



# Hệ thống AI hỗ trợ sáng tác nhạc

(AI-powered Support System for Music Composition)

**Giảng viên hướng dẫn:** TS. Trần Duy Hoàng  
**Giảng viên phản biện:** TS. Lê Khánh Duy

**Số thứ tự:** 7 – **Loại đề tài:** Nghiên cứu  
**20120406** Phạm Quốc Vương  
**20120486** Ngô Phi Hùng

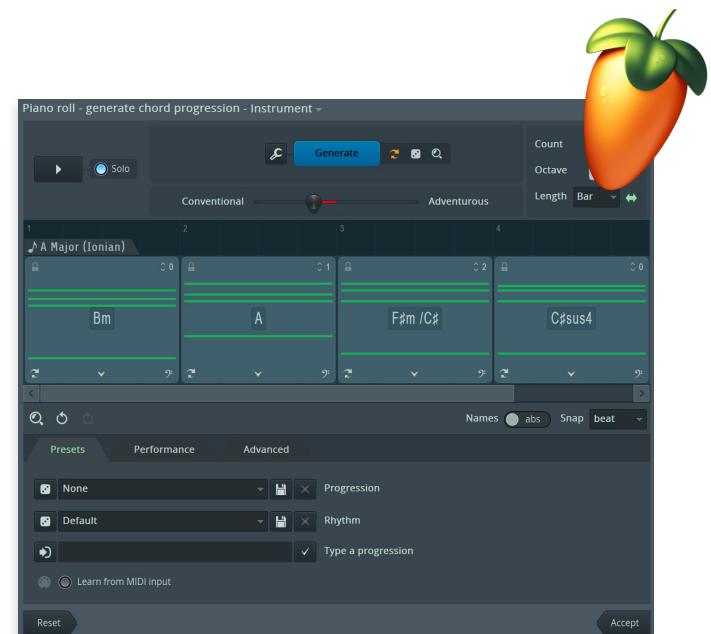
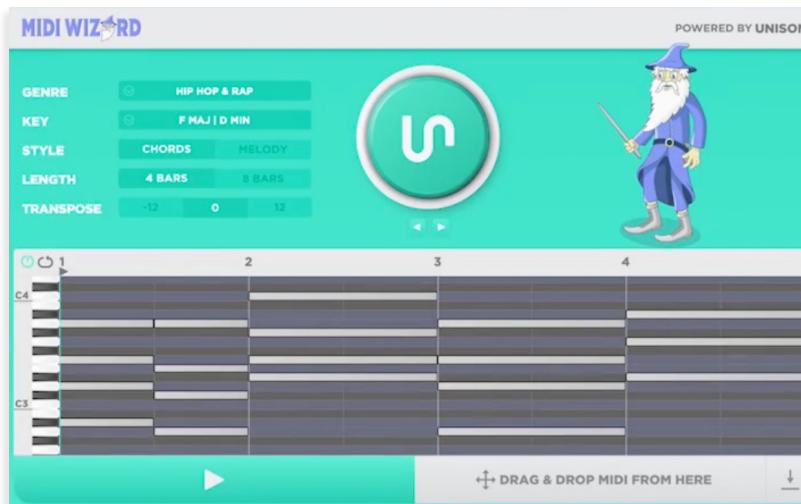
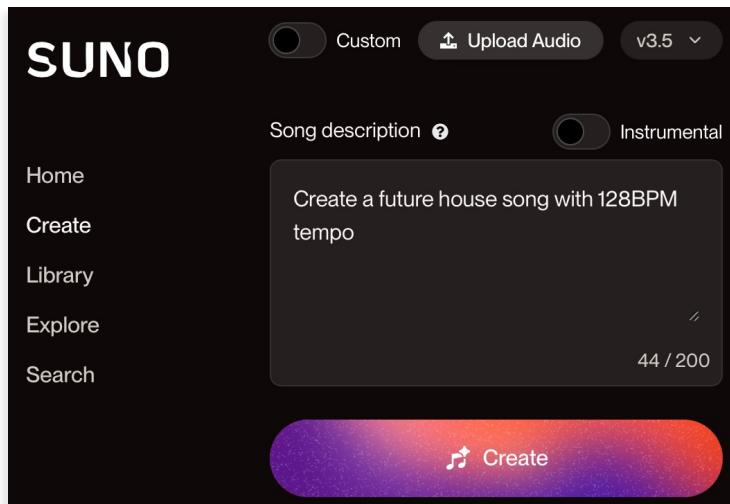
# Nội dung

1. Giới thiệu đề tài
2. Các công trình liên quan
3. Phương pháp đề xuất
4. Kết quả thí nghiệm
5. Kết luận

# 1. Giới thiệu đề tài

## 1.1. Đặt vấn đề

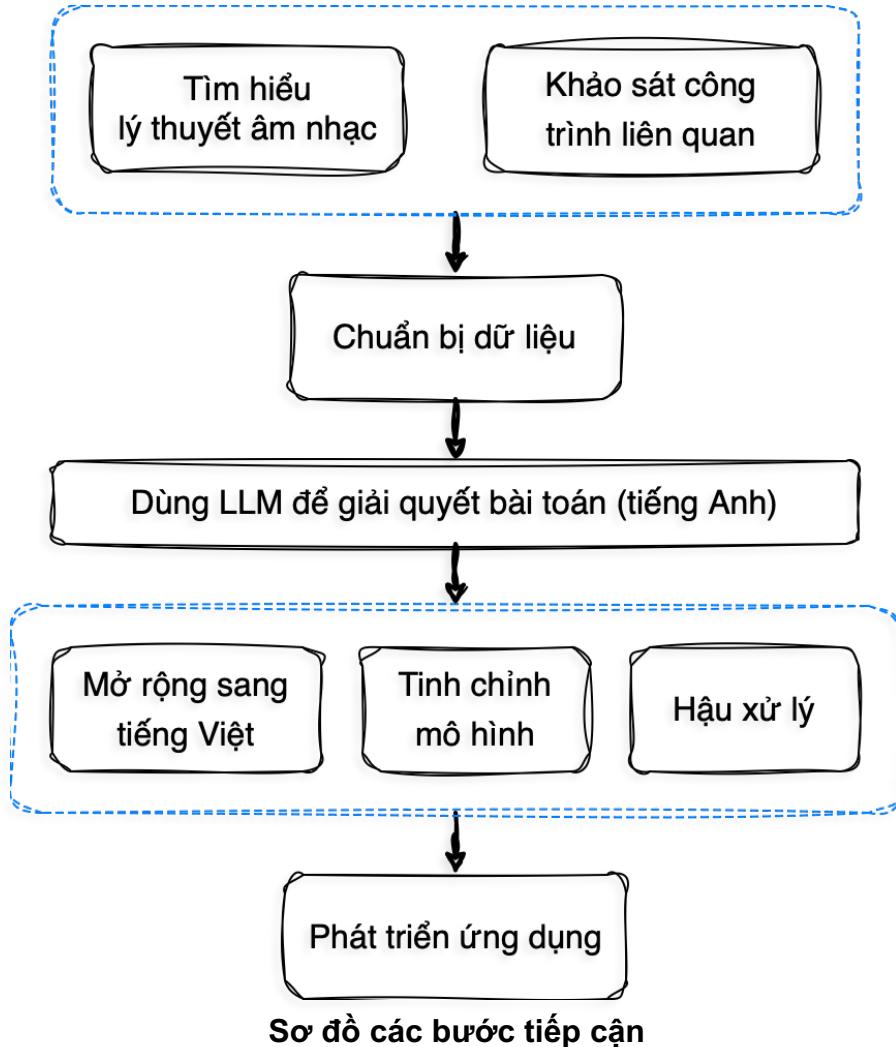
- Bài toán sinh nhạc: text-to-music (text-to-audio và **text-to-midi**)
- Hiểu text mang cảm tính (vấn đề của một số nghiên cứu)
- Sinh nhạc không đúng với mô tả
- MIDI: Dễ sửa, data nhẹ, train nhẹ, sinh nhạc



Một số giải pháp phần mềm cho bài toán sinh nhạc

# 1. Giới thiệu đề tài

## 1.3. Các bước tiếp cận



## 1.4. Đóng góp

- Mô hình sinh nhạc với đầu vào là mô tả tiếng Anh hoặc tiếng Việt.
- Phần mềm sinh nhạc theo yêu cầu.
- Giải pháp hậu xử lý dữ liệu âm nhạc (dạng MIDI).
- Kỹ thuật cào dữ liệu.
- Bài báo khoa học nộp đến **The 2<sup>nd</sup> International Conference on Intelligent Systems and Data Science (ISDS 2024)** (*xem thông tin hội nghị ở trang sau*).

## Call for papers

Follow success of [the first ISDS 2023](#) organized at Can Tho University, the second ISDS 2024 will be organized at Nha Trang University. Objectives of this international conference is to attract domestic and foreign researchers to participate and present outstanding and recent research in the field of ICT. This is an opportunity for scientists to meet, exchange, and cooperate. The ISDS 2024 is also a place for students to report and learn new results in the field of ICT. This ISDS conference looks at state-of-the-art and original research issues (in the topics of intelligent systems and data science).

Topics of the conference relate to (but not limited to):

- Track 1: Intelligent Systems & Recommender Systems
- Track 2: Data Science & Machine Learing
- Track 3: Image Processing & Pattern Recognition
- Track 4: Natural Language Processing

## Important dates

- Deadline for submission: 31-07-2024
- Acceptance notification: 26-08-2024
- Deadline for final papers: 06-09-2024
- Conference dates: 09-11-2024 – 10-11-2024

## Submission guidelines

All papers must be original and not simultaneously submitted to another journal or conference. Authors are invited to electronically submit full papers in English. The submitted papers must be in PDF in the [LNCS/CCIS one-column page format](#). The length of submitted papers should be from 12-15 pages (for long papers) and 6-8 pages (for short papers). All papers have to be written in the English language.

Authors are invited to submit their papers at the EasyChair web site using the following URL: <https://easychair.org/conferences/?conf=isds2024>

## Publications

All accepted papers will be published in one of the following methods (based on the review results):

- **Proceedings:** Papers with **acceptance rate less than 40%** will be published by [Springer Verlag in Communications in Computer and Information Science \(CCIS - indexed in Scopus\)](#).

Registration fee for each paper/author in Proceedings:

- + For Vietnamese authors: 4.000.000 VND
- + For Foreigner authors: 300 USD

- **Special issue in journal:** Papers with **acceptance rate from 40% to 60%** will be published by [CTU Journal of Innovation and Sustainable Development \(CTUJoISD\)](#). These papers will be converted to the CTUJoISD template by the authors.

Registration fee for each paper/author in the CTUJoISD:

- + For Vietnamese authors: 1.500.000 VND
- + For Foreigner authors: 150 USD

Moreover, the selected papers, after further revision and extension (at least 30%), will be considered for publication in special issues of the [Springer Nature Computer Science \(SNCS\) journal](#). SNCS is a broad-based, peer reviewed journal that publishes original research in all the disciplines of computer science including various interdisciplinary aspects. SNCS is indexed and abstracted in Scopus, ACM Digital Library, DBLP, Google Scholar, etc.



## 2. Công trình liên quan

### 2.1. Các nghiên cứu

- **MuseCoCo**<sup>[1]</sup>: text-to-midi, văn bản mô tả mang tính kỹ thuật và cảm tính.
- **MusicLM**<sup>[2]</sup>: text-to-audio, văn bản mô tả mang tính kỹ thuật lẫn ngũ cành.
- **MUGEN**<sup>[3]</sup>: text-and-video-to-audio, văn bản mô tả ngũ cành và đoạn video về game.

### 2.2. Dữ liệu

- **MuseCoCo**:
  - Các cặp “text – encoded text”;
  - Các bản nhạc ở định dạng MuseCoCo.
- **ChatGPT**:
  - Hỗ trợ tạo các mẫu câu mô tả âm nhạc.
- **Hooktheory**: Dữ liệu âm nhạc có thể chuyển về dạng MIDI.

## 2. Công trình liên quan

### 2.3. Cơ sở lý thuyết

- Nhạc lý
- Kỹ thuật huấn luyện:
  - Masked Language Modelling
  - Casual Language Modelling
  - LoRA
  - Fairseq

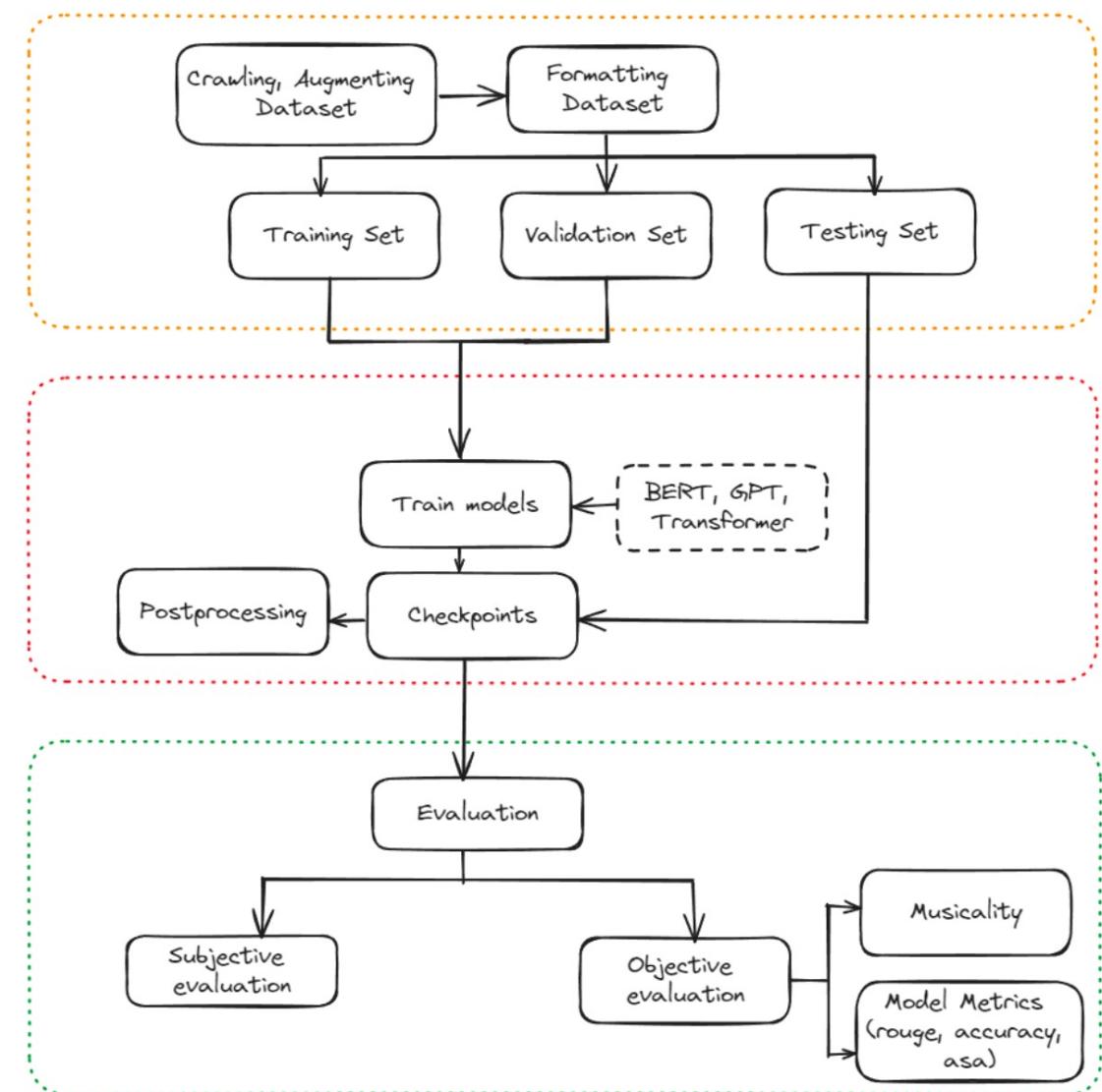
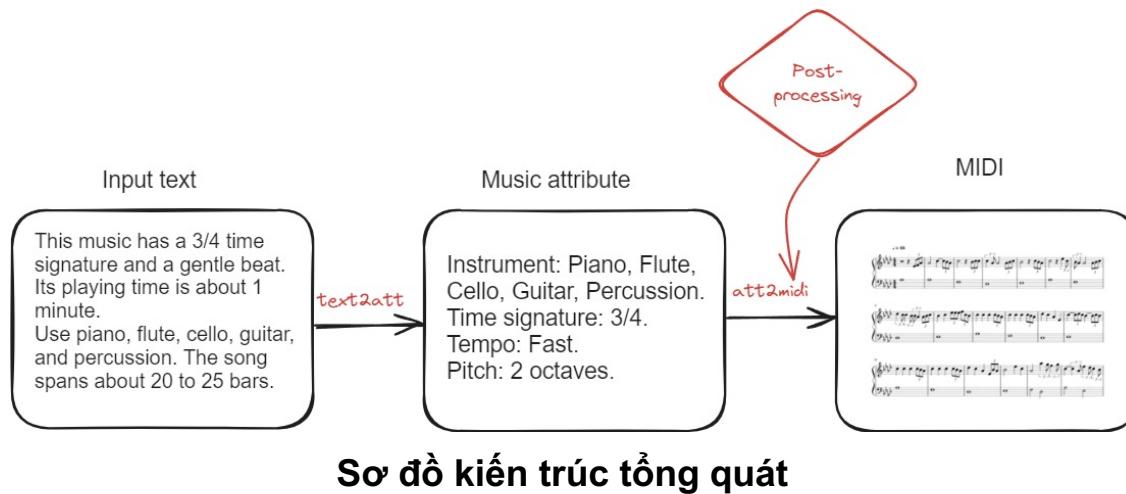
### 2.4. Mô hình ngôn ngữ lớn

- Transformer
- BERT
- GPT-2

# 3. Phương pháp đề xuất

## 3.1. Tổng quan về phương pháp

- Kiến trúc tổng quát:
  - ❑ Mô hình **text-to-attribute**;
  - ❑ Mô hình **attribute-to-midi**;
  - ❑ LoRA.
  - ❑ Mỗi mô hình mô hình áp dụng **quy trình thực nghiệm** như mô tả ở hình bên phải.



### 3. Phương pháp đề xuất

#### 3.2. Chuẩn bị dữ liệu

➤ Dữ liệu cho mô hình **text2att**:

**Kết quả chuẩn bị:**

- Các câu template mô tả âm nhạc
- Số mẫu dữ liệu tiếng Anh: 14,900  
(trong đó khoảng 4,800 mẫu từ MuseCoCo)
- Số mẫu dữ liệu tiếng Việt: 14,900

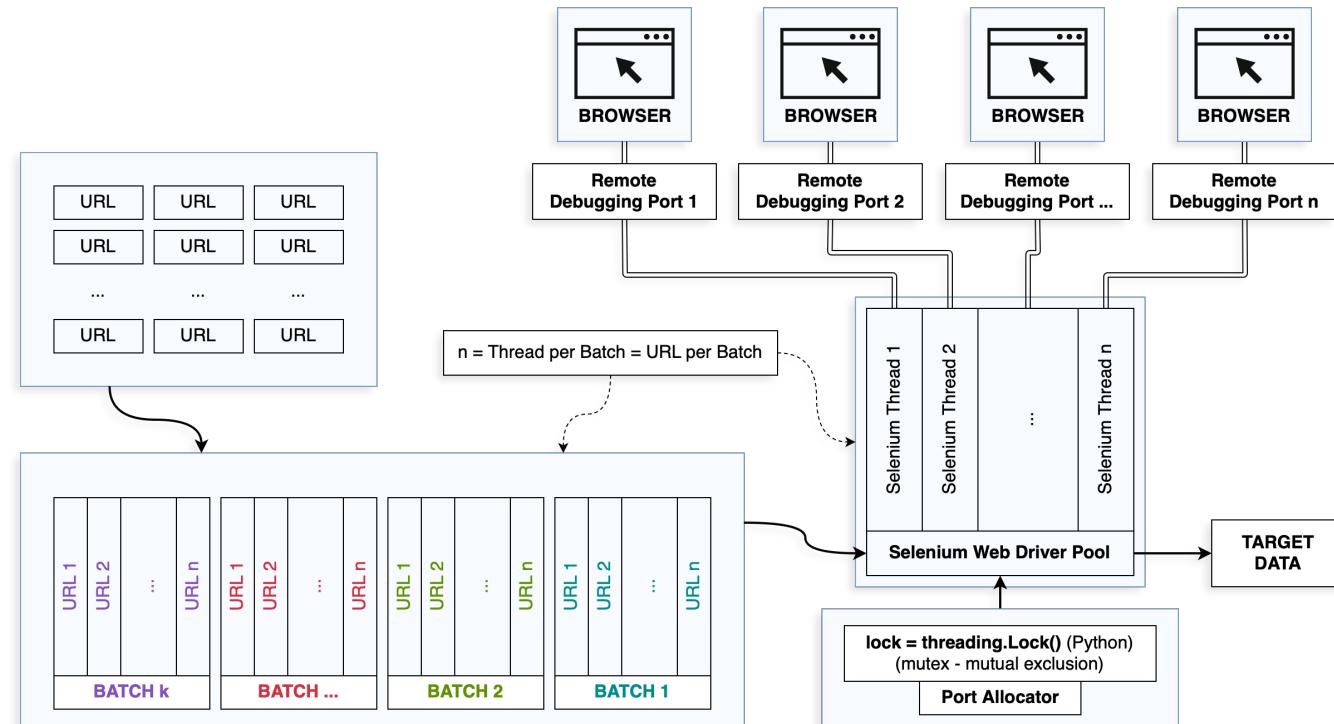
# 3. Phương pháp đề xuất

## 3.3. Chuẩn bị dữ liệu

- Dữ liệu cho mô hình **att2midi**:

- **Kết quả chuẩn bị:**

- 300 cặp “**command – music**” công khai của MuseCoCo.
    - 29,000 cặp thu thập từ Hooktheory và xử lý.

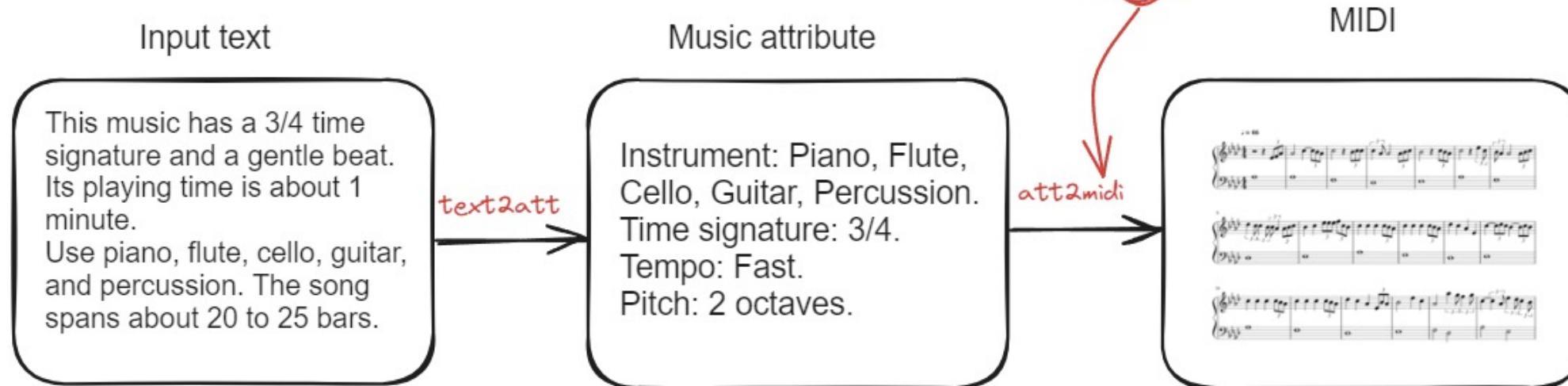


Kiến trúc cào dữ liệu áp dụng 3 kỹ thuật:

**Batch Processing, Multi-threading và Browser Automation**

# 3. Phương pháp đề xuất

## 3.4. Kiến trúc mô hình

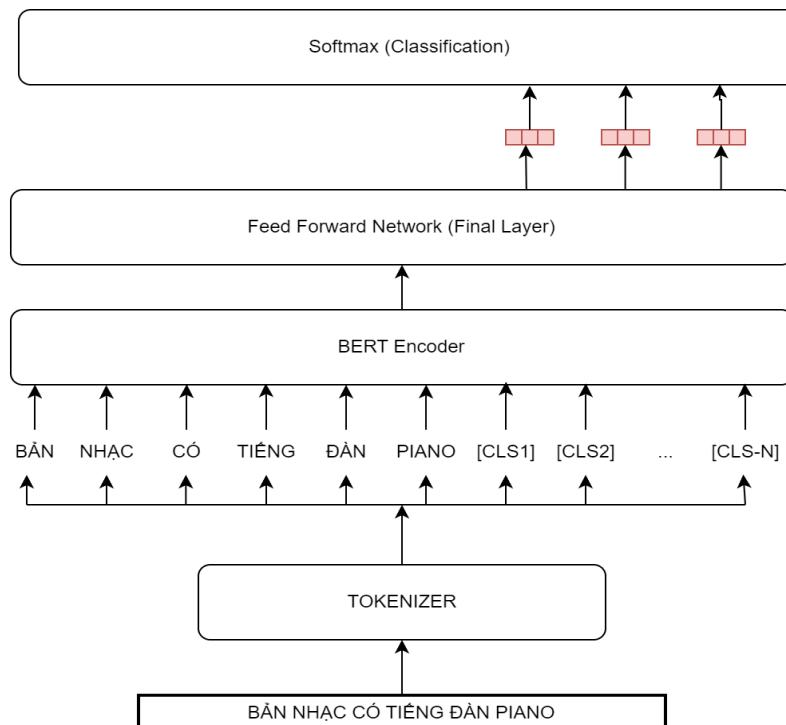


- Mô hình gồm 2 stage: text2att và att2midi.
- Lựa chọn kiến trúc trên phù hợp vì:
  - Khối lượng tài nguyên tính toán;
  - Khối lượng dữ liệu để huấn luyện;
  - Kiểm soát đầu ra;
  - Lượng từ vựng cần xử lý;
  - Tách biệt khâu xử lý văn bản và âm nhạc.

# 3. Phương pháp đề xuất

## 3.5. text2att

- Checkpoint gốc: bert-uncased-multilangage (168M params).
- Thêm nhiều [CLS] đại diện cho mỗi nhãn để học được bối cảnh trong câu.



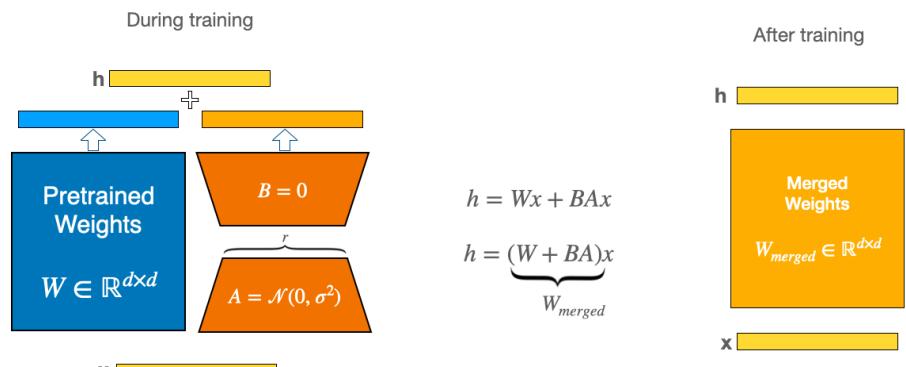
Mô tả flow text2att

Tên nhãn	Giá trị
Instrument	28 instruments: piano, keyboard, percussion, organ, guitar, bass, violin, viola, cello, harp, strings, voice, trumpet, trombone, tuba, horn, brass, sax, oboe, bassoon, clarinet, piccolo, flute, pipe, synthesizer, ethnic instrument, sound effect, drum. Mỗi nhạc cụ: 0: Được chơi, 1: Không được chơi, 2: NA
Pitch	Range: 0-11: octaves, 12: NA.
Rhythm	0: danceable, 1: not danceable, 2: NA.
Danceability	
Bar	0: 1-4 bars, 1: 5-8 bars, 2: 9-12 bars, 3: 13-16 bars, 4: NA.
Time	0: 4/4, 1: 2/4, 2: 3/4, 3: 1/4, 4: 6/8, 5: 3/8, 6: các nhịp khác, 7: NA.
Signature	
Key	0: major, 1: minor, 2: NA.
Tempo	0: chậm (<=76 BPM), 1: trung bình (76-120 BPM), 2: nhanh (>=120 BPM), 3: NA.
Time	0: 0-15s, 1: 15-30s, 2: 30-45s, 3: 45-60s, 4: >60s, 5: NA.

Tên và giá trị tương ứng của các nhãn trong mỗi câu template  
từ bài báo MuseCoCo (đã được rút gọn)

# 3. Phương pháp đề xuất

## 3.6. att2midi - Mô hình 1: GPT2 LoRA



- 203 Triệu tham số.
- Lựa chọn các tham số tuyến tính để huấn luyện.
- 3.5 Triệu tham số được huấn luyện.

### 3. Phương pháp đề xuất

#### 3.6. att2midi - Mô hình 1: GPT2 LoRA

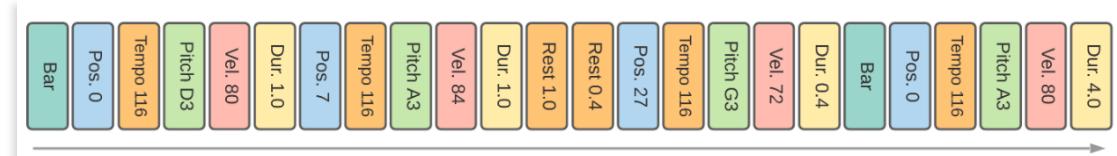
➤ Chuẩn bị bộ tokenizer cho mô hình sinh nhạc:

□ Bộ từ vựng 1253 từ bao gồm:

- Gồm từ vựng âm nhạc (REMI)
- Từ vựng command (metadata)
- Từ vựng khác: special token (CLS, UNK, SEP,...)

Ký hiệu	Mô tả
I1s2	Instrument
I4	Main Instrument
R3	Rhythm Intensity
B1s1	Bar
TS1s1	Time Signature
K1	Key
T1s1	Tempo
P4	Pitch Range
TM1	Time

Bảng 3.2: Mô tả các khoá trong command



**Minh họa phương pháp token hoá âm nhạc REMI**

(Nền tảng cho phương pháp của MuseCoCo)

# 3. Phương pháp đề xuất

## 3.6. att2midi - Mô hình 1: GPT2 LoRA



Mô tả dạng huấn luyện command-music, att2midi

```
I1s2 : (11, 4)
I4 : (11, False)
R3 : 1
B1s1 : (16, 3)
TS1s1 : (4, 4)
K1 : major
T1s1 : (114.03503592196344, 1)
P4 : 3
TM1 : (33.45463058047076, 2)
```

### Minh họa một command

```
train.txt x
Users > 4rr311 > Documents > VectorA > KHTN > Nam4 > HKII > Thesis > Brainstorming
1 s-9 o-0 t-33 i-57 p-50 d-12 v-20 p-50 d-12 v-20 o-12 t-33 i-57 p-5
2 s-9 o-0 t-31 i-35 p-43 d-9 v-25 i-78 p-79 d-6 v-25 i-128 p-197 d-3
3 s-9 o-0 t-27 i-0 p-74 d-3 v-12 p-53 d-27 v-12 i-32 p-38 d-12 v-4 i
4 s-9 o-0 t-39 i-128 p-173 d-6 v-20 o-6 t-39 i-56 p-72 d-6 v-20 i-57
5 s-9 o-0 t-35 i-0 p-36 d-6 v-20 o-6 t-35 i-0 p-43 d-6 v-20 o-12 t-3
6 s-9 o-0 t-30 i-0 p-59 d-6 v-20 p-36 d-18 v-16 i-96 p-55 d-6 v-16 i
7 s-9 o-0 t-35 i-14 p-62 d-12 v-20 i-48 p-53 d-6 v-20 p-50 d-6 v-20
8 s-9 o-0 t-38 i-12 p-78 d-3 v-20 p-68 d-3 v-20 p-56 d-3 v-20 p-54 d
9 s-23 o-0 t-35 i-0 p-59 d-14 v-24 p-47 d-4 v-28 p-40 d-4 v-28 o-6 t
10 s-9 o-0 t-33 i-30 p-40 d-27 v-25 p-28 d-27 v-25 i-35 p-40 d-25 v-2
11 s-9 o-0 t-38 i-0 p-74 d-11 v-20 p-65 d-11 v-20 p-50 d-6 v-20 o-6 t
12 s-9 o-0 t-33 i-56 p-79 d-6 v-20 i-57 p-60 d-6 v-20 i-58 p-45 d-6 v
13 s-9 o-0 t-23 i-56 p-57 d-6 v-20 i-64 p-72 d-9 v-20 o-6 t-23 i-56 p
14 s-17 o-0 t-35 i-40 p-64 d-3 v-28 i-41 p-52 d-6 v-31 i-42 p-52 d-6
```

### Ví dụ về dữ liệu âm nhạc

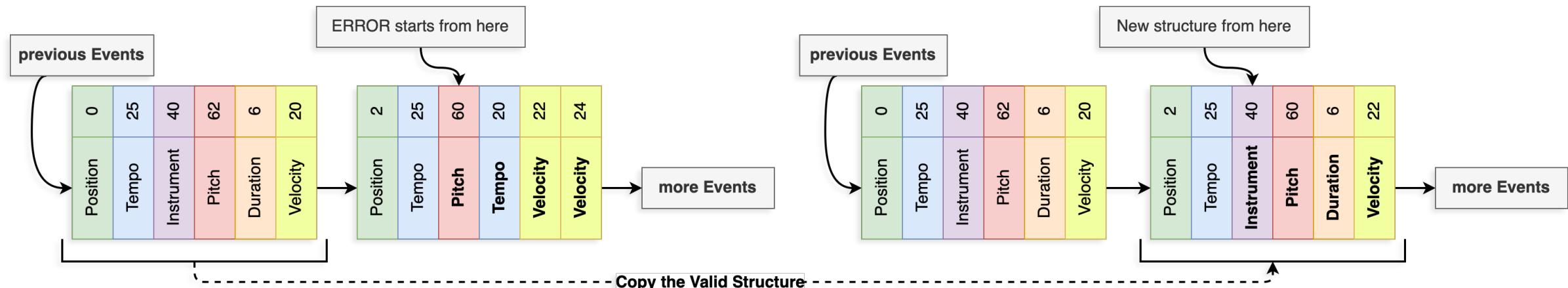
# 3. Phương pháp đề xuất

## 3.7. Hậu xử lý dữ liệu

- **Vấn đề:** Rủi ro dữ liệu sinh ra sai cấu trúc.
- **Giải pháp:**
  - ❑ Dữ liệu = time series = {Các sự kiện MIDI};
  - ❑ Tìm vị trí lỗi dựa trên đặc tả.
  - ❑ Khôi phục cấu trúc tại điểm lỗi;
  - ❑ Điền giá trị cho các thành phần cấu trúc.

Thuộc tính	Tên đầy đủ	Mô tả	Thuộc tính liền sau
s	Time Signature	Nhip của bản nhạc	b, o.
o	Position	Thời điểm xuất hiện của sự kiện	t.
t	Tempo	Tốc độ bản nhạc	i.
i	Instrument	Nhạc cụ	p.
p	Pitch	Cao độ nốt nhạc	d.
d	Duration	Độ dài nốt nhạc	v.
v	Velocity	Tốc độ nhấn/Lực nhấn phím đàn	i, b, p, o.
b	Bar	Vạch nhịp	s.

Bảng 2.2: Đặc tả dữ liệu âm nhạc dạng văn bản của MuseCoco



Ví dụ minh họa quá trình hậu xử lý

# 4. Kết quả thí nghiệm

## 4.1. Đánh giá mô hình

➤ **Khách quan:**

- Micro:** Tính accuracy cho từng nhãn riêng lẻ.
- Macro:** Tính accuracy trên toàn bộ các dự đoán.

$$\text{accuracy} = \frac{\text{số dự đoán đúng}}{\text{tổng số dự đoán}}$$

- ASA - Average Sample-wise Accuracy:** Do MuseCoCo đề xuất; Trung bình các giá trị  $A_i$  trên toàn bộ mẫu, với  $A_i$  = Tỷ lệ thuộc tính dự đoán đúng của mẫu i.
- Đánh giá dựa trên các đặc tính kỹ thuật.

➤ **Chủ quan:**

- Đánh giá dựa trên cảm nhận.

# 4. Kết quả thí nghiệm

## 4.2. Baseline cho att2midi – MuseCoCo (dựng lại)

- Dùng để so sánh với mô hình của nhóm.
- 200 Triệu tham số.
- Cùng một số hyperparameter với mô hình GPT2 - LoRA.

**facebookresearch/  
fairseq**

Facebook AI Research Sequence-to-Sequence  
Toolkit written in Python.

310  
Contributors

3k  
Used by

30k  
Stars

6k  
Forks



Công cụ fairseq do Facebook phát triển

Tham số	Giá trị	Mô tả
n_layer	20	Số lớp trong Transformer.
n_head	16	Số lượng head attention.
n_emb	1024	Kích thước vector embeddings.
FFN size	2048	Kích thước của Feed Forward Network.
dropout ratio	0.1	Tỷ lệ dropout để tránh overfitting.
optimizer	AdamW	Optimizer sử dụng để huấn luyện mô hình.
$\beta_1$	0.9	Tham số $\beta_1$ của Adam optimizer.
$\beta_2$	0.98	Tham số $\beta_2$ của Adam optimizer.
$\epsilon$	$10^{-9}$	Tham số $\epsilon$ của Adam optimizer.
learning rate	$2 \times 10^{-4}$	Tốc độ học của mô hình.
warmup steps	2000	Số bước đầu tiên để tốc độ học tăng dần đến giá trị tối đa đã định.

**Các tham số của mô hình  
Casual Linear Transformer dựa trên MuseCoCo**

# 4. Kết quả thí nghiệm

## 4.3. Kết quả huấn luyện

- LoRA GPT-2 **tốt hơn 9.84%**  
(ASA) so với kiến trúc MuseCoCo

**Table 4.** Accuracy of each instrument for the Instrument attribute

Instrument	Accuracy (ENG)	Accuracy (VIE)
accordion	0.94	0.93
brass	0.98	0.96
celesta	0.91	0.92
choir	0.95	0.97
guitar	0.99	0.93
harmonica	0.97	0.94
organ	0.90	0.91
piano	0.96	0.95
synth	0.92	0.94
viola	0.91	0.90
violin	0.93	0.92
voice	0.95	0.96

**Table 5.** Accuracy of other attributes

Attribute	Accuracy (ENG)	Accuracy (VIE)
Rhythm Danceability	0.85	0.87
Bar	0.91	0.89
Time Signature	0.94	0.92
Key	0.88	0.90
Tempo	0.93	0.85
Pitch Range	0.90	0.94

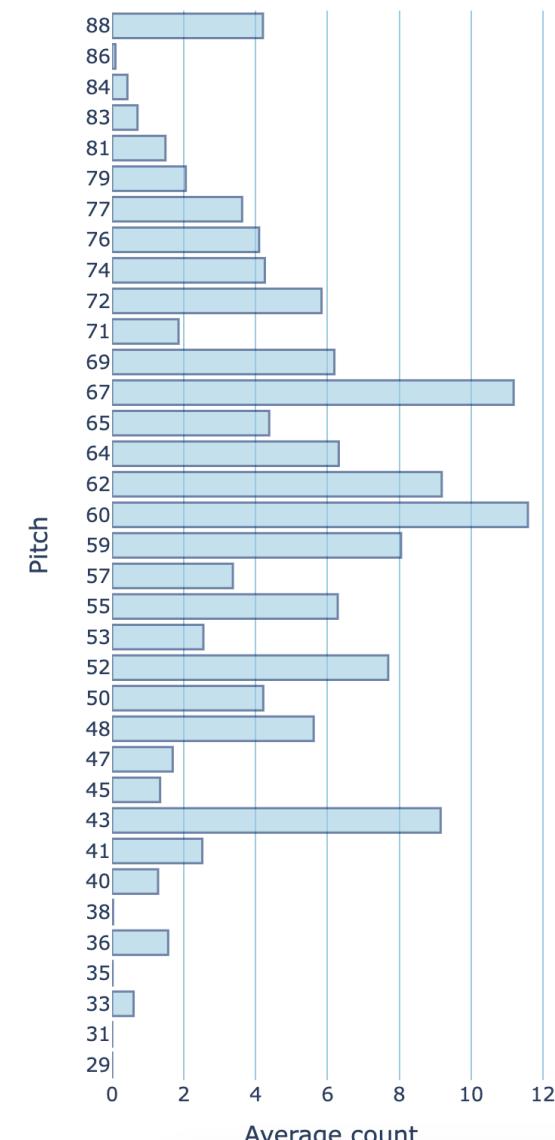
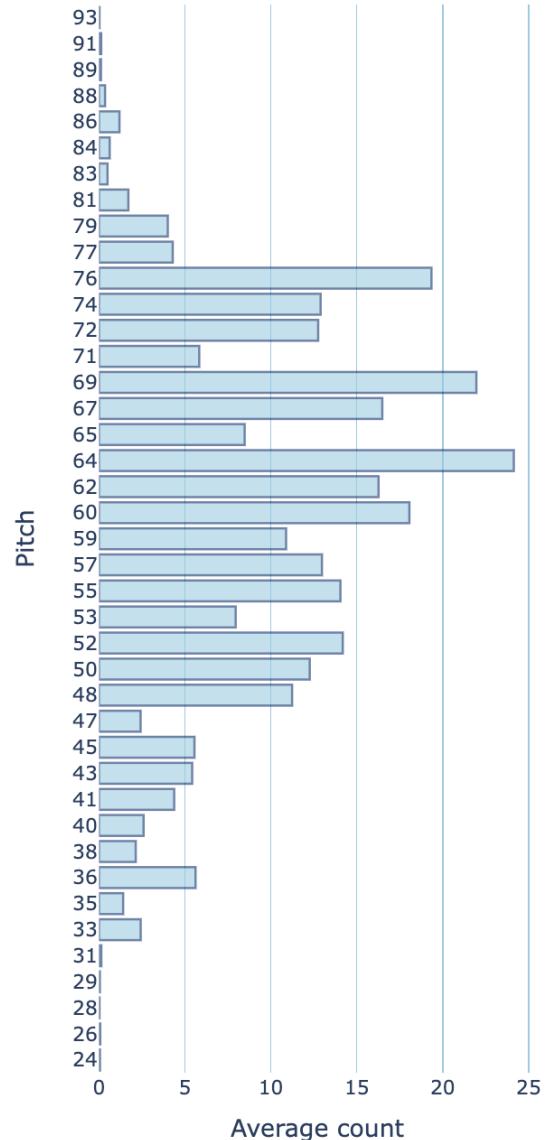
**Table 6.** Comparison of results between the two models

Metric	LoRA GPT-2	Reconstructed MuseCoCo
ASA	<b>0.67</b>	0.61
Instrument	0.68	0.60
Pitch Range	0.68	0.52
Rhythm Danceability	0.96	0.89
Bar	0.51	0.42
Time Signature	0.60	0.72
Key	0.57	0.51
Tempo	0.69	0.61

# 4. Kết quả thí nghiệm

## 4.3. Kết quả huấn luyện

- **Main Key:** C major/A minor
- **Min pitch:** 24 vs. 29
- **Max pitch:** 93 vs. 88
- **Đặc điểm:** melody ở high và high-mid; chord ở high-mid, mid, và low-mid; bass ở low-mid và low
- **Hình dáng phân phối:** mịn hơn vs. có outlier (43 và 88)

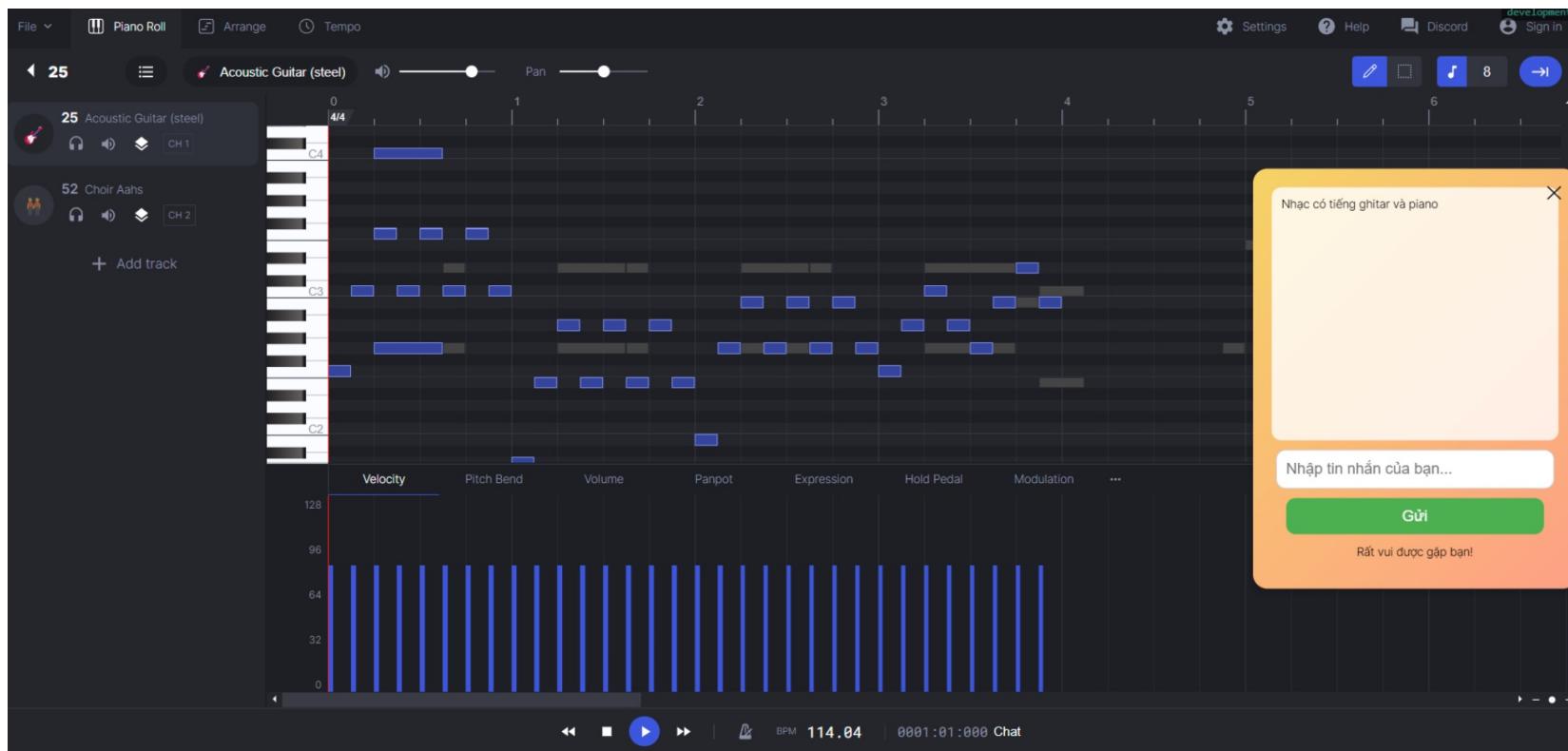


Pitch average count (LoRA GPT-2: trái, MuseCoCo: phải)

# 4. Kết quả thí nghiệm

## 4.4. Phần mềm demo

- Link: <https://youtu.be/vOBuDsyuA-s> (rút gọn: [bom.so/ai-music-demo](http://bom.so/ai-music-demo))

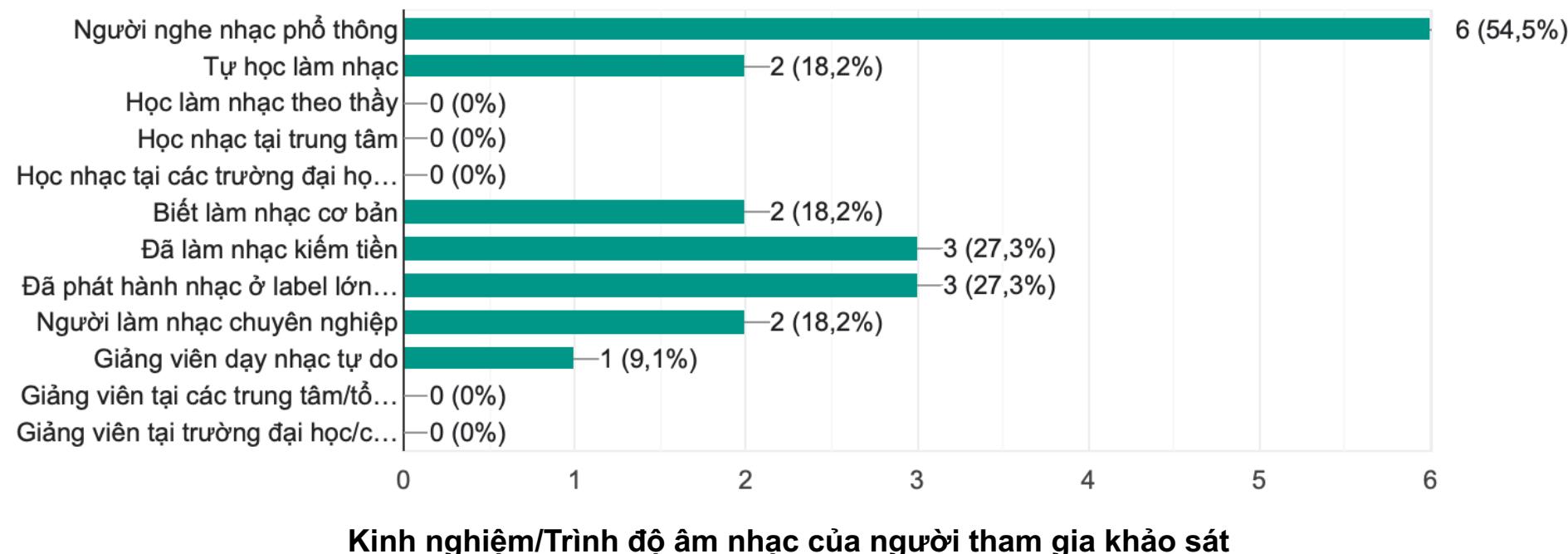


Giao diện phần mềm demo

# 4. Kết quả thí nghiệm

## 4.5. Nhận xét của cộng đồng

- Giao diện: 7.8/10
- Prompt dễ hiểu, dễ viết theo: 7/10
- Độ hay và giống prompt của nhạc: 6.5/10



## 5. Kết luận

- Kiến trúc hai lớp có hiệu quả đáng kể (9.84% hiệu năng).
- Dễ lựa chọn mô hình cho mỗi tác vụ.
- Làm giàu dữ liệu bằng ChatGPT (prompt engineering).
- Áp dụng LoRA tối ưu chi phí.
- Mang lại lợi ích cho người làm nhạc.

# TÀI LIỆU THAM KHẢO

1. Lu, Peiling et al. MuseCoco: Generating Symbolic Music from Text. 2023. arXiv: 2306.00110 [cs.SD].
2. Agostinelli, Andrea et al. MusicLM: Generating Music From Text. 2023. arXiv: 2301.11325 [cs.SD].
3. Hayes, Thomas et al. MUGEN: A Playground for Video-Audio-Text Multimodal Understanding and GENeration. 2022. arXiv: 2204.08058 [cs.CV].

**CẢM ƠN QUÝ THẦY CÔ ĐÃ LẮNG NGHE**