Generating music with Large Language Models

Filip Ręka[1]

[1]AGH University of Krakow Faculty of Computer Science

Introduction

Text and music share a fundamental structural similarity, relying on patterns and sequences to convey and evoke emotion. Both utilize a hierarchical organization: sentences form paragraphs, while notes and chords create melodies and harmonies. Just as text is structured according to the rules of syntax and grammar, music follows principles of rhythm and harmony. Music, in addition to its temporal dimension (horizontal), also has a vertical aspect—the possibility of several notes being played simultaneously as a chord. This is an element that does not occur in written text in the Latin alphabet, making it important that the processing of musical notation considers this phenomenon.

Music representation

In Western music notation, music is traditionally written on a staff (also known as a pentagram), which consists of five horizontal lines that represent different pitches. Unfortunately, this method of storing music is not very conducive to processing and analyzing it using algorithms.

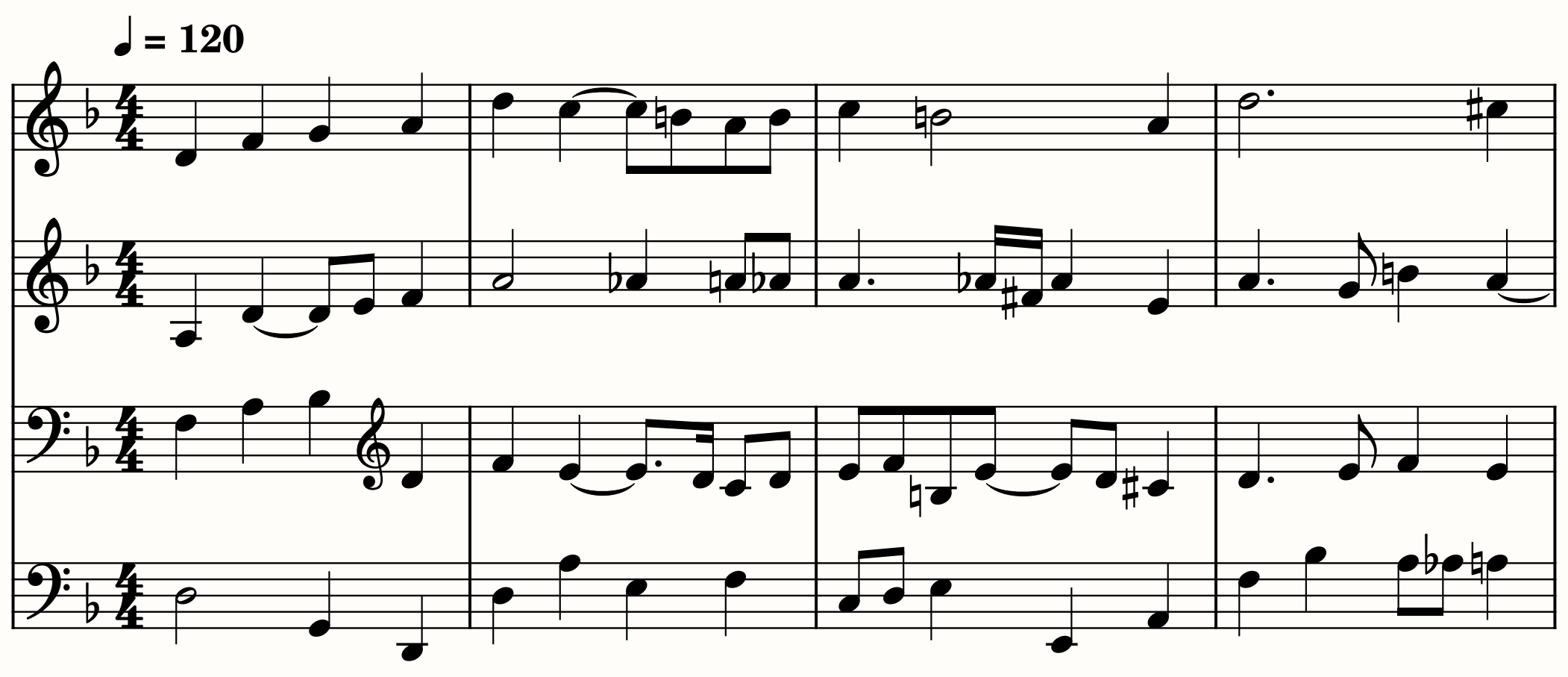

Figure 1: Music in traditional western notation.

Over the years, many file formats have been created to help computer scientists work with music more easily. The most notable ones are MusicXML, WAV and MP3 (which stores recordings rather than notation), MIDI, and ABC. The last two were the most important during my research. The MIDI file format represents music as a series of events consisting of fields like pitch, duration, and loudness. A typical way of viewing such a file is on a piano roll, where the X-axis represents time and the Y-axis represents pitch.

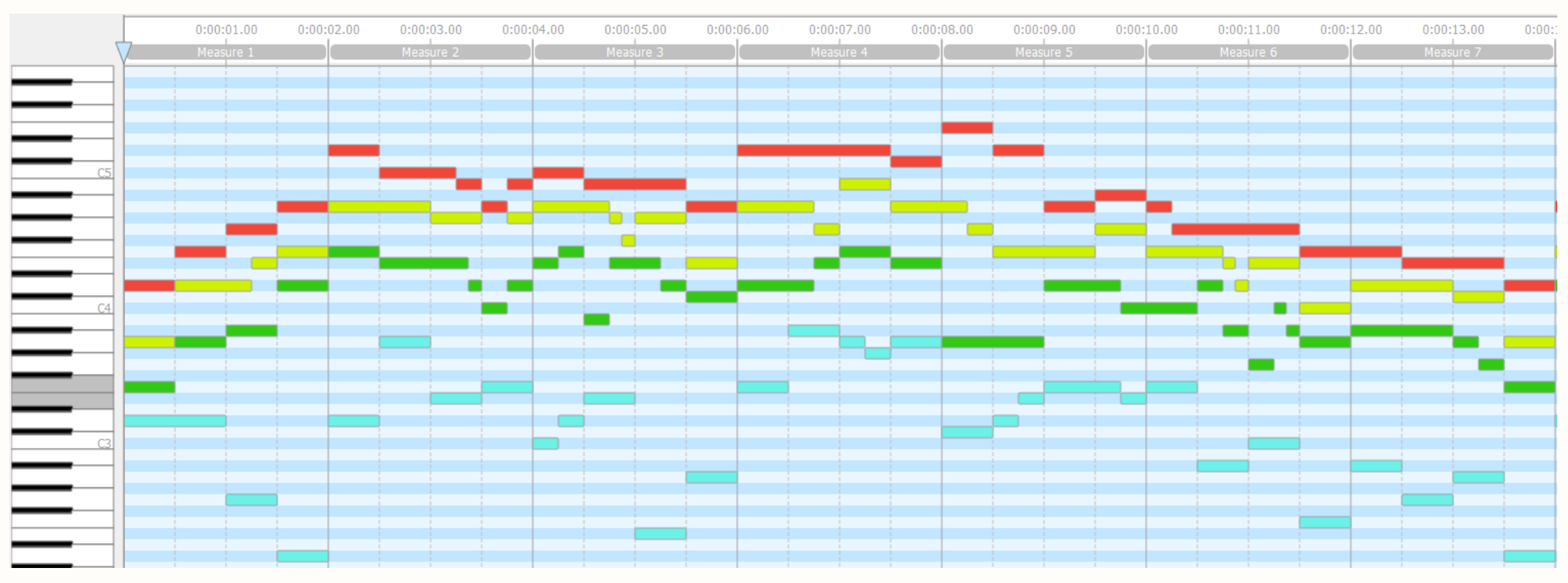

Figure 2: Music in MIDI notation.

The ABC format was developed as a way to represent traditional Irish music. It later evolved, largely with the help of Chris Walshaw, who initially used it before fully understanding traditional Western notation. ABC notation uses only ASCII characters to represent any given music track. The basic notation allows the user to specify key, meter, and note length with the tags K, M, and L, respectively. With the help of additional tags, it can also represent multiple voices together. An extreme example of this is the notation of the Second Movement of Ludwig van Beethoven's 7th Symphony, which consists of 19 different voices.

```
X:1
Q:1/4=120
V:1
L:1/16
M:4/4
K:C clef=G2
D4F4G4A4|d4c6B2A2B2|c4B8A4|d12^c4|
```

Figure 3: Music in ABC notation.

In the context of language models, working with ABC data does not require any additional preprocessing compared to traditional text. However, to work with MIDI data, we first need to tokenize it. There are many tokenizers available, but in this example, ReMI (Revamped MIDI) was used [1]. ReMI represents notes with Pitch, Velocity, and Duration tokens, and it represents time with Bar and Position tokens, which indicate when a new bar begins and the specific position within the bar.

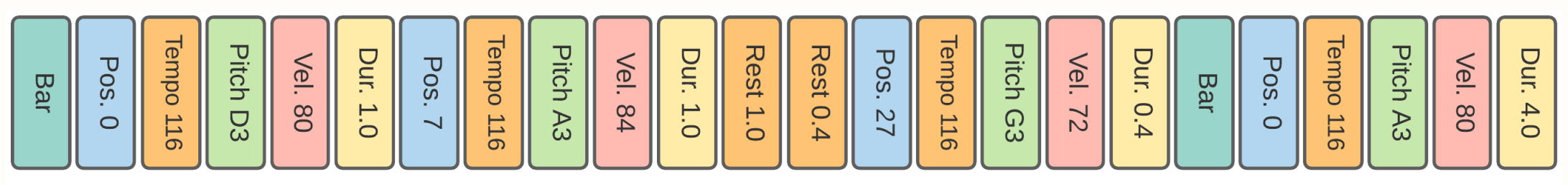

Figure 4: Sequence of REMI tokens.

Transformer and Mamba models

Music, like text, is generated using a process called Causal Language Modeling (CLM), which predicts the next note or word in a sequence based on the preceding ones, ensuring that the generated output is coherent and follows the logical flow of the prior content. Currently, the most acclaimed models use transformer architecture, which employs an attention mechanism to track dependencies between each token in the context window. Although the transformer consists of two parts—the encoder and the decoder—the encoder is often omitted in CLM tasks because it usually provides additional context to the decoder that is not necessary for CLM. When working with decoder-only models, a special mask is used to ensure that tokens cannot "look ahead" and thus avoid cheating while predicting the sequence. One of the most problematic aspects of transformer-based models is their complexity during inference. Generating a sequence of length L requires L^2 computations, which can be costly for long sequences. This challenge has led to the development of new model types, one of which is the *State Space Model* (SSM). With the addition of the HiPPO matrix, selective scan algorithm, and convolutional representation, the new model called Mamba has been introduced [2]. In language modeling tasks, it surpasses transformer models by providing reduced complexity, thereby improving training and inference speed.

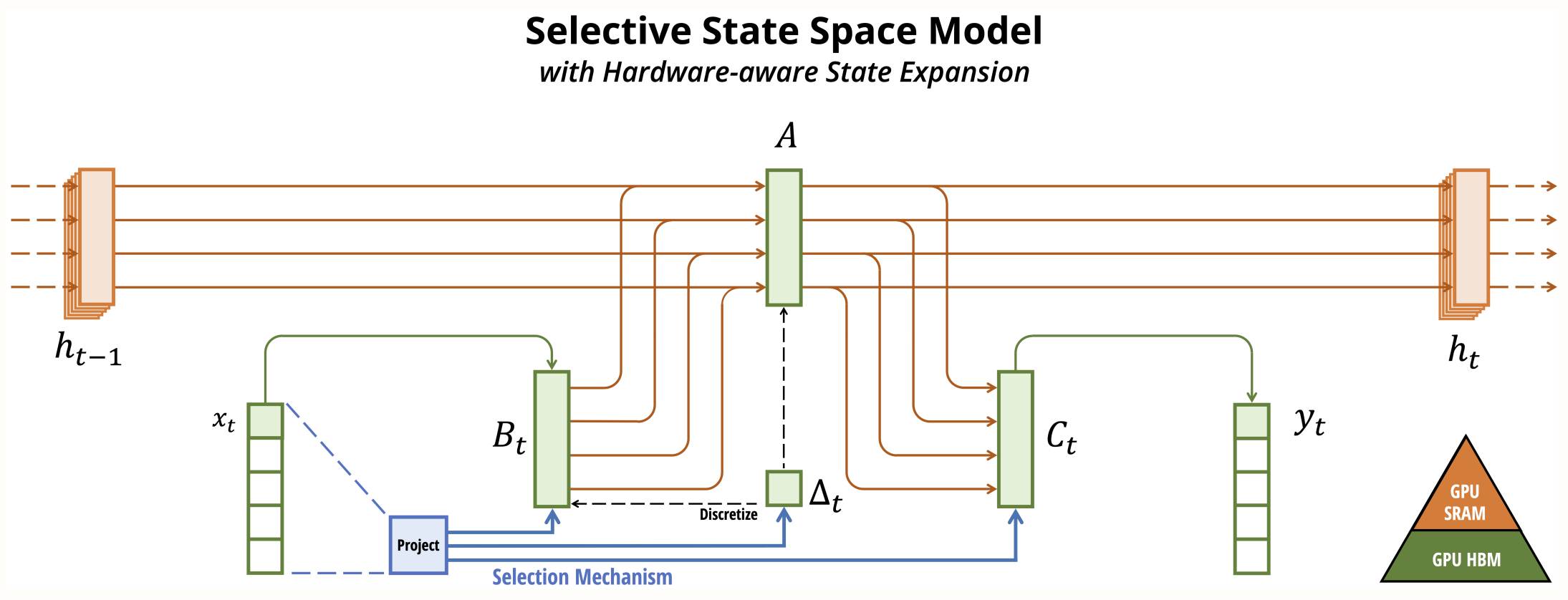

Figure 5: Scheme of Mamba model.

Experiments

To compare both architectures in the context of music generation, numerous experiments have been conducted. GPT-2 was chosen as the transformer-based model. Two variants of each model were trained: one with 6 million parameters and another with 60 million parameters. Although both are on the smaller side, larger models tend to overfit given the training data.

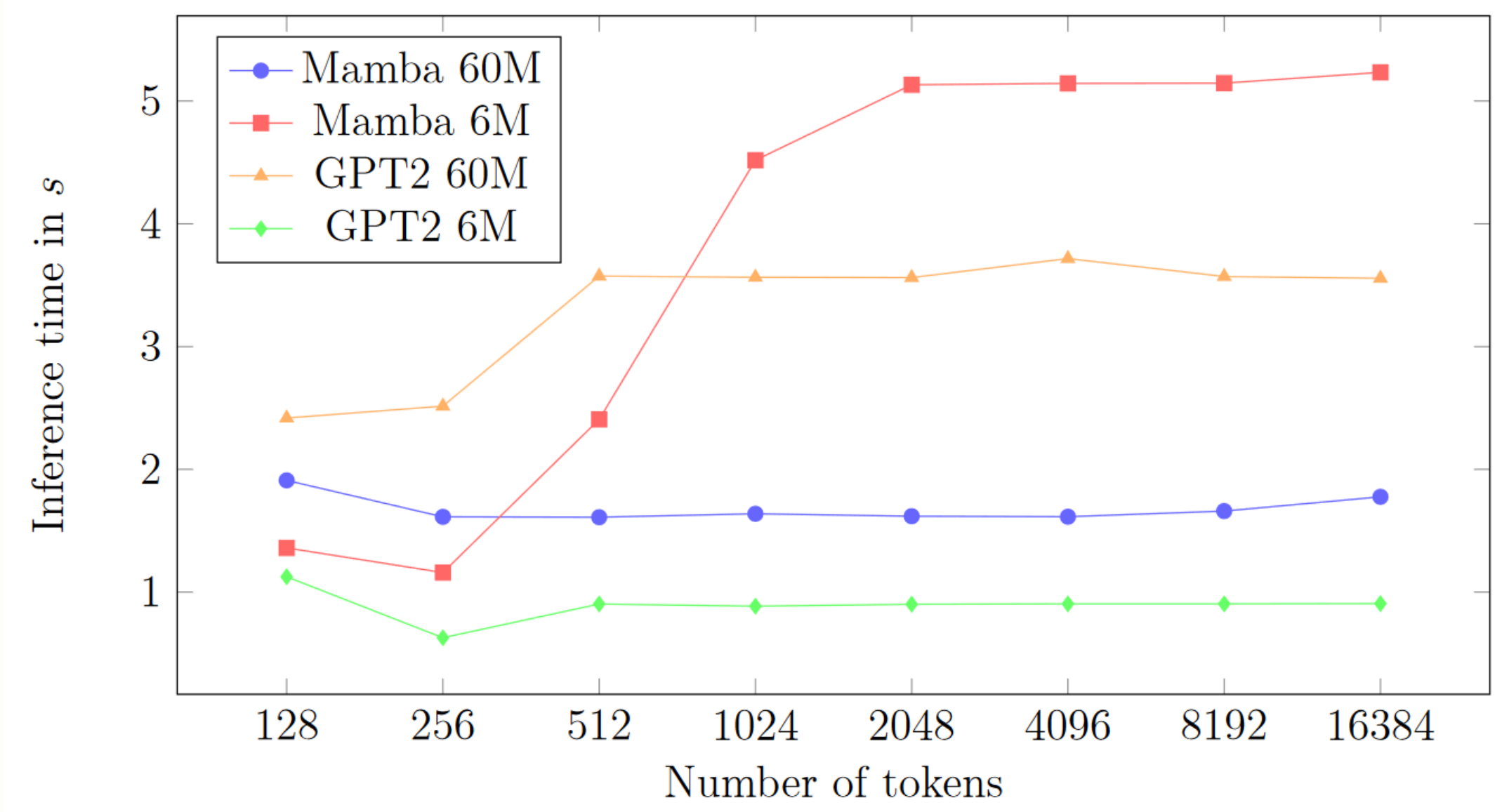

Figure 6: Inference time of different models on NVIDIA A100 GPU.

Based on Figure 6, the Mamba 60M model was twice as fast during inference compared to GPT-2 with the same number of parameters. The Mamba 6M model exhibited an unexpected spike after 256 tokens, which might be due to its architecture's inability to effectively utilize its features on a smaller scale.

Using LLMs for scoring generated music

One problem encountered during the research was the lack of a clear and accurate benchmark for scoring generated music. While music is a highly subjective medium, it still needs to adhere to certain rules. One approach to scoring generated music was to use a specially trained language model. An example of such a model is *ChatMusician* [3], which is a fine-tuned *Llama-2-7B* model trained on a dataset of music theory texts and music in ABC format. However, the model proved to be only somewhat useful, as it often produced very general information and did not identify harmonically incorrect sequences. In comparison, tests with the general GPT-4o model, despite not being specifically trained on music, turned out to be much more effective. GPT-4o analyzed each part of the melody using proper musical terminology and identified errors in harmony. Although it remained somewhat overly positive in its analysis, it opened up possibilities for further research into the musical reasoning capabilities of language models.

Listen to it yourself!

Music is a medium that is best experienced while listening. Scanning the QR will lead you to YouTube playlist with couple of generated fragments so you can decide for yourself if music generated by models sounds like music that is written by humans.

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid HPC Center: ACK Cyfronet AGH for providing computer facilities and support.

References

[1] Yu-Siang Huang and Yi-Hsuan Yang.
Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions.
In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1180–1188, New York, NY, USA, 2020. Association for Computing Machinery.

[2] Albert Gu and Tri Dao.
Mamba: Linear-time sequence modeling with selective state spaces, 2024.

[3] Ruibin Yuan.
Chatmusician: Understanding and generating music intrinsically with llm, 2024.