more active, playing a steady eighth-note pattern. The Piano line continues with complex chordal textures and moving lines, including some triplets and sixteenth-note patterns.

Bass

Piano

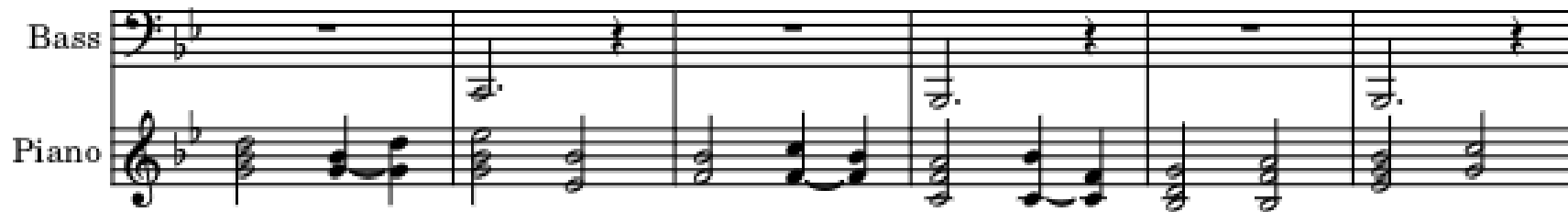


19

This system shows measures 13 through 18. The Bass line maintains its rhythmic pattern. The Piano line features a mix of chords and moving lines, with some instances of sixteenth-note runs and complex chordal structures.

Bass

Piano



This system shows the final six measures of the piece. The Bass line continues with its rhythmic pattern. The Piano line concludes with a series of chords and moving lines, ending with a final chord.

# Discussion

- Model performance is unimpressive
- The model learns some larger-scale information about the reference piece but does not react to smaller-scale changes
- Possible reasons:
  - Small/bad-quality dataset
  - Training parameters
  - Tokenization/model parameters
  - Inference method
- Model could potentially perform better if these reasons are addressed