



# Scaling Open Foundation Models for Music

-- *MERT, ChatMusician, and YuE*

Ruixin Yuan

# About HKUST-Audio Group

## Machine Creativity Lab

Our lab is working on Open Audio AI, led by Prof. Wei Xue. Work include:

- Music
  - YuE
  - ChatMusician
  - ComposerX
- Speech
  - Xcodec
  - LLaSA
  - ComoSpeech
- General Audio
  - Audio-FLAN



香港生成式人工智能研發中心  
Hong Kong  
Generative  
AI | Research  
& Development  
Center



We are a group of researchers working towards AI4Art, especially AI Music, pre-training open foundation models, and beyond.

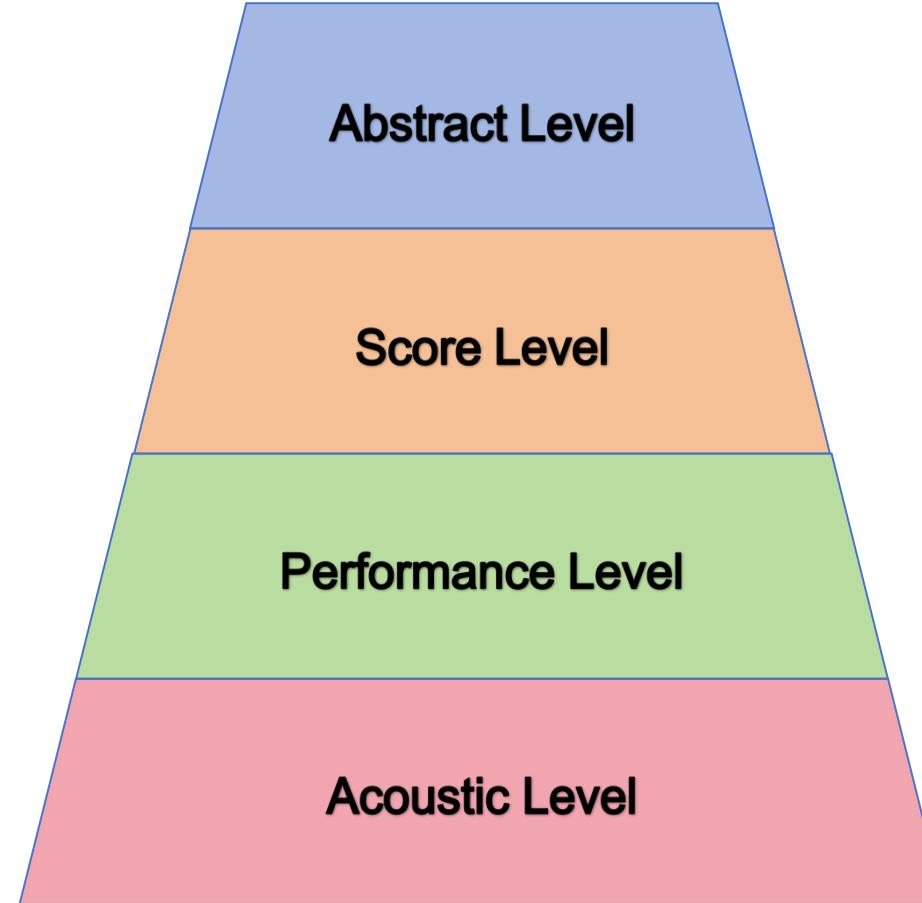
Our vision for Music x AGI:

- Open-source & Inclusive
- Empower Creativity: appreciate, understand, and create
- A Repository of Human Audio Arts and Beyond

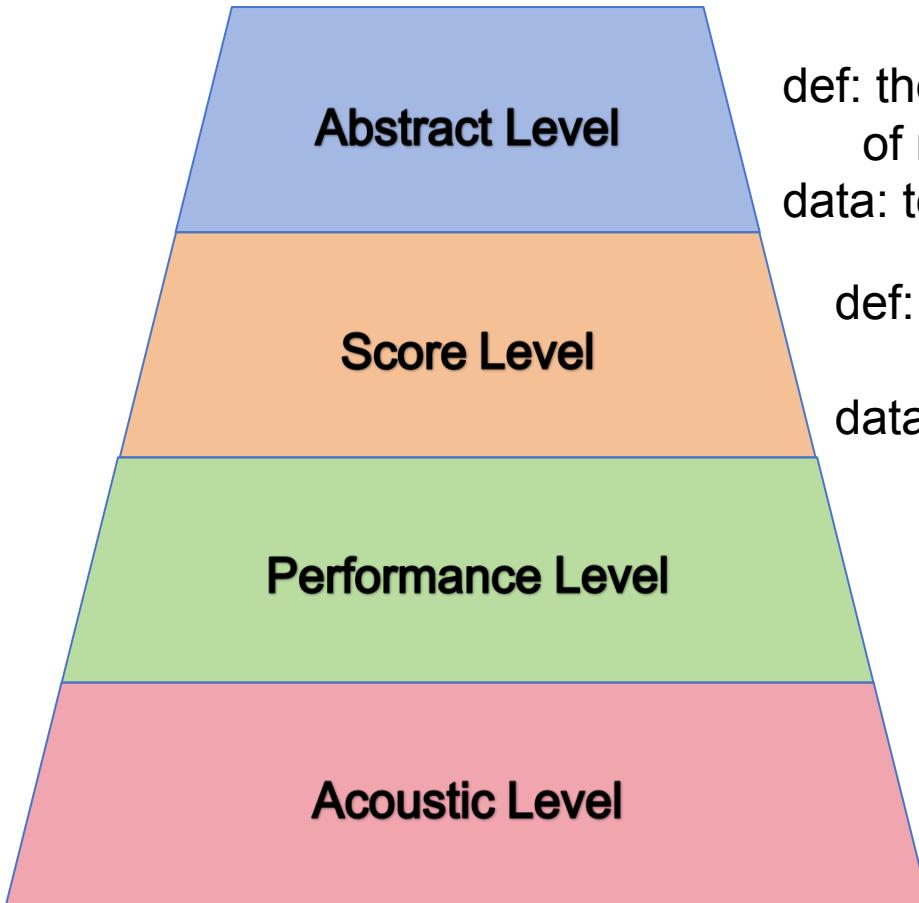
**What are we modeling in music?**

# Music Information Hierarchy: A Performer/generator's Perspective

A musically trained person usually has knowledge of all 4 levels, explicitly (embodied) models score level and performance level music info.



# Music Information Hierarchy: A Performer/generator's Perspective



def: the abstract, emotional, aesthetic, cultural, and historical interpretations of music, involves perception, knowledge, and reasoning of all lower level info  
data: text/language

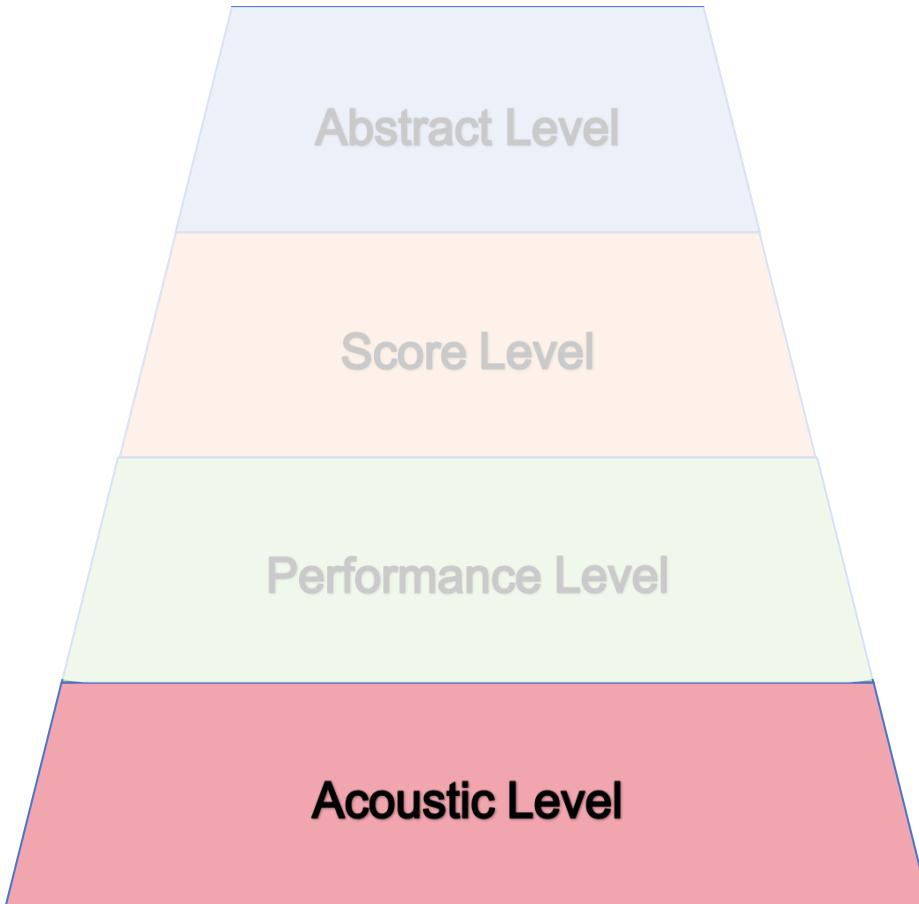
def: musical notation, the quantized control grouped and written by human, describes musical entities

data: image scores, score midi, abc notation, music xml, guitar tabs...

def: the detailed control of acoustic level features, or control of an instrument, usually encodes the expressiveness  
data: performance midi, waveform, spectrogram...

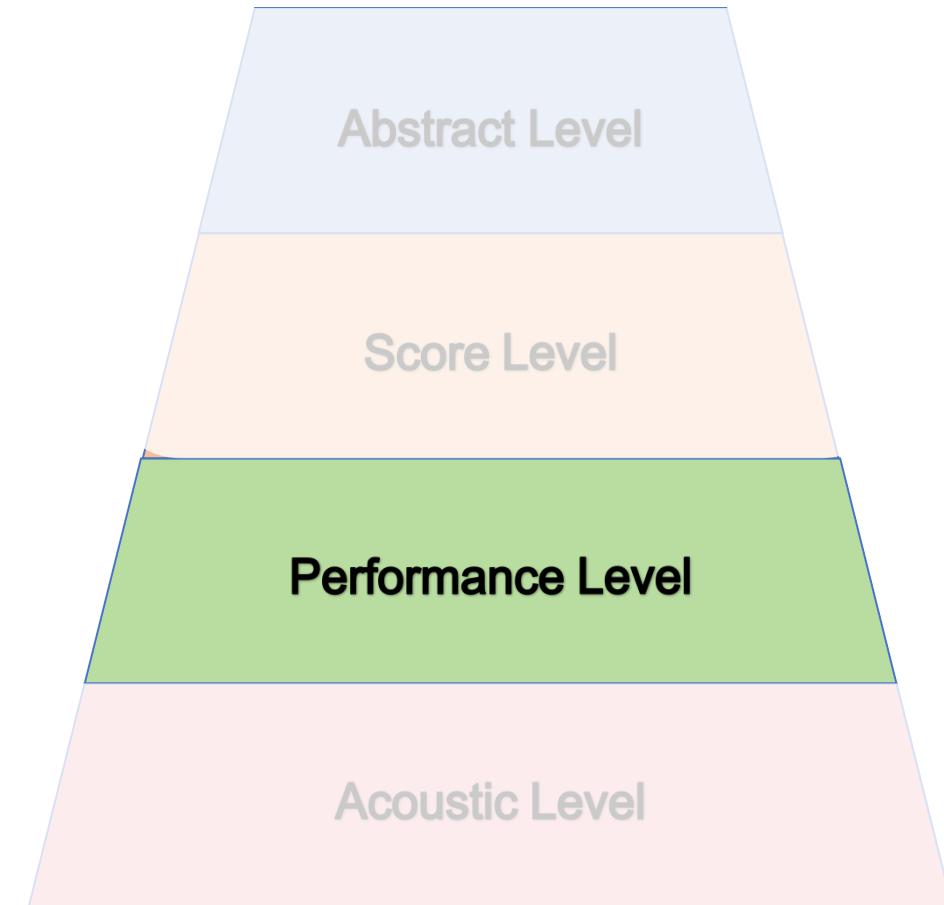
def: frequency, amplitude, duration, timbre, effects( $h(t)$ )...  
data: waveform, spectrogram..

# Music Information Hierarchy: A Performer/generator's Perspective



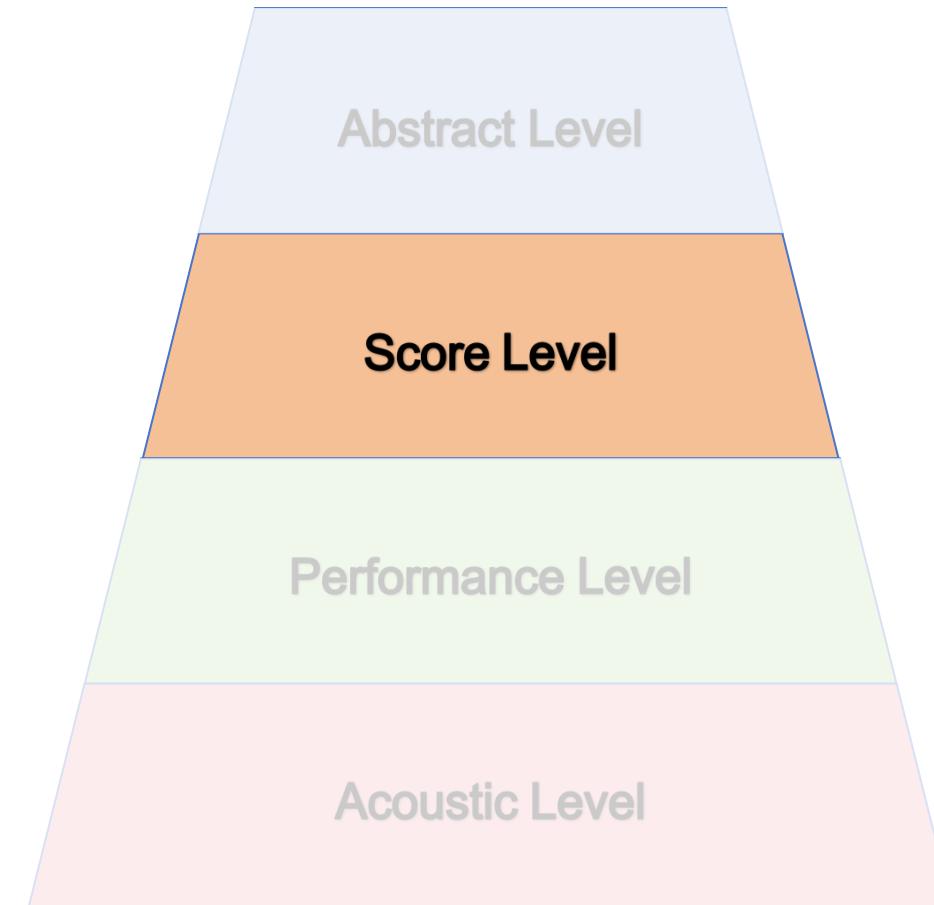
- def: frequency, amplitude, duration, timbre, effects( $h(t)$ )...
  - data: waveform, spectrogram..
  - MIR tasks: instrument id, singer id, source sep, pitch track...
- 
- context len: frame level(~50ms)
  - data sparsity: lots of data since most music are recorded into waveform
  - info rate: **1411kbps** for 44.1khz stereo

# Music Information Hierarchy: A Performer/generator's Perspective



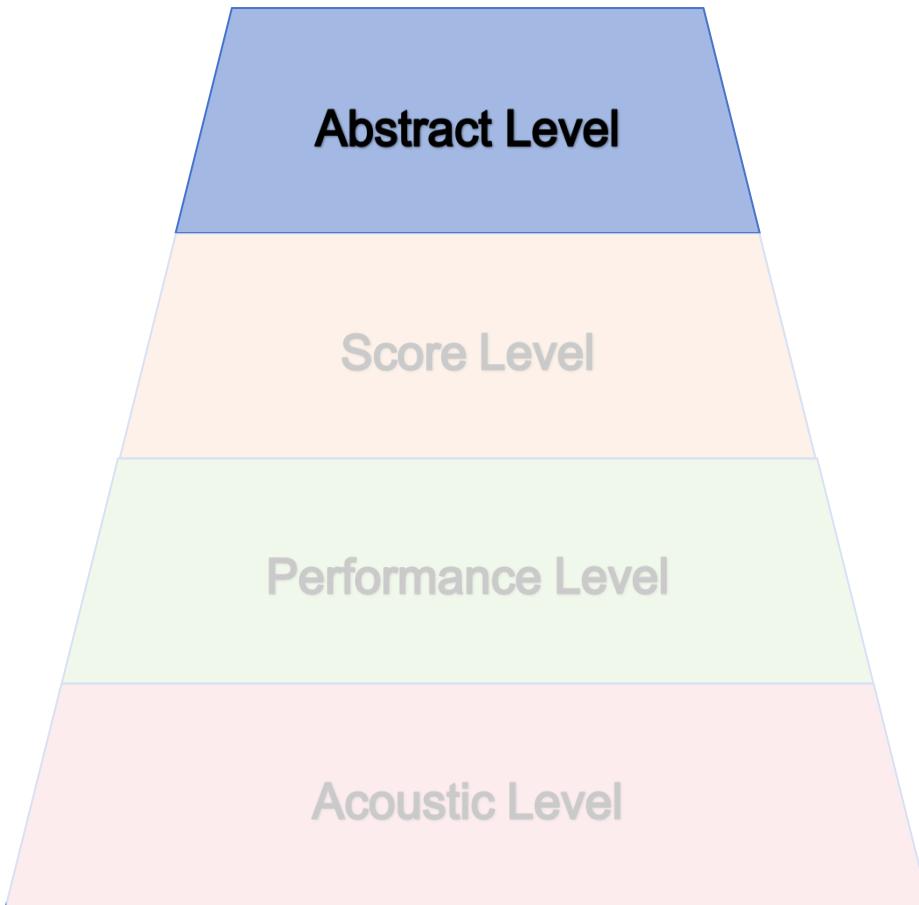
- def: the detailed control of acoustic level features, or control of an instrument, usually encodes the expressiveness
- data: performance midi, waveform, spectrogram..
- MIR tasks: technique recog/eval, fingering pred,...
- context len: note level (.3sec ~ several sec)
- data sparsity: **sparse**, since performance midi is way less than waveform, but you can extract some control signals from waveform
- info rate: **0.48kbps** (120bpm, quarter note, 10 instrument)

# Music Information Hierarchy: A Performer/generator's Perspective



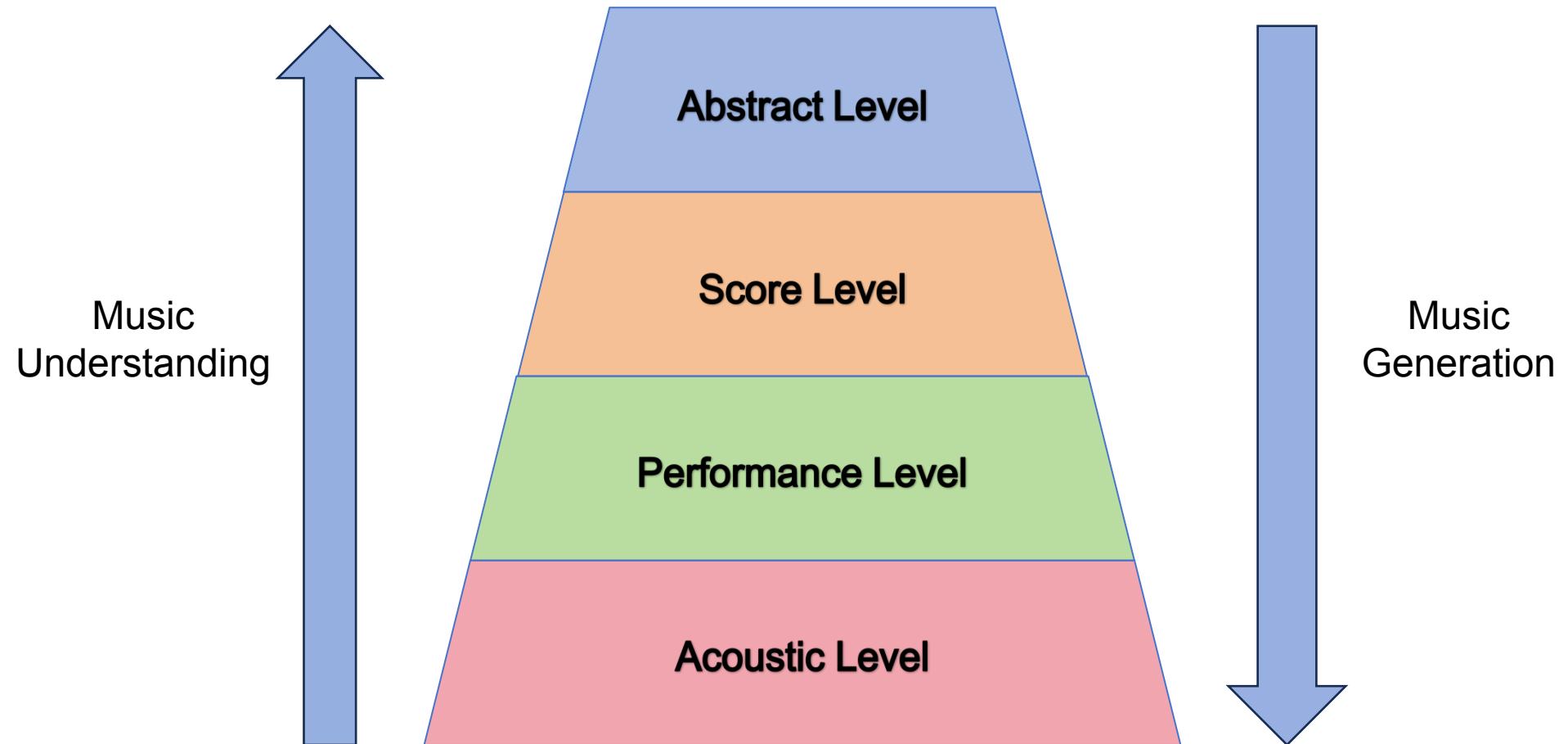
- def: musical notation, the quantized control grouped and written by human, describes musical entities
  - data: image scores, score midi, abc notation, music xml, guitar tabs...
  - MIR tasks: transcription, beat tracking, p. midi to score, score following...
- 
- context len: several notes to song level
  - data sparsity: **very sparse**, only the most popular pieces have scores
  - info rate: **0.013kbps** (300bytes, 3min song, abc notation)

# Music Information Hierarchy: A Performer/generator's Perspective

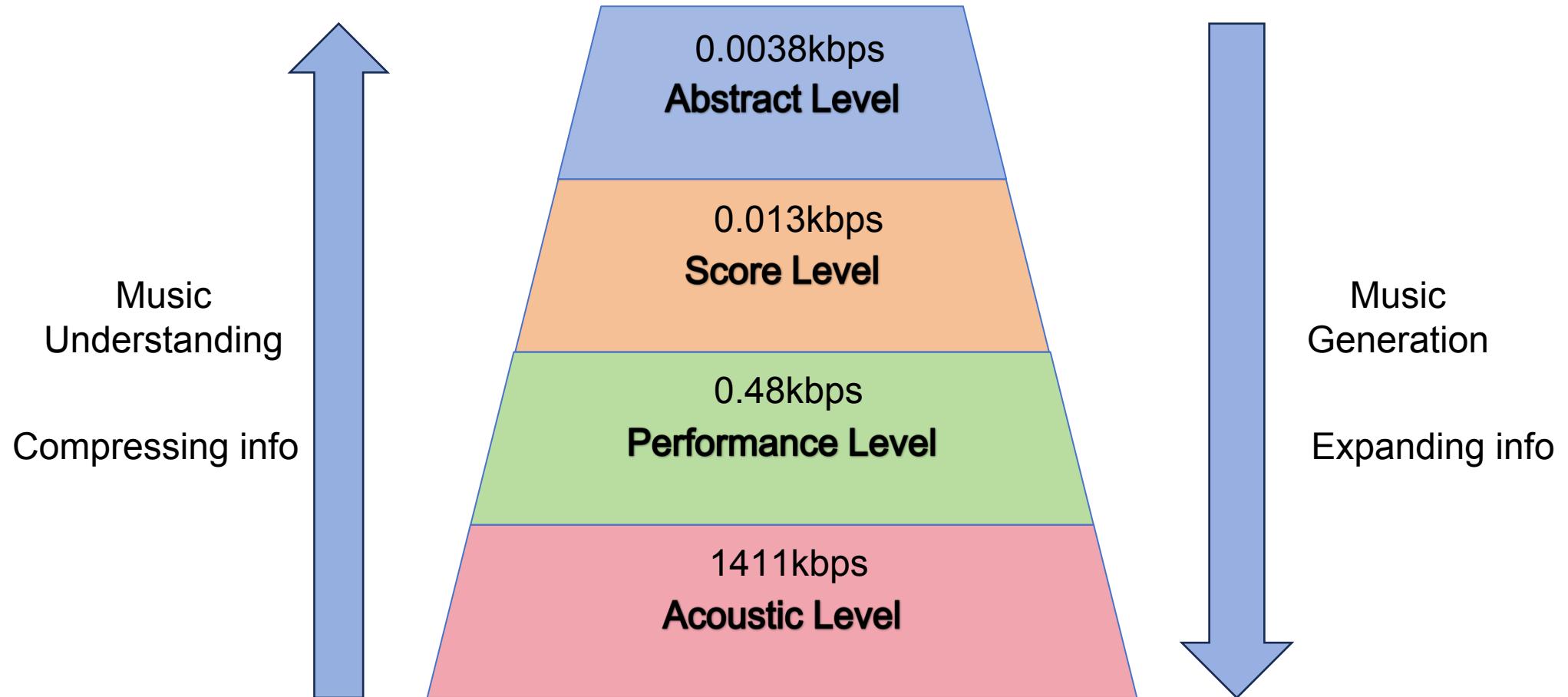


- def: the abstract, emotional, **aesthetic**, cultural, and historical interpretations of music, involves perception, knowledge, and reasoning of all lower level info
- data: text/language...
- MIR tasks: tagging, captioning, genre cls, emo recog...
- context len: song level to database level
- data sparsity: text is everywhere
- info rate: **0.0038kbps** (15words, 3min song)

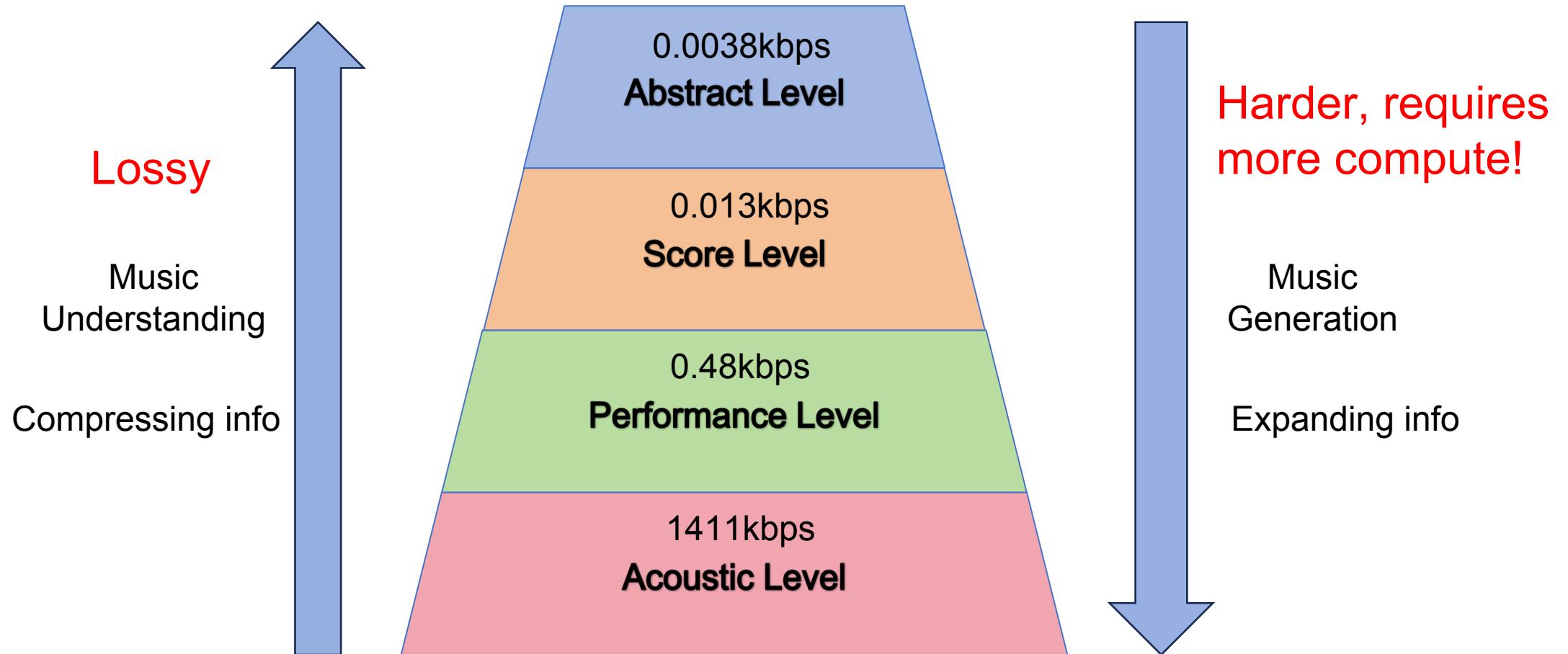
# Music Information Hierarchy: A Performer/generator's Perspective

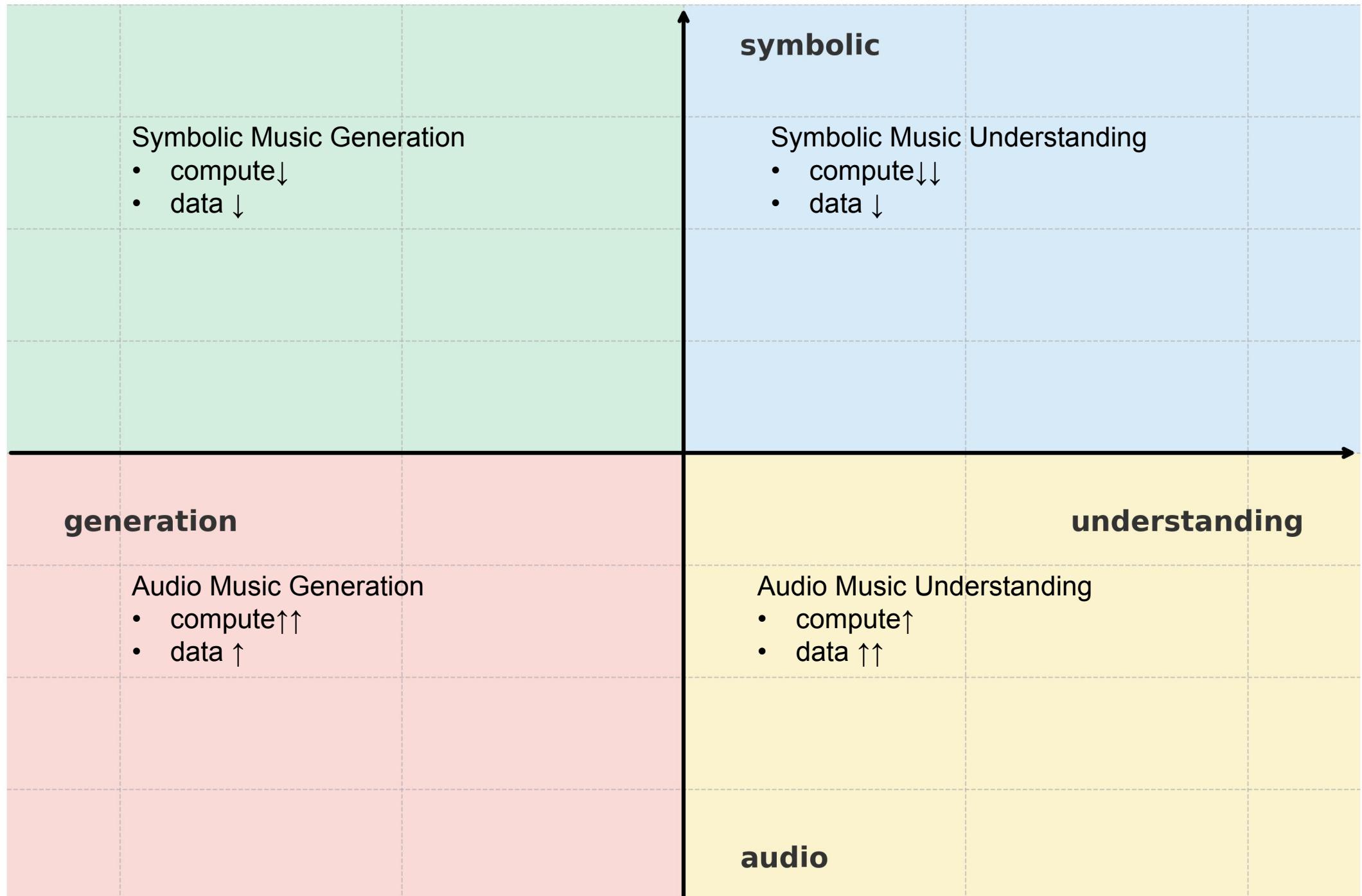


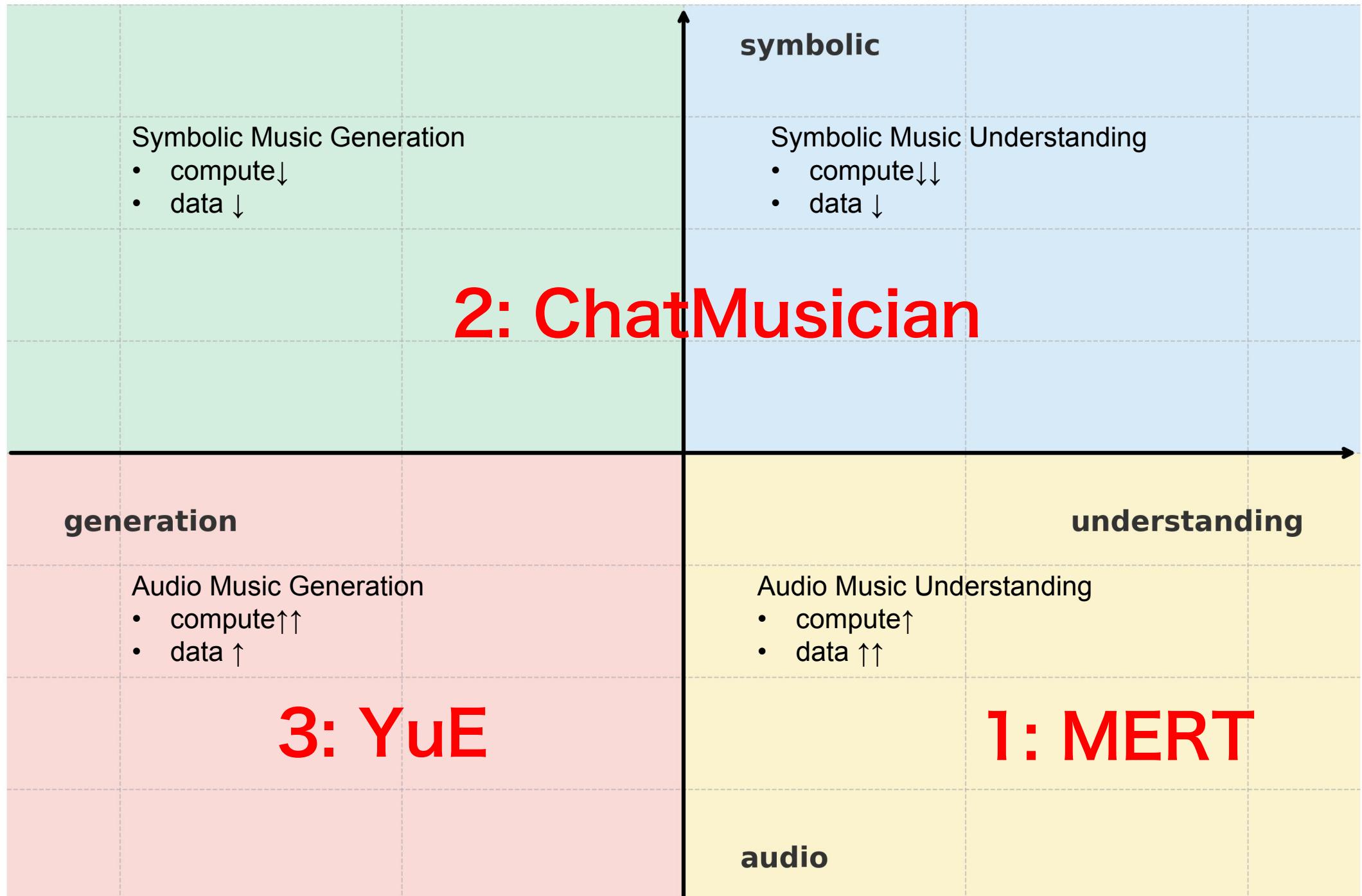
# Music Information Hierarchy: A Performer/generator's Perspective



# Music Information Hierarchy: A Performer/generator's Perspective









# MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training

# Setup

- 4 core developers, several contributors on data and eval
- initially 8xA100s, later scaled to 32-64xA100s
- 160khr audio data
- <=6 months

# MERT

- Acoustic teacher: Encodex

- Musical teacher: CQT

- Paradigm: Masked Language Modeling

- Training setting

1k ~ 160k hours of music

5 sec context

95M ~ 330M parameter size

64 x A100-40G with fp16

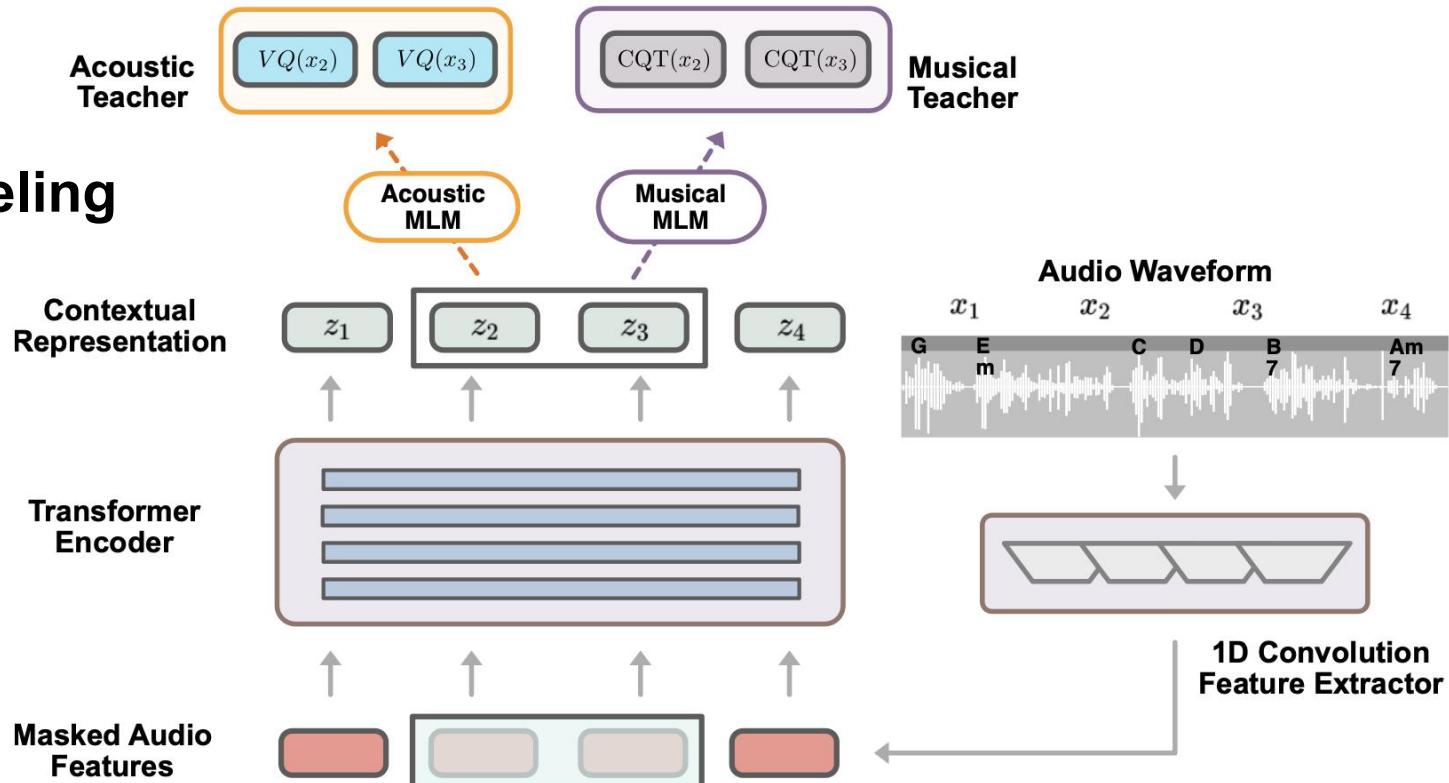
- Variations

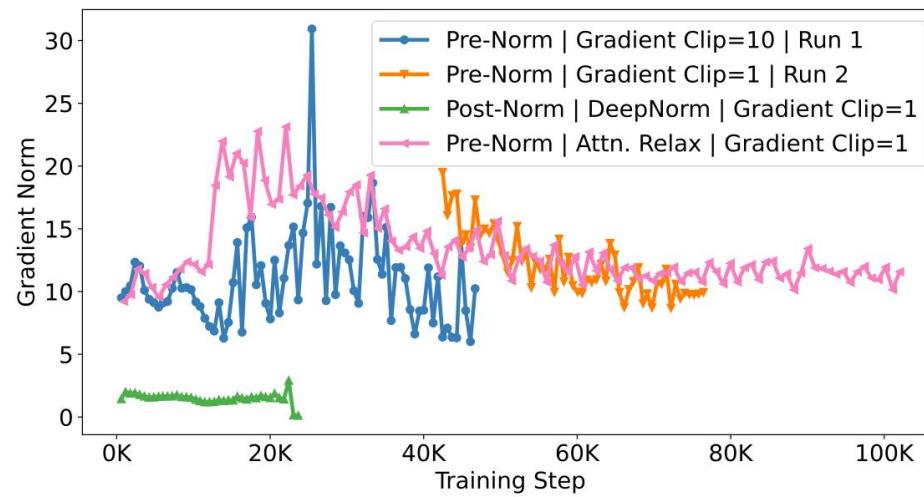
**MERT(-v1)-95M (non-k-means)**

**MERT(-v1)-330M (non-k-means)**

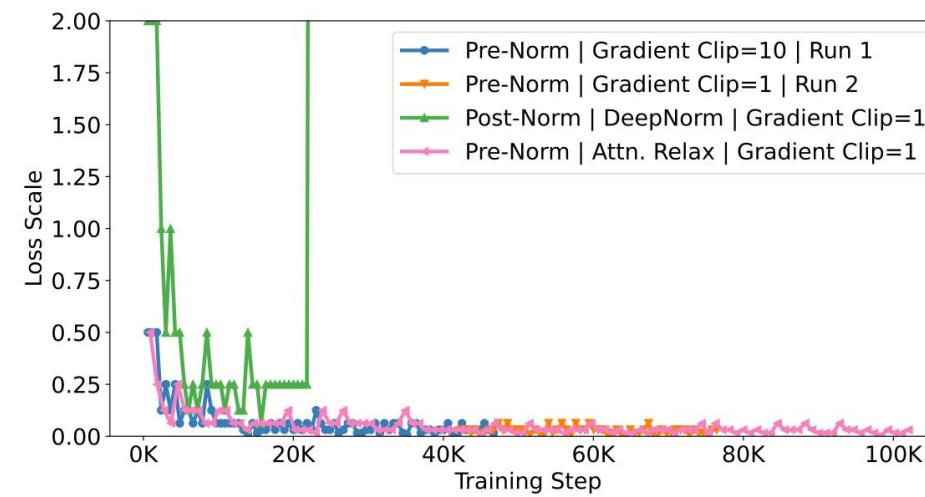
**MERT(-v0)-95M-public (k-means)**

**MERT(-v0)-95M (k-means)**

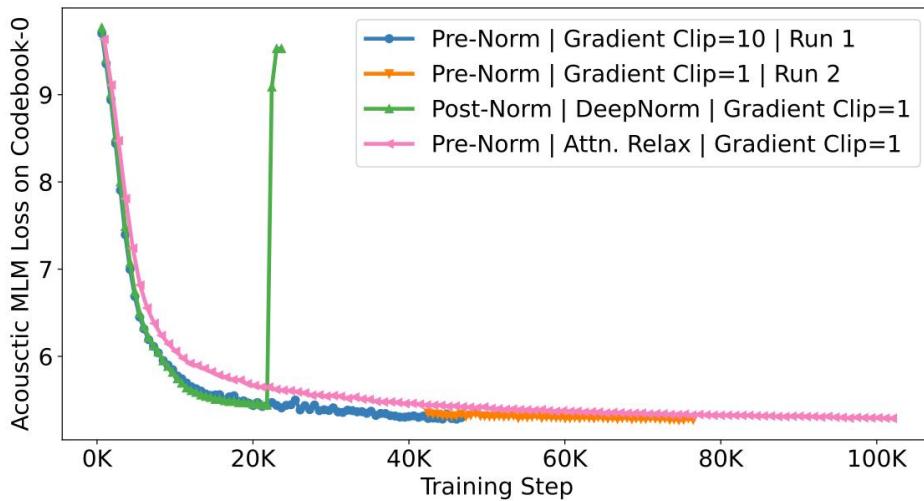




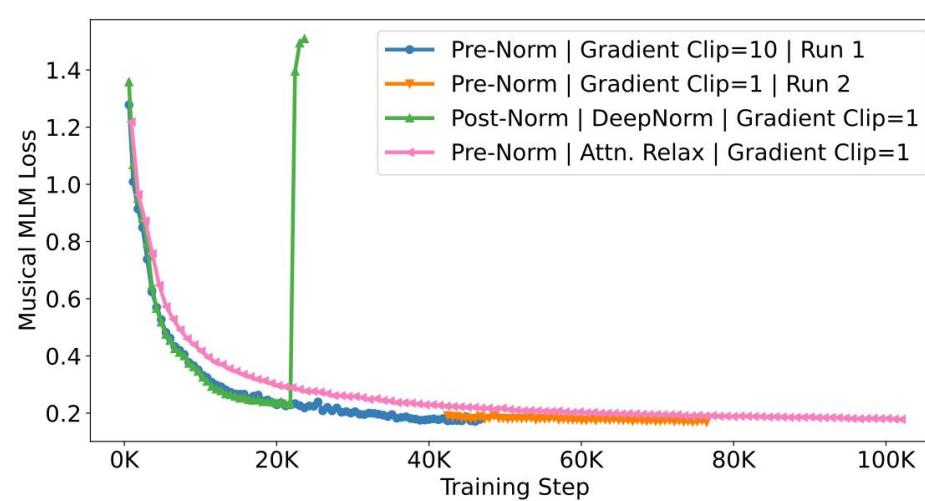
(a) Gradient Norm



(b) Loss Scale



(c) Acoustic MLM Loss on Codebook-0



(d) Music MLM Loss

Figure 2: Illustration of the Training Curves of Trials on Large (330M) Models. Only the acoustic MLM loss on codebook 0 in the RVQ-VAE is shown as the other seven show similar trends.

# MERT: Downstream Evaluation

Dataset Task	MTT Tagging		GS Key	GTZAN Genre	GTZAN Rhythm	EMO Emotion		Nsynth		VocalSet Tech	VocalSet Singer	
Metrics	ROC	AP	Acc <sup>Refined</sup>	Acc	F1 <sup>beat</sup>	R2 <sup>V</sup>	R2 <sup>A</sup>	Acc	Acc	Acc	Acc	
MusiCNN [54]	90.6*	38.3*	12.8*	79.0*	-	46.6*	70.3*	72.6	64.1	70.3	57.0	
CLMR [64]	89.4*	36.1*	14.9*	68.6*	-	45.8*	67.8*	68.3	47.0	60.0	50.7	
Jukebox-5B [11, 76]	91.5*	41.4*	66.7*	79.7*	-	61.7*	72.1*	70.0	90.9	77.6	81.4	
MULE [47]	91.4*	40.4*	66.7*	73.5*	-	57.7*	70*	74.0*	89.2*	75.5	87.8	
HuBERT-base <sup>music</sup> [28]	90.2	37.7	14.7	70.0	<b>88.6</b>	42.1	66.5	69.3	77.4	65.9	75.3	
data2vec-base <sup>music</sup> [2]	89.1	35.1	19.0	59.7	33.5	38.5	61.9	69.4	88.9	68.3	69.5	
MERT-95M <sup>K-means</sup>	90.6	38.4	65.0	78.6	88.3	53.1	68.7	71.3	91.5	74.6	77.2	
MERT-95M-public <sup>K-means</sup>	90.7	38.4	66	71.4	88.1	53.2	71.5	69	91.1	75.5	78.2	
MERT-95M <sup>RVQ-VAE</sup>	91	39.3	63.3	78.6	88.3	60	76.4	69	91.7	74.2	83.7	
MERT-330M <sup>RVQ-VAE</sup>	91.3	40.2	65.6	79.3	87.9	61.2	74.7	71.3	<b>92.4</b>	78.3	87.3	
Previous SOTA	<b>92.0</b> [29]	<b>41.4</b> [11]	<b>74.3</b> [37]	<b>83.5</b> [47]	80.6 [27]	<b>61.7</b>	72.1 [11]	<b>78.2</b> [70]	89.2 [47]	65.6 [74]	80.3 [50]	
Dataset Task	MTG Instrument		MTG MoodTheme		MTG Genre		MTG Top50		MUSDB Source Separation			
Metrics	ROC	AP	ROC	AP	ROC	AP	ROC	AP	SDR <sup>vocals</sup>	SDR <sup>drums</sup>	SDR <sup>bass</sup>	SDR <sup>other</sup>
MusiCNN [54]	76.2	18.6	74.7	12.8	86.0	17.5	82.0	27.3	-	-	-	-
CLMR [64]	73.5	17.0	73.5	12.6	84.6	16.2	81.3	26.4	-	-	-	-
Jukebox-5B [11, 76]	-	-	-	-	-	-	-	-	5.1*	4.9*	4.1*	2.7*
MULE [47]	76.7	19.6	78.1	15.3	88	20.4	83.7	30.7	-	-	-	-
HuBERT-base <sup>music</sup> [28]	75.5	17.8	76.0	13.9	86.5	18.0	82.4	28.1	4.7	3.7	1.8	2.1
data2vec-base <sup>music</sup> [2]	73.1	16.3	74.3	12.2	85.2	16.5	81.4	26.2	5.5	5.5	4.1	3.0
MERT-95M <sup>K-means</sup>	77.2	19.6	75.8	13.6	87.0	18.6	82.8	29.4	5.6	5.6	4.0	3.0
MERT-95M-public <sup>K-means</sup>	76.9	19.2	76.2	13.4	87.2	18.9	82.9	28.7	5.5	5.5	3.7	3.0
MERT-95M <sup>RVQ-VAE</sup>	76.5	19.2	76.5	13.6	87.0	18.8	82.7	28.3	5.5	5.5	3.8	3.1
MERT-330M <sup>RVQ-VAE</sup>	77.0	19.7	76.7	14	87.0	18.6	83.5	29.7	5.3	5.6	3.6	3.0
Previous SOTA	<b>78.8</b>	<b>20.2</b> [1]	<b>78.6</b>	<b>16.1</b> [47]	<b>87.7</b>	<b>20.3</b> [1]	<b>84.3</b>	<b>32.1</b> [47]	<b>9.3</b>	<b>10.8</b>	<b>10.4</b>	<b>6.4</b> [58]

Avg.

64.5

# MARBLE Benchmark: Models

Method	MusiCNN	CLMR	Jukebox	MULE	MAP-Music2Vec	MAP-MERT-v0		MAP-MERT-v1	
	MSD-big					base	base-public	base	large
Network	CNN	9-Conv	3-Conv, 36-Trans	22-Conv, 2-Trans	7-Conv, 12-Trans	7-Conv, 12-Trans	7-Conv, 12-Trans	7-Conv, 12-Trans	7-Conv, 12-Trans
#Params	8M	2.5M	5B	62.4M	95M	95M	95M	95M	330M
Input	log-mel	waveform	waveform	log-mel	waveform	waveform	waveform	waveform	waveform
Stride	3s	2.69s	23.78s	2s	20ms	20ms	20ms	13.3ms	13.3ms
Context Length	3s	2.69s	23.78s	3s	30s	5s	5s	5s	5s
Data (hour)	10~20k	1.7k	60~120k	117.5k	1k	1k	0.9k	17k	160k
Pre-training Task	Music Tagging	Contrastive Learning	CALM	Contrastive Learning	MLM Bootstrapping	MLM Clustering	MLM Clustering	MLM Clustering	MLM Clustering

# Visualization on MARBLE

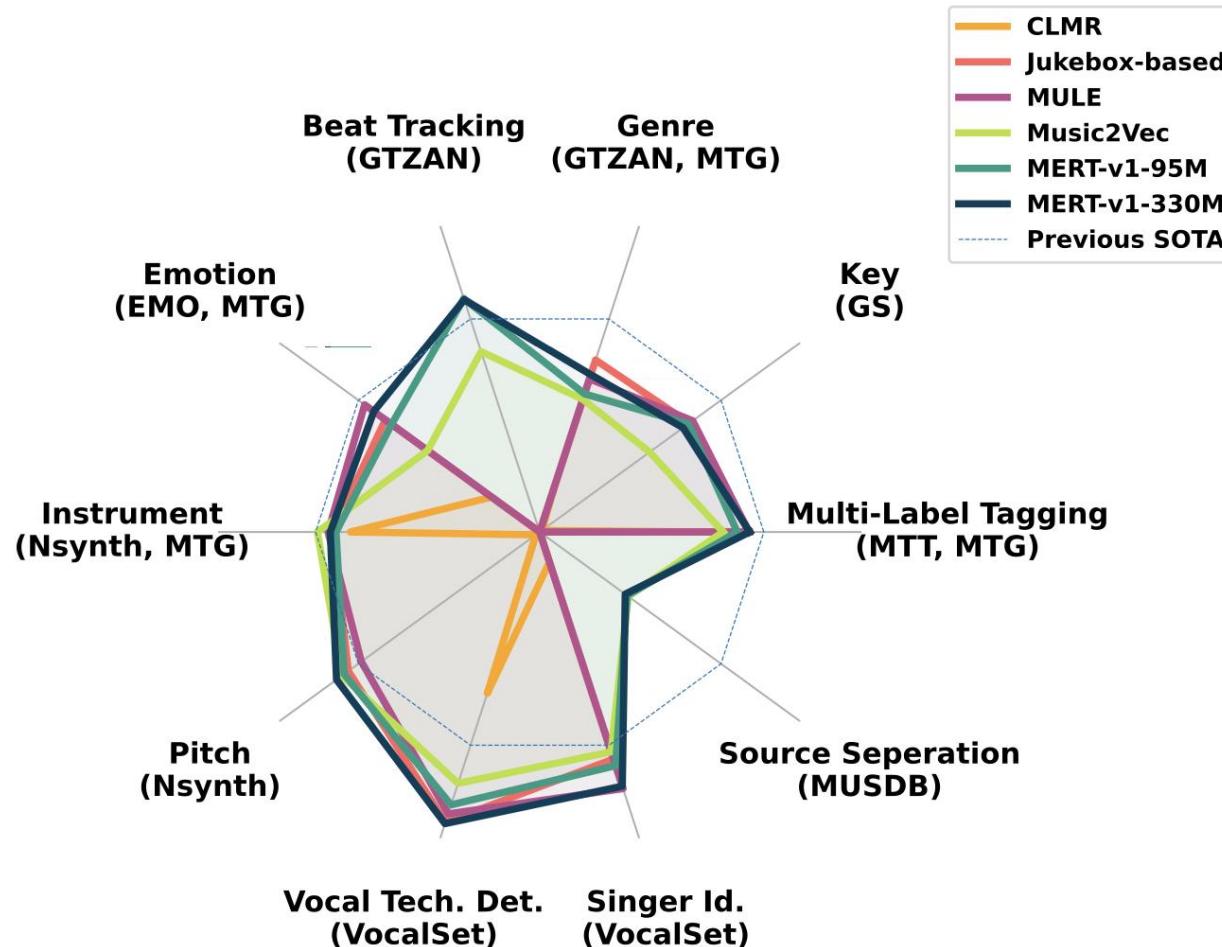
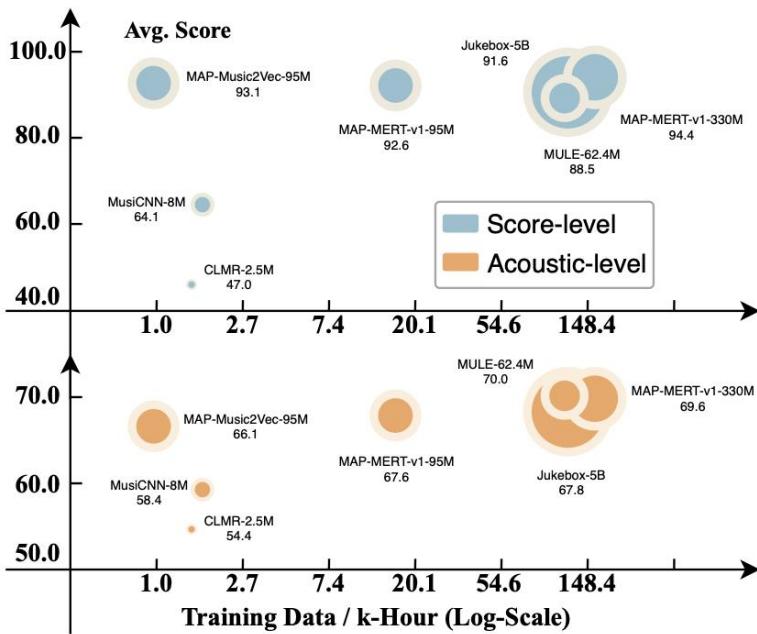
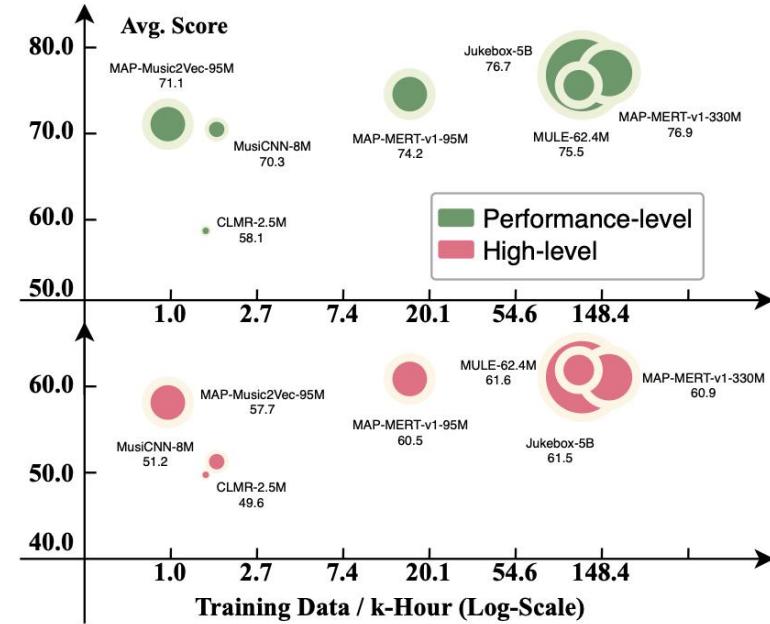


Figure 1: SSL Baselines Compared to previous SOTA. The performances of the tasks are merged according to the task types demonstrated in Tab. 1. Results not applicable are set to 0.

# Visualization on MARBLE



(a) Scores at Acoustic-level and Score-level.



(b) Scores at Performance-level and High-level.

Figure 2: Results Analysis Regarding to Training Data Size. Since some models are not applicable to the sequence labelling tasks, the performances of *source separation* and *beat tracking* tasks are excluded on acoustic-level and score-level average score calculation correspondingly. The radii of the scatter points are isometrically log scaling with the parameter sizes.

# Rethinking MERT

# Rethinking MERT

- **Scaling brings better representation, but we are facing stability issues with this paradigm**
- **All representations have learned multiple levels of knowledge**
- **Fully supervision may overfit the supervision signal**
- **SSL can generalize well**

# Rethinking MERT

- Scaling brings better representation, but we are facing stability issues with this paradigm
- All representations have learned multiple levels of knowledge
- Fully supervision may overfit the supervision signal
- SSL can generalize well

SSLs are used as semantic tokens nowadays, because...

Supervised Tokenizer for disentanglement

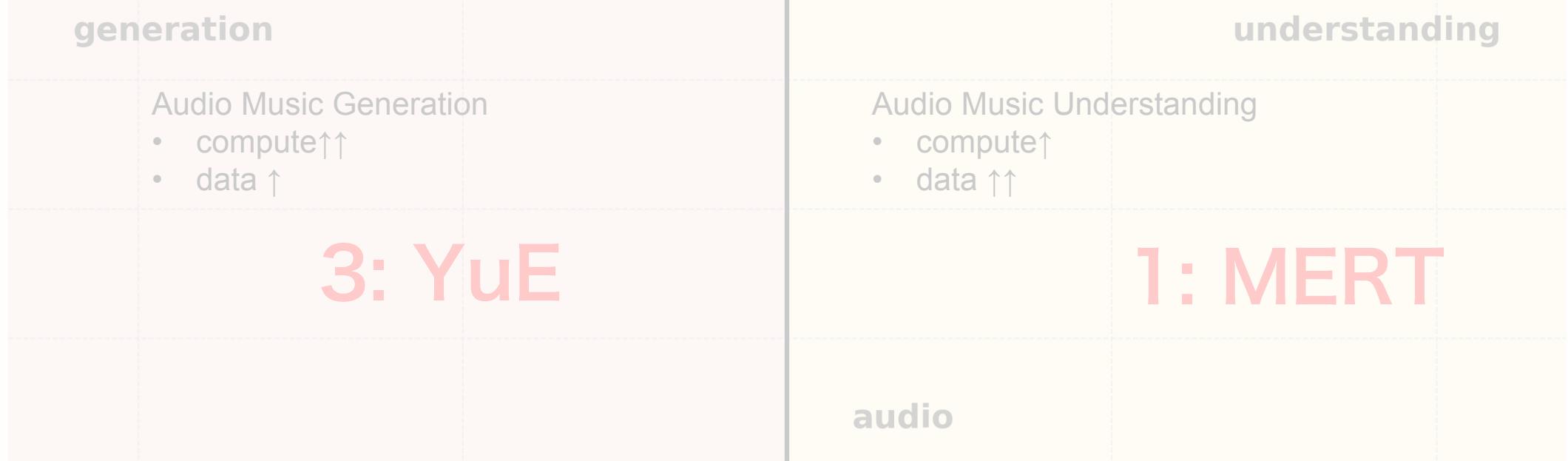
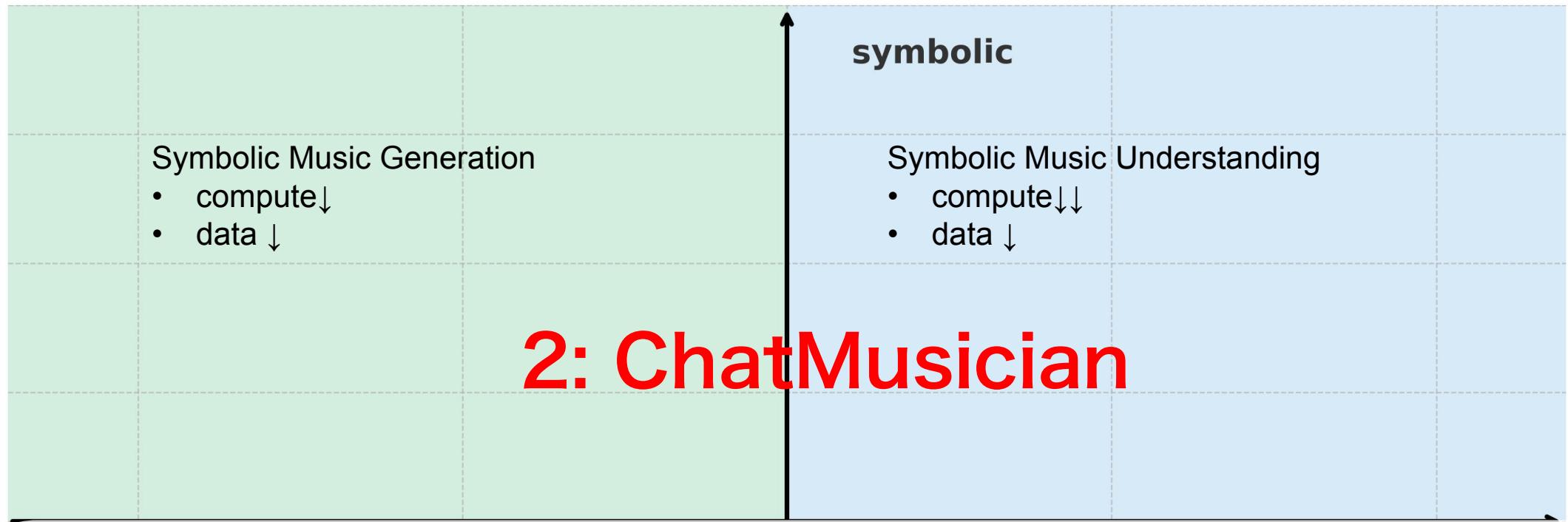
# Rethinking MERT

- **Scaling brings better representation, but we are facing stability issues with this paradigm**
- **All representations have learned multiple levels of knowledge**
- **Fully supervision may overfit the supervision signal**
- **SSL can generalize well**

SSLs are used as semantic tokens nowadays, because...

Supervised Tokenizer for disentanglement

Downstreams: fadtk, mullama, semantic token, clamp3



# **ChatMusician: Understanding and Generating (Symbolic) Music Intrinsically with LLM**

# Introduction

- LLMs are now good at text generation, solving STEM problems.
- But how about creative activities?
- e.g. writing music?



<https://neilpatel.com/blog/bert-google/>



<https://ai.meta.com/blog/large-language-model-llama-meta-ai/>

# Introduction

- LLMs are now good at text generation, solving STEM problems.
- But how about creative activities?
- e.g. writing music?

**Actually GPT-4 do compose  
music~**



<https://neilpatel.com/blog/bert-google/>



<https://ai.meta.com/blog/large-language-model-llama-meta-ai/>

# Introduction

## Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck      Varun Chandrasekaran      Ronen Eldan      Johannes Gehrke  
Eric Horvitz      Ece Kamar      Peter Lee      Yin Tat Lee      Yuanzhi Li      Scott Lundberg  
Harsha Nori      Hamid Palangi      Marco Tulio Ribeiro      Yi Zhang

Microsoft Research

# Introduction

## Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck      Varun Chandrasekaran      Ronen Eldan      Johannes Gehrke  
Eric Horvitz      Ece Kamar      Peter Lee      Yin Tat Lee      Yuanzhi Li      Scott Lundberg  
Harsha Nori      Hamid Palangi      Marco Tulio Ribeiro      Yi Zhang

Microsoft Research

"In summary, the model was able to produce valid tunes in ABC notation and, to some extent, explain and manipulate their structure. "

# Introduction

## Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck      Varun Chandrasekaran      Ronen Eldan      Johannes Gehrke  
Eric Horvitz      Ece Kamar      Peter Lee      Yin Tat Lee      Yuanzhi Li      Scott Lundberg  
Harsha Nori      Hamid Palangi      Marco Tulio Ribeiro      Yi Zhang

Microsoft Research

"In summary, the model was able to produce valid tunes in ABC notation and, to some extent, explain and manipulate their structure. "

**How does it sound like?**

# The Cherry-picked sample in Sparks of AGI Paper

GPT-4

Can you compose a short tune (say four to eight bars) using ABC notation?  
There are many possible ways to compose a short tune in abc notation, but here is one example:

X:1  
T:Example Tune  
M:4/4  
L:1/8  
K:C  
[:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :|  
:c2de f2ed | c2de f2f2 | g2fe d2cB | c4-c2z2 :|

This tune has two four-bar phrases, each repeated once, forming an A part and a B part. The key is C major.

Can you describe the tune in musical terms?

Here is one possible description:

The tune starts with a rising arpeggio of the tonic chord (C major), followed by a descending scale that returns to the tonic. This pattern is repeated, but the second time the scale continues to the higher tonic note. The first phrase ends with a long tonic note.

The second phrase begins with a similar pattern, but using the subdominant chord (F major) instead of the tonic. The second time, the scale continues to the higher subdominant note. The phrase ends with a stepwise descent to the tonic, followed by a long tonic note.

The A part and the B part have a similar structure, but use different chords and ranges. The tune has a simple and catchy melody, with a clear sense of tonality and rhythm.

I like part A, but maybe part B is just a bit too similar to part A, making the song repetitive. How about beginning part B with a descending arpeggio instead of a rising one?

Sure, that would create some contrast and variation. Here is one possible way to do that:

X:1  
T:Example Tune  
M:4/4  
L:1/8  
K:C  
[:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :|  
:f2ed c2BA | f2ed c2c2 | g2fe d2cB | c4-c2z2 :|

That's much better. Now how about making this into a duet, can you add a second staff for this with a bass accompanying the melody?

One possible way to write a second staff for this tune is to use root notes and fifths of the chords, as well as some passing notes, to create a simple accompaniment. Here is an example in abc notation, with the second hand bass on the lower staff:

X:1  
T:Example Tune  
M:4/4  
L:1/8  
K:C  
V:1  
[:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :|  
:f2ed c2BA | f2ed c2c2 | g2fe d2cB | c4-c2z2 :|  
V:2 clef=bass  
[:C,2C,2 G,2G,2 | C,2C,2 G,2G,2 | F,2F,2 E,2E,2 | C,4-C,2z2 :|  
:F,2F,2 C,2C,2 | F,2F,2 C,2C,2 | G,2G,2 F,2F,2 | E,2C,2 G,2C,2 :|

The sheet music of the final tune:

20

Figure 2.9: Generating and modifying a tune in ABC notation.

X:1

T:Example Tune

M:4/4

L:1/8

K:C

V:1

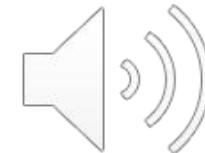
[:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-G2z2 :|

[:f2ed c2BA | f2ed c2c2 | g2fe d2cB | c4-c2z2 :|

V:2 clef=bass

[:C,2C,2 G,2G,2 | C,2C,2 G,2G,2 | F,2F,2 E,2E,2 | C,4-C,2z2 :|

[:F,2F,2 C,2C,2 | F,2F,2 C,2C,2 | G,2G,2 F,2F,2 | E,2C,2 G,2C,2 :|



# Non-cherry-picked Sample by Me

You  
Generate a song in ABC notation

< 1 / 2 >

ChatGPT

Creating a song in ABC notation involves specifying the melody, rhythm, and key signature in a text-based format that is both compact and readable. ABC notation is widely used in music transcription for its simplicity and effectiveness in representing musical scores. Here's a simple example of a song in ABC notation, demonstrating the structure and components such as the header (with fields like X for index number, T for title, M for meter, L for default note length, K for key), followed by the musical notes:

```
abc
X:1
T:Simple Melody
M:4/4
L:1/4
K:Cmaj
C2 D E F | G4 | A2 G F E | D4 | C2 D E F | G4 | A2 G F E | D4 |]
```

This song is in the key of C major, has a 4/4 meter (common time), and a default note length of a quarter note. The melody starts with a simple ascending scale from C to F, holds a G note for a full measure, descends from A back to E, and repeats a similar pattern before ending. Each measure is separated by a vertical bar '|', and the song concludes with a double bar '| |'.

To expand on this concept and inspire creativity in AI and music, consider how algorithms could generate more complex compositions by varying meter, rhythm, key signatures, and incorporating chord progressions. The exploration of generative music through AI not only demonstrates the potential for AI in artistic domains but also offers insights into the nature of creativity and the structure of musical expression. Encouraging AI to collaborate with human artists could lead to novel and inspiring compositions, blurring the lines between human and machine creativity.



X:1

T:Simple Melody

M:4/4

L:1/4

K:Cmaj

C2 D E F | G4 | A2 G F E | D4 | C2 D E F | G4 | A2 G F E | D4 |



# Non-cherry-picked Sample by Me

You  
Generate a song in ABC notation

< 1 / 2 >

ChatGPT

Creating a song in ABC notation involves specifying the melody, rhythm, and key signature in a text-based format that is both compact and readable. ABC notation is widely used in music transcription for its simplicity and effectiveness in representing musical scores. Here's a simple example of a song in ABC notation, demonstrating the structure and components such as the header (with fields like X for index number, T for title, M for meter, L for default note length, K for key), followed by the musical notes:

```
abc
X:1
T:Simple Melody
M:4/4
L:1/4
K:Cmaj
C2 D E F | G4 | A2 G F E | D4 | C2 D E F | G4 | A2 G F E | D4 |
```

This song is in the key of C major, has a 4/4 meter (common time), and a default note length of a quarter note. The melody starts with a simple ascending scale from C to F, holds a G note for a full measure, descends from A back to E, and repeats a similar pattern before ending. Each measure is separated by a vertical bar '|', and the song concludes with a double bar '| |'.

To expand on this concept and inspire creativity in AI and music, consider how algorithms could generate more complex compositions by varying meter, rhythm, key signatures, and incorporating chord progressions. The exploration of generative music through AI not only demonstrates the potential for AI in artistic domains but also offers insights into the nature of creativity and the structure of musical expression. Encouraging AI to collaborate with human artists could lead to novel and inspiring compositions, blurring the lines between human and machine creativity.



X:1

T:Simple Melody

M:4/4

L:1/4

K:Cmaj

C2 D E F | G4 | A2 G F E | D4 | C2 D E F | G4 | A2 G F E | D4 |



## How do you feel?

# **Trivial!**

# What is Missing?

- Big tech companies focus too much on STEM
- They forget to teach LLMs Art and Creativity!
- Maybe it's our music people's duty to inject musical abilities into AGI

# What is Missing?

- Big tech companies focus too much on STEM
- They forget to teach LLMs Art and Creativity!
- Maybe it's our music people's duty to inject musical abilities into AGI

Quote from ChatMusician paper: "While Large Language Models (LLMs) demonstrate impressive capabilities in text generation, we find that their ability has yet to be generalized to music, humanity's creative language. "

# ChatMusician: Overview

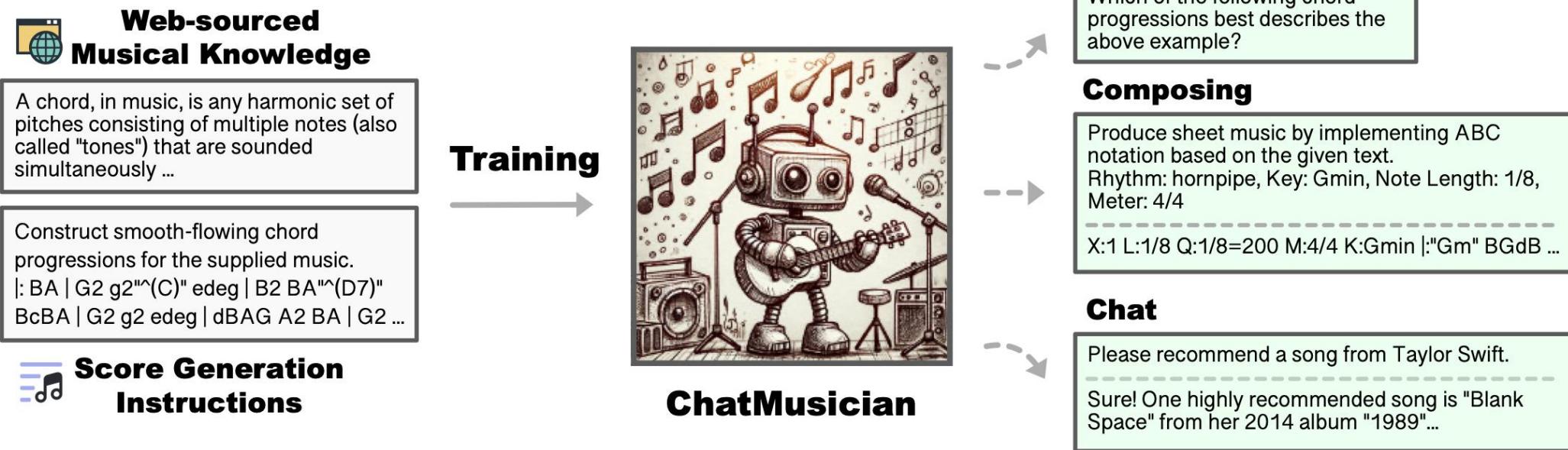


Figure 1: ChatMusician learns from web-sourced musical knowledge and handcrafted music score generation instructions, unifies music generation and music understanding, and can chat, compose, and answer college-level music theory questions.

# Setup

- 5 core developers, dozens contributors on data and eval
- 16xA100s, a dozen V100s
- 5B tokens
- <=6 months

# Music Style Transfer

## INPUT

Construct smooth-flowing chord progressions for the supplied music.

|: g2 a g e8 z4 | d2 e2 c4 z8 | A4 g2 e2 d2 c2 A4 | G12 z4 |  
g2 a g e8 z4 | C'4 a2 g6 z4 | A4 g2 e2 e2 d2 c4 | d12 z4 |



## OUTPUT

X:1

L:1/16

M:3/4

K:Cmin

"A" |:"Cm" g2 a g e8 z4 |"Fm" d2 e2 c4 z8 |"G7" A4 g2 e2  
d2 c2 A4 |"Cm" G12 z4 |  
"Cm" g2 a g e8 z4 |"Fm" C'4 a2 g6 z4 |"G7" A4 g2 e2 e2  
d2 c4 |"G7" d12 z4 :|"B" |:"G7" d8"Eb" e8 |  
f2 gf e2d2"Cm" c8 |"G7" Bcde cdBc"Cm" ABGA G4 |"G7"  
GABC dedc"Cm" =Bcd=B c4 |"G7" =Bcd\_e fe\_dc"Cm"  
\_BAGF E4 |  
"G7" F2 G2 D4"Cm" E8 |"G7" F2 G2 D4"Cm" C8 |"G7"  
GABC de=Bc"Cm" cdAc B4 |"G7" GABC defg"Cm" a8 |  
"G7" b2 ag ^f2g2"Cm" g8 |"G7" a2 g^f =efge"Cm" de=Bc  
A4 |"G7" GABC d2e2"Cm" c8 |"G7" GABC d2e2"Cm" c8 :|



# ChatMusician: Data

Datasets	Sourced from	Tokens	# Samples	Category	Format
Pile (Gao et al., 2020)	public dataset	0.83B	18K	general	article
Falcon-RefinedWeb (Penedo et al., 2023)	public dataset	0.80B	101K	general	article
Wikipedia (Wikipedia contributors, 2023)	public dataset	0.39B	588K	general	article
OpenChat (Wang et al., 2023a)	public dataset	62.44M	43K	general	chat
LinkSoul (LinkSoul-AI, 2023)	public dataset	0.6B	1.5M	general	chat
GPT4-Alpaca (Peng et al., 2023)	public dataset	9.77M	49K	general	chat
Dolly (Conover et al., 2023)	public dataset	3.12M	14K	general	chat
Irishman (Wu and Sun, 2023)	public dataset + Human-written Instructions	0.23B	868K	music score	chat
KernScores (CCARH at Stanford University, 2023)	public dataset + Human-written Instructions	2.76M	10K	music score	chat
Bach (Wu et al., 2023)	public dataset + Human-written Instructions	0.44M	349	music score	chat
synthetic music chat☆	public dataset + Human-written Instructions	0.54B	50K	music score	chat
music knowledge☆	Generated w/ GPT-4	0.22B	255K	music verbal	chat
music summary☆	Generated w/ GPT-4	0.21B	500K	music verbal	chat
GSM8k (Cobbe et al., 2021)	public dataset	1.68M	7K	math	chat
math (Kenney, 2023)	public dataset	7.03M	37K	math	chat
MathInstruct (Yue et al., 2023)	public dataset	55.50M	188K	math	chat
Camel-Math (Li et al., 2023)	public dataset	27.76M	50K	math	chat
arxiv-math-instruct-50k (Kenney, 2023)	public dataset	9.06M	50K	math	chat
Camel-Code (Li et al., 2023)	public dataset	0.13B	366K	code	chat
OpenCoder (Wang et al., 2023a)	public dataset	36.99M	28K	code	chat
Total		4.16B	5.17M		

Table 1: Overview of MusicPile. ☆means synthesis from music score data and general data. ★means with NEW rationales curated by us by prompting GPT-4.

# ChatMusician: Data

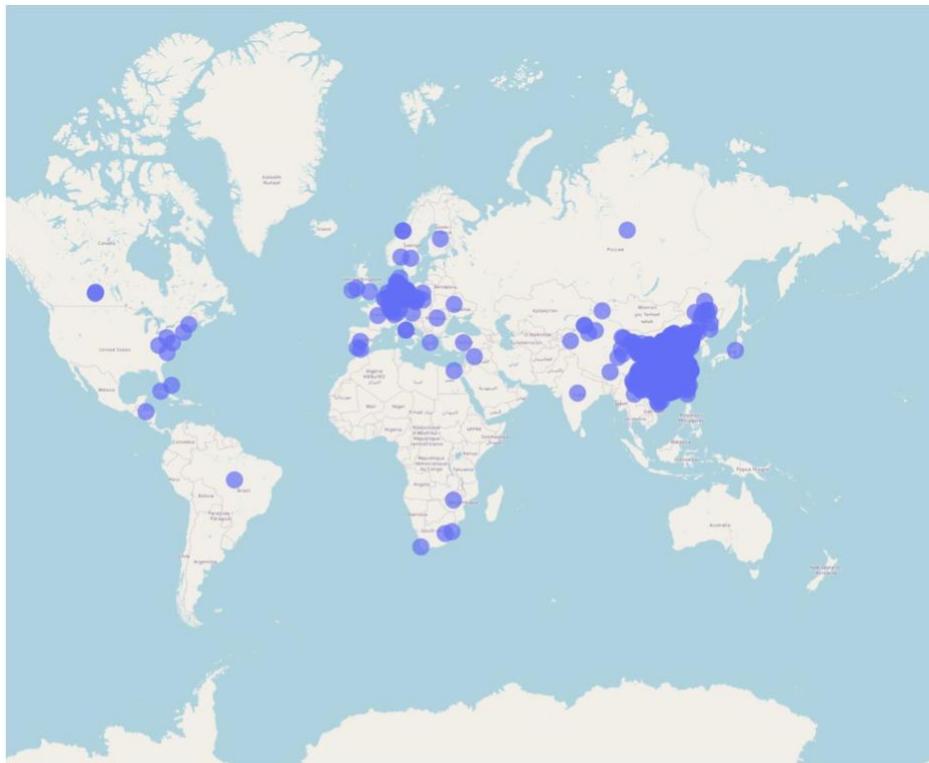


Figure 3: We included diverse music scores from around the world in **MusicPile**. The distribution of a portion of music scores containing regional information has been marked with blue points on the world map.

# ChatMusician: Advanced Music Understanding

- We collected the 1<sup>st</sup> college-level music understanding benchmark, **MusicTheoryBench**
- We defined music knowledge and music reasoning

**Question:** "Which of the following statements about triads is correct?",

**Options:**

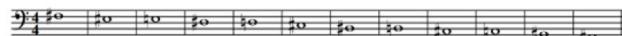
- A: "A triad can only be composed of three notes.",
- B: "All triads are consonant chords.",
- C: "When a triad is inverted, its properties and consonance remain consistent with its original position.",
- D: "A triad is defined as a chord formed by stacking three notes in a third relationship."},

**Answer:** D

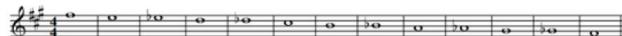
**Question:** "Which of the following is a descending natural minor scale with E as the leading tone?",

**Options:**

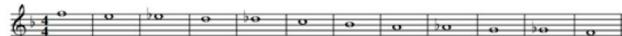
- A: "L:1/4\nM:4/4\nK:C\n^F,4 | ^E,4 | =E,4 | ^D,4 | =D,4 | ^C,4 | ^B,,4 | =B,,4 | ^A,,4 | =A,,4 | ^G,,4 | ^F,,4 ] %12",



- B: "L:1/4\nM:4/4\nK:A\nf4 | e4 | \_e4 | d4 | \_d4 | c4 | B4 | \_B4 | A4 | \_A4 | G4 | \_G4 | F4 ] %13",



- C: "L:1/4\nM:4/4\nK:F\nf4 | e4 | \_e4 | d4 | \_d4 | c4 | B4 | A4 | \_A4 | G4 | \_G4 | F4 ] %12",



- D: "None of the first three options are correct"},

**Answer:** D

Figure 4: Simple examples of (a) music knowledge and (b) music reasoning from MusicTheoryBench. Question a. mainly includes concepts that can be answered through memorizing them. Question b. requires the knowledge of *descending, natural minor scale* and *leading tone*, and inference based on the musical score.

# ChatMusician: Advanced Music Understanding

- We collected the 1<sup>st</sup> college-level music understanding benchmark, **MusicTheoryBench**
- We defined music knowledge and music reasoning

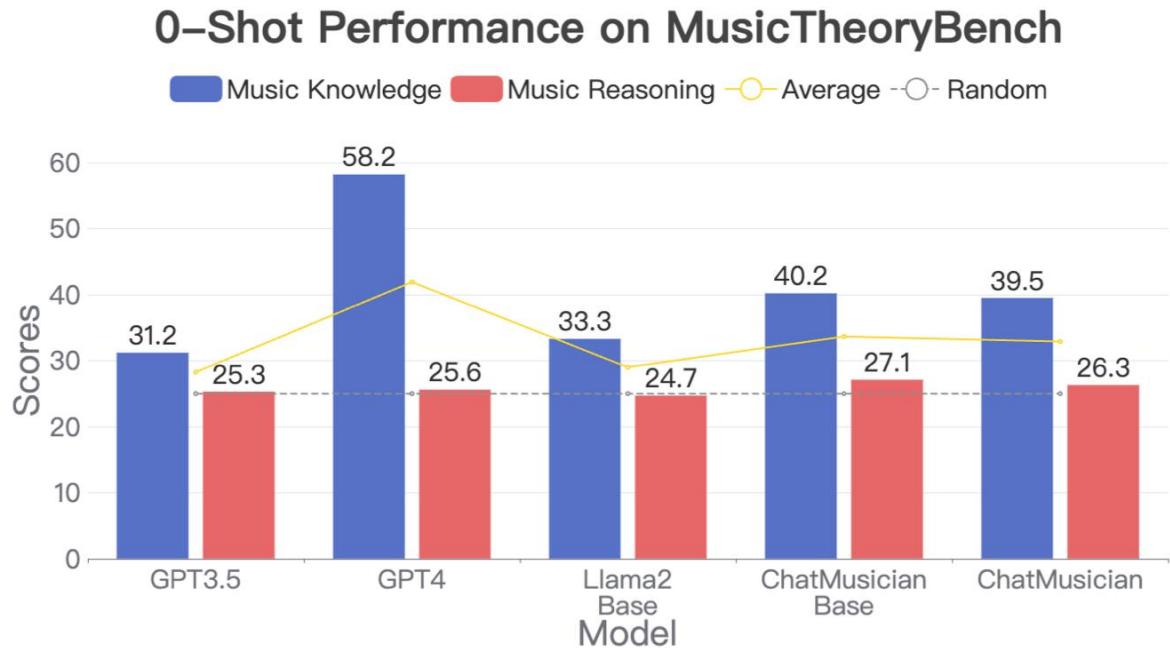


Figure 5: Zero-shot accuracy on MusicTheoryBench. We included GPT-3.5, GPT-4, LLaMA2-7B-Base, ChatMusician-Base, and ChatMusician. The blue bar represents the performance on the music knowledge metric, and the red bar represents the music reasoning metric. The dashed line corresponds to a random baseline, with a score of 25%.

# ChatMusician: Advanced Music Understanding

- We collected the 1<sup>st</sup> college-level music understanding benchmark,  
**MusicTheoryBench**
- We defined music knowledge and music reasoning
- GPT-4 is still the best system for language-like sequence understanding
- But not good enough

Method	Mus. Knowledge	Mus. Reasoning
GPT4-0-shot	58.2	25.6
+5-shot ICL	64.1	38.0
GPT4-RolePlay	68.3	36.6
+5-shot ICL	68.8	<b>39.5</b>
GPT4-CoT	68.4	36.7
+5-shot ICL	<b>69.9</b>	34.9

Table 3: We further conducted prompt engineering on GPT-4 to check the upper limit on MusicTheoryBench. We included the techniques of chain-of-thoughts, role-play, and 5-shot in-context-learning. The highest score we achieved on music knowledge metric is 69.9, and 39.5 on music reasoning metric.

# ChatMusician: Advanced Music Understanding

- We collected the 1<sup>st</sup> college-level music understanding benchmark,  
MusicTheoryBench
- We defined music knowledge and music reasoning
- GPT-4 is still the best system for language-like sequence understanding
- But not good enough

Sadly, we find that ChatMusician shows weak ICL and CoT capabilities.  
Probably need to improve the data quality.

Method	Mus. Knowledge	Mus. Reasoning
GPT4-0-shot	58.2	25.6
+5-shot ICL	64.1	38.0
GPT4-RolePlay	68.3	36.6
+5-shot ICL	68.8	<b>39.5</b>
GPT4-CoT	68.4	36.7
+5-shot ICL	<b>69.9</b>	34.9

Table 3: We further conducted prompt engineering on GPT-4 to check the upper limit on MusicTheoryBench. We included the techniques of chain-of-thoughts, role-play, and 5-shot in-context-learning. The highest score we achieved on music knowledge metric is 69.9, and 39.5 on music reasoning metric.

# ChatMusician: Music Generation

Task Name	Type	Example Instruction
Chord Conditioned Music Generation	G	Develop a musical piece using the given chord progression. [CHORDS]
Musical Form Conditioned Music Generation	G	Craft a musical work that incorporates the given musical pattern as a central element. [MUSICAL FORMS]
Alphabetic Musical Form and Motif Conditioned Music Generation	G	Develop a musical piece employing the provided motif and an alphabet-based structure. [MUSICAL FORMS A] [MOTIF]
Terminology Musical Form and Motif conditioned Music Generation	G	Create tunes by incorporating the provided motif in the specified composition structure. [MUSICAL FORMS T] [MOTIF]
Melody Harmonization	G	Formulate chord combinations to increase the harmonic complexity of the specified musical excerpt. [MELODY]
Bach's Style Music Generation	G	Provide a musical piece that draws inspiration from Bach's compositions.
Motif Extraction	U	Analyze the musical work and pinpoint the consistent melodic element in every section. [MUSIC]
Musical Form Extraction	U	Investigate the attributes of this musical creation and identify its arrangement using suitable music-related terms. [MUSIC]

Table 2: Handcrafted musical tasks in MusicPile, including 6 generation tasks (Type:G) and 2 understanding tasks (Type:U), and provide an example prompt for each task. In the examples, we use tokens in square brackets to represent information other than natural language instruction ([MUSICAL FORM A] represents musical form in alphabets and [MUSICAL FORM T] represents musical form in terminology. [MOTIF], [MUSIC] and [MELODY] are represented in ABC notation. [CHORD] is represented in chord symbols.)

# ChatMusician: Music Generation

X:1  
L:1/8  
M:2/4  
K:F

F/G/ |: "F" BA"C7" GG | "F" FA"C7" G2 | "F" F>G"C7" AB |  
"Am" cA"C7" GF/G/ | "F" BA"C7" GG | "F" FA"C7" G2 | "F" F>G"Bb" Bd |  
1"C7" cE"F" FF/G/ :| 2"C7" cE"F" F z |: "F" f3 (c/d/)(d/e/) |  
"Gm" (e/f/)(f/g/) g>ec | "C7" e/d/ d/c/c/B/ B/A/A/G/ | "F" GA/B/ c/d/e/f/ | f3 (c/d/)(d/e/) |  
"Gm" (e/f/)(f/g/) g>ec | "C7" e/d/ d/c/c/B/ B/A/A/G/ | "F" FA/c/ f z :|

Figure 7: ABC notation and corresponding staff notation of a generated music. Repetition symbols are marked blue in both notations and demonstrate a clear phrase-level repetition. Red and yellow rectangles mark clear motif-level repetition in both sections. Green rectangles mark variation notes following the motif of the first section.

# ChatMusician: Music Generation

- As a score-level music representation, ABC is quite good at compressing data,
- and is text-compatible.

Format	Tokenizer	Tok./Song	Tok./Sec.
ABC	LLaMA Tokenizer	288.21	5.16
MIDI	REMI( <a href="#">Huang and Yang, 2020b</a> )	753.41	12.84
MIDI	MIDI-like( <a href="#">Oore et al., 2018</a> )	728.60	12.42
WAV	EnCodec( <a href="#">Défossez et al., 2022</a> )	12577.46	200.00

Table 4: The average number of tokens per song (Tok./Song) and tokens per second (Tok./Sec) on 1000 songs with different encoding methods. ABC notation achieves the best compression ratio.

# ChatMusician: Music Generation

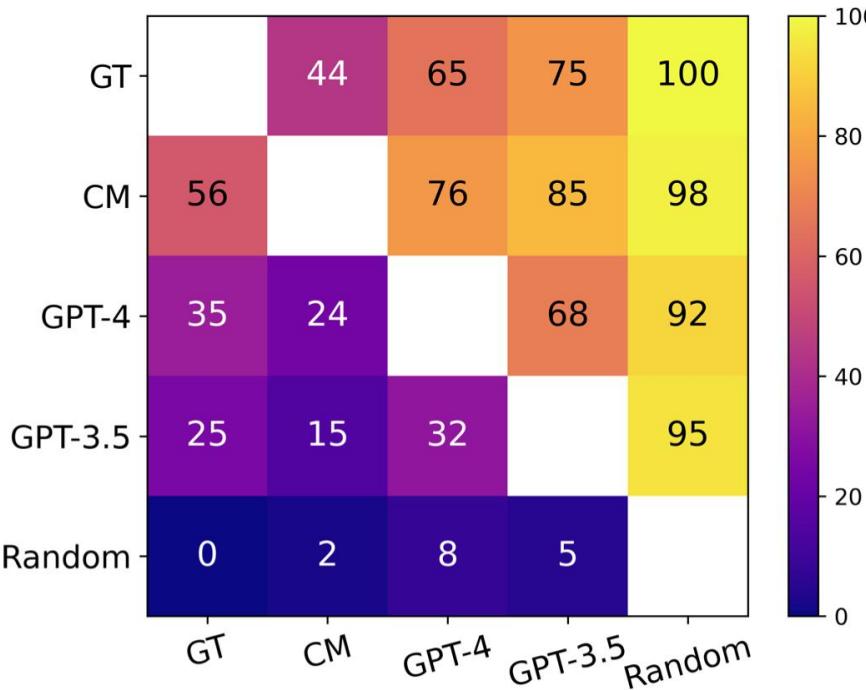


Figure 6: Results from our qualitative study where listeners judged pairs of music come from two different sources. Each row indicates the % of times listeners preferred instrumentals from that system compared to those from each system individually ( $N = 80$ ). ChatMusician is denoted by CM. i.e. 76 means that listeners preferred ChatMusician over GPT-4 in 76% of cases.

# ChatMusician: General Language Ability

System	MMLU Score(%)	Mozart effect?
ChatMusician-Base	<b>48.50</b>	
ChatMusician	46.80	
LLaMA2-7B-Base	46.79	

Table 7: MMLU score of ChatMusicians and LLaMA2-7B-Base.

**It seems that learning music does not harm language ability, even improve it slightly~**

# ChatMusician: Memorization Effect

## 5.4 Memorization Effect of ChatMusician

We analyze the memorization abilities of ChatMusician following (Copet et al., 2023). We randomly select 500 samples from our training set and we feed the model with an instruction prompt. We compare the generated ABC notations with the ground truth. The fraction of examples where the generated and ground truth tokens are identical for the entire sequence is 0.02%. Furthermore, partial matches occur in 0.24% of the training examples, where the generated and ground truth sequences share at least 80% of their tokens.



# YuE: Scaling Open Foundation Models for Long- Form Music (Audio) Generation

# Motivation

- Full song generation with lyrics condition (lyrics2song) is still a challenging problem, only a few close source systems achieve this.
- The challenge mainly lies in
  - the long context nature of music
  - the complexity of the music signal compared to others (speech, sound effects)
  - the distorted linguistic content
  - lack of parallel data

# **Demo Time!**

# Default Mode



YUE

OPEN MUSIC FOUNDATION MODELS FOR FULL-SONG GENERATION

# Voice Cloning

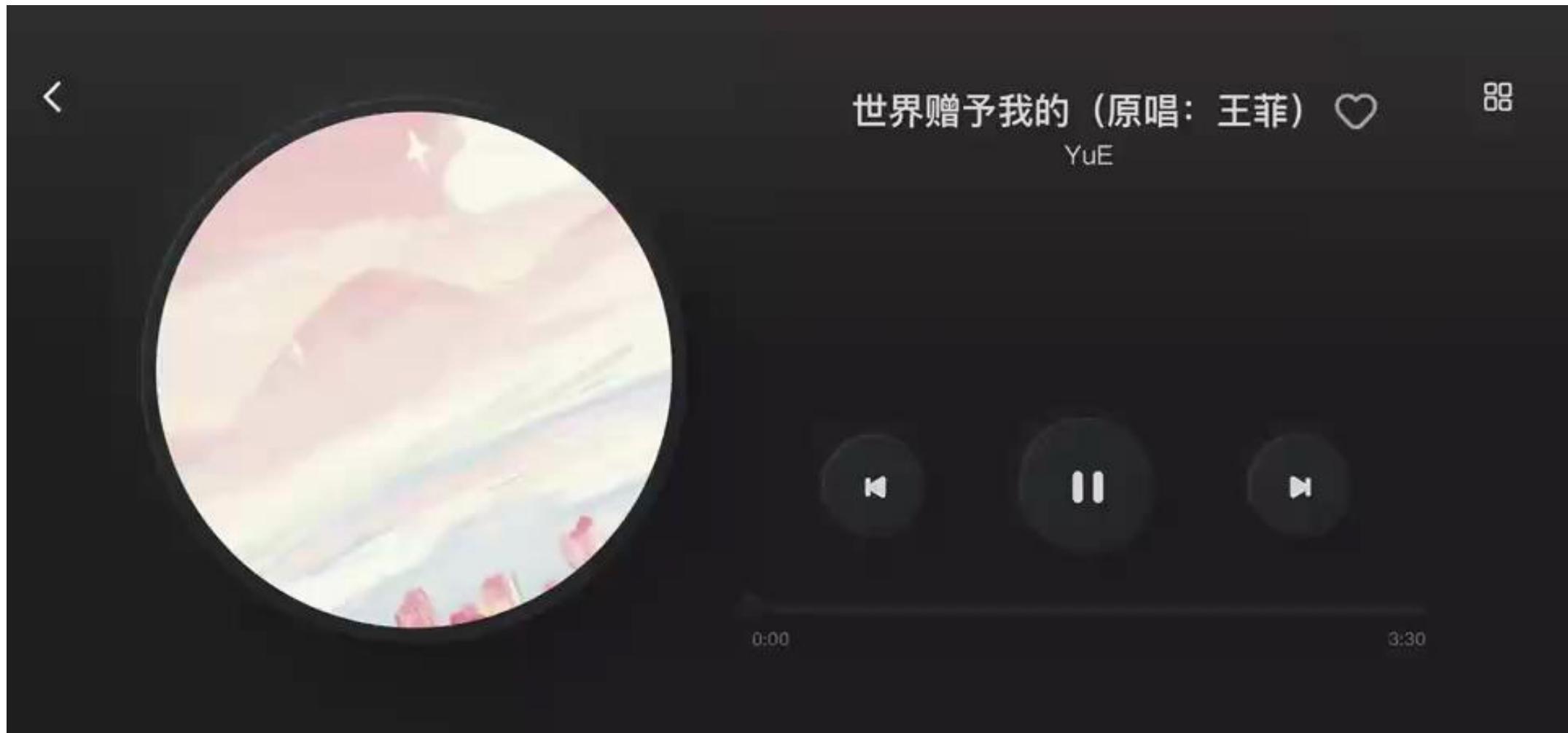


Not So Open

AI

Billie Eilish

# Voice Cloning



# Style Transfer

Genres

Lyrics

[Intro]  
Work it, make it  
Do it, makes us  
Harder, better  
Faster, stronger  
More than, hour  
Hour, never  
Ever, after  
Work is, over  
Work it, make it  
Do it, makes us  
Harder, better  
Faster, stronger

[Chorus]  
Work it harder, make it better  
Do it faster, makes us stronger  
More than ever, hour after hour  
Work is never over

Number of Sequences (paragraphs in Lyrics, the higher this number, the higher the VRAM consumption)

2

1 10

Max new tokens to process in a row (trade off quality / VRAM consumption ?)

3000

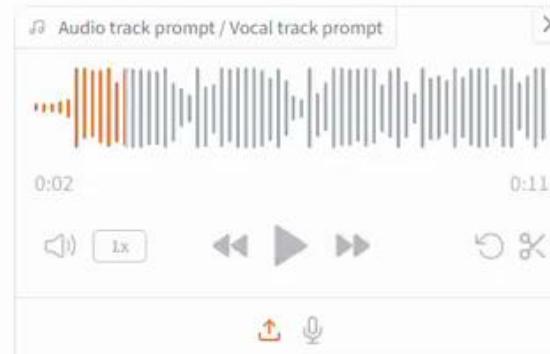
100 5000

Seed (0 for random) 0

0 999999999



Generate



Intrumental track prompt (optional if Vocal track prompt set)

Drop Audio Here

- Or -

Click to Upload

# Emergent Spontaneous Performance & Advanced Vocal Techniques

- 京剧+陕北民歌 
- Scatting (2:20) 
- Death Growl (0:50) 
- Mix Voice 
- Powerful Belt, Riffs and Runs 
- A cappella 
- Harmonica solo improvisation (2:00) 

# **How do we train YuE?**

# Setup

- 5 core developers, dozens contributors on data and eval, some help babysit the pre-training, 1 on vocoder
- (most of the time) small=16xH800s, (some time) medium=96xH800s, (a few times) large=512xH800s
- 700khr data
- <=6 months

# Setup

- 31 exps, hundreds of runs
- scaled up 4 times, only the final run worked
- most exps are at 0.5B scale



# Data

- ~700khr rawdata (music:speech=6:1)
- all data is tagged with qwen2audio, in terms of genre, instrumental, mood, vocal gender, vocal timbre
- 260khr source sep with UVR
- 66khr has matched lyrics on the Internet (in-the-wild)
- structure labels are obtained by all-in-one

# On Data Quality

- Internet data is super noisy
- irrelevant texts, too short lyrics are removed by rule based filter
- manual review and exclude some of the data sources that are too noisy
- manual inspection shows ~80% match rate (most of the content in the audio is reflected in the lyrics transcript)

# Architecture

- 2 LMs, LLaMA2 structure
- stage-1 LM: music language modeling (7B)
- stage-2 LM: residual modeling (2B)
- RVQ tokenizer
- vocoder upsampler

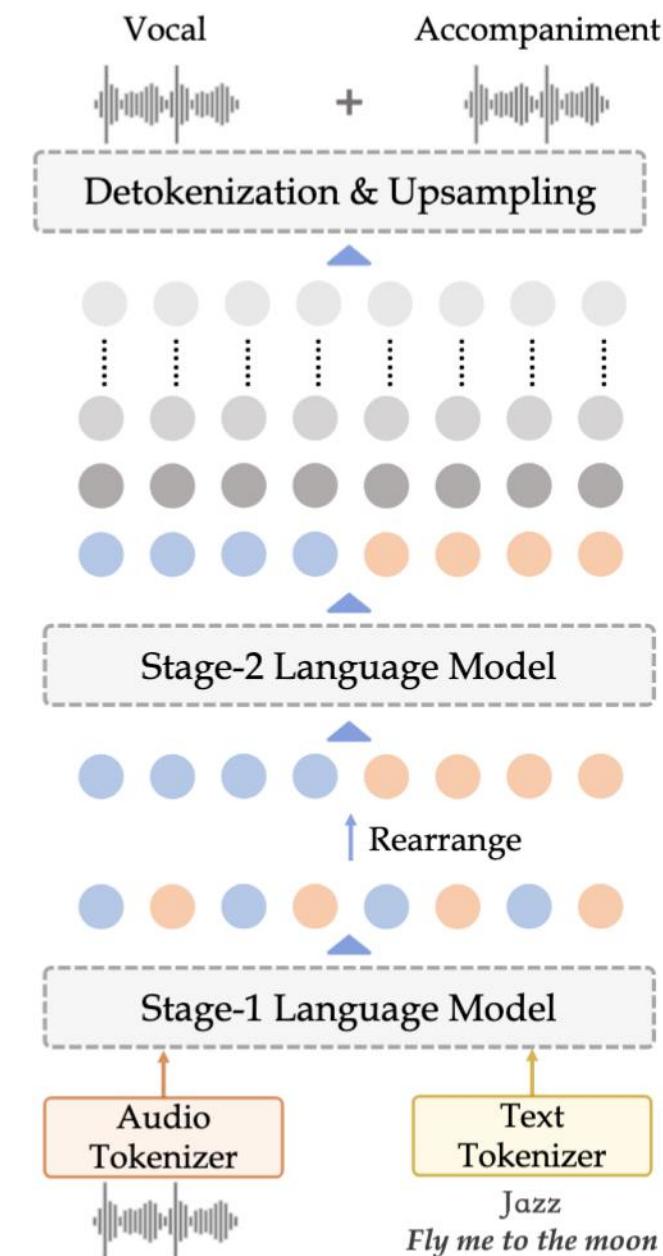


Figure 1. The Overall Framework of YuE.

# YuE: Tokenization

- We tried 4 popular audio tokenizers on our data

Type	Name	Reconstruction	LM Converge	Invalid prob
acoustic	encodec32k	✓	✗	90%+
acoustic	academicodec	✓	✗	90%+
semantic-acoustic fused	semanticodec	✓	✓	>50%
semantic-acoustic fused	xcodec	✓	✓	<10%

# YuE: Tokenization

- Acoustic tokens generally not working
- We even scaled up encodec32k to 1T token 7B size, not working
- BERT based representation is better than MAE based representation
  - probably due to frame level alignment

Type	Name	Reconstruction	LM Converge	Invalid prob
acoustic	encodec32k	✓	✗	90%+
acoustic	academicodec	✓	✗	90%+
semantic-acoustic fused	semanticodec	✓	✓	>50%
semantic-acoustic fused	xcodec	✓	✓	<10%

# YuE: Tokenization

- We use llama tokenizer (bpe) on text, no phonemes
- We introduce some special tokens
- llama tokenizer is expanded to support audio range
- A vocos vocoder is applied to complete 16khz-44.1khz frequency

Table 1. Special tokens and their descriptions.

Token	Description
<EOD>	End of document
<SOA>	Start of audio
<EOA>	End of audio
<stage_1>	Start of Stage 1
<stage_2>	Start of Stage 2
<encodec32k>	Tokenizer type (Encodec 32k)
<xcodec>	Tokenizer type (XCodec)
<semanticcodec>	Tokenizer type (SemanticCodec)
<hificodec>	Tokenizer type (HiFiCodec)

# Stage-1: Music Language Modeling

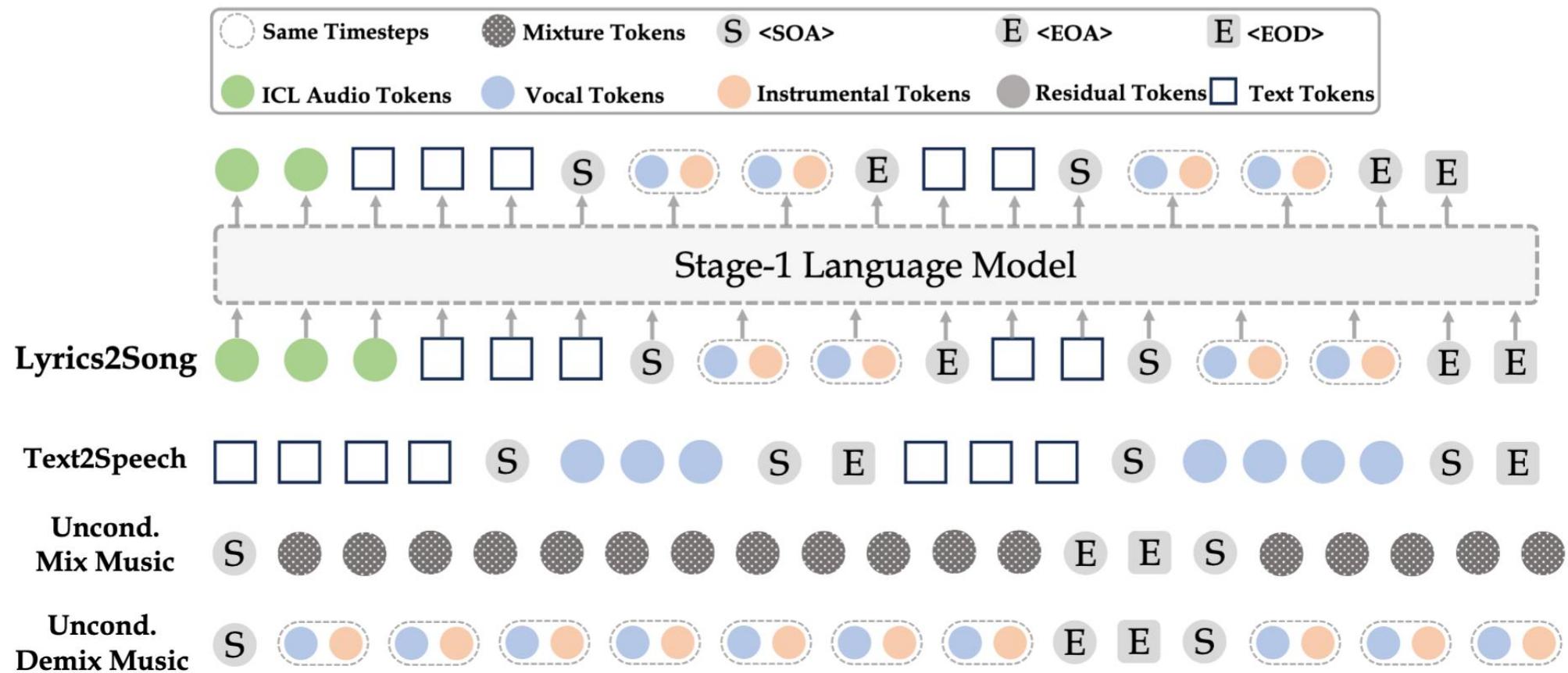


Figure 2. The Stage-1 Framework of YuE.

# Single Token v.s. Source Separate Prior

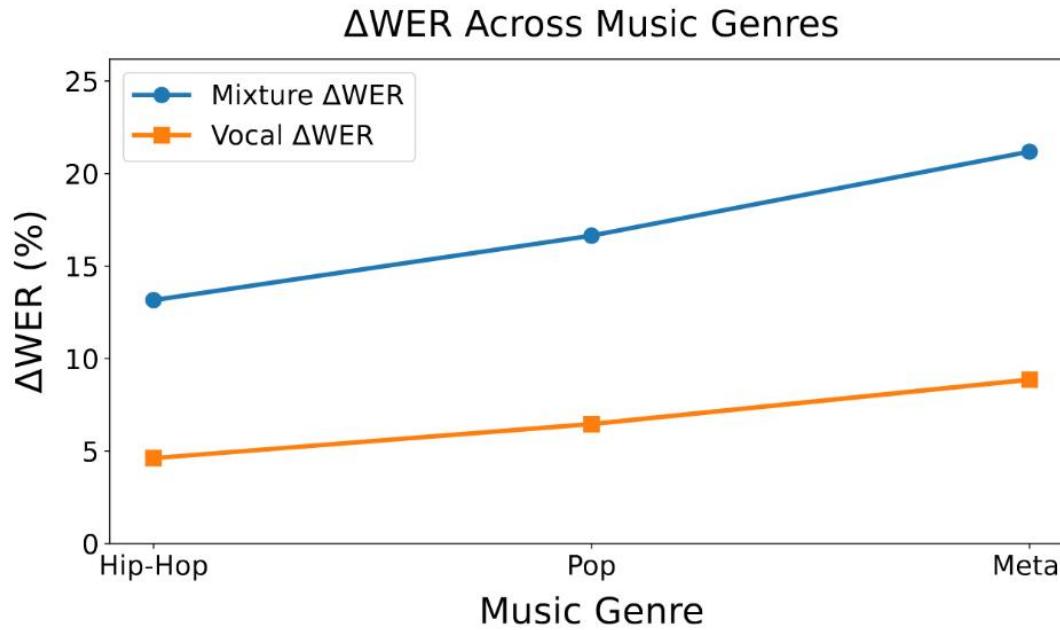


Figure 3. Comparison of  $\Delta\text{WER}$  across different music genres for mixture and vocal-only tracks.  $\Delta\text{WER}$  represents the increase in lyrics WER (obtained by a fine-tuned whisper) after tokenizer reconstruction, defined as  $\Delta\text{WER} = \text{WER}_{\text{recon}} - \text{WER}_{\text{ori}}$ . A higher  $\Delta\text{WER}$  indicates a greater loss of linguistic information. The vocal-only tracks consistently exhibit lower  $\Delta\text{WER}$  across genres, suggesting that source separation is beneficial.

# Dual-NTP

- Use 2 tokens instead of one to model vocal and accompaniment
- Frame-level interleaved
- Can be implemented as standard NTP objective for simplicity

$$\left( \underbrace{v_1}_{\text{vocal}}, \underbrace{a_1}_{\text{accomp.}}, \underbrace{v_2}_{\text{vocal}}, \underbrace{a_2}_{\text{accomp.}}, \dots, \underbrace{v_T}_{\text{vocal}}, \underbrace{a_T}_{\text{accomp.}} \right)$$

# Dual-NTP

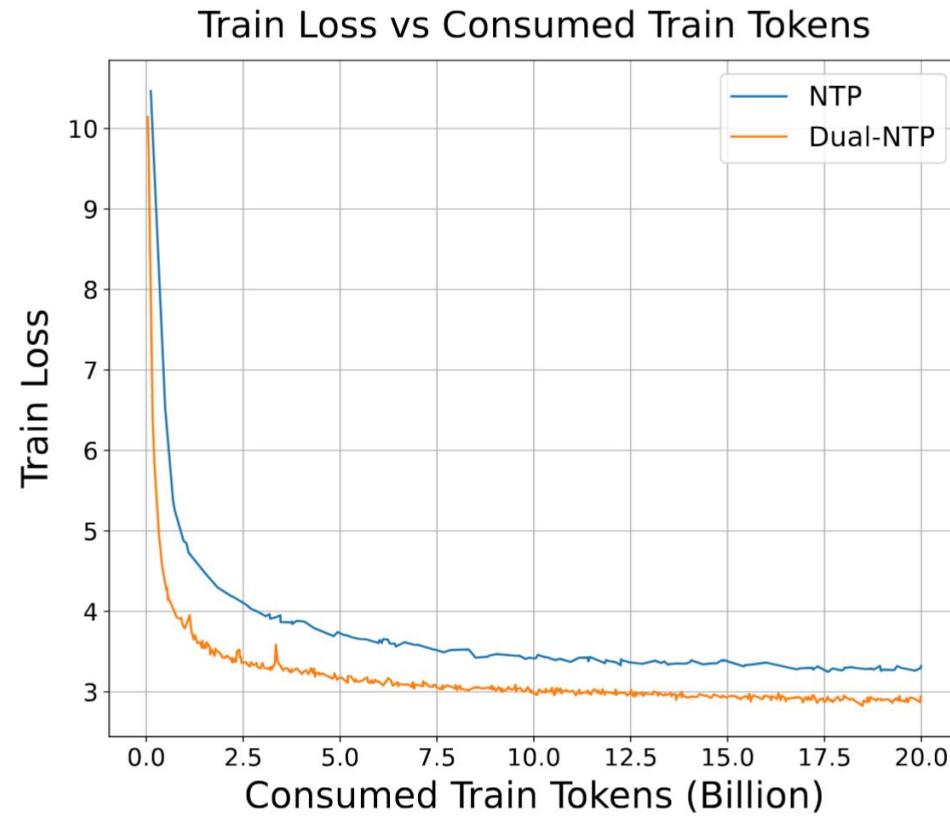


Figure 9. Training Loss over Consumed Train Tokens for NTP and Dual-NTP.

VAR: vocal-to-accompaniment ratio

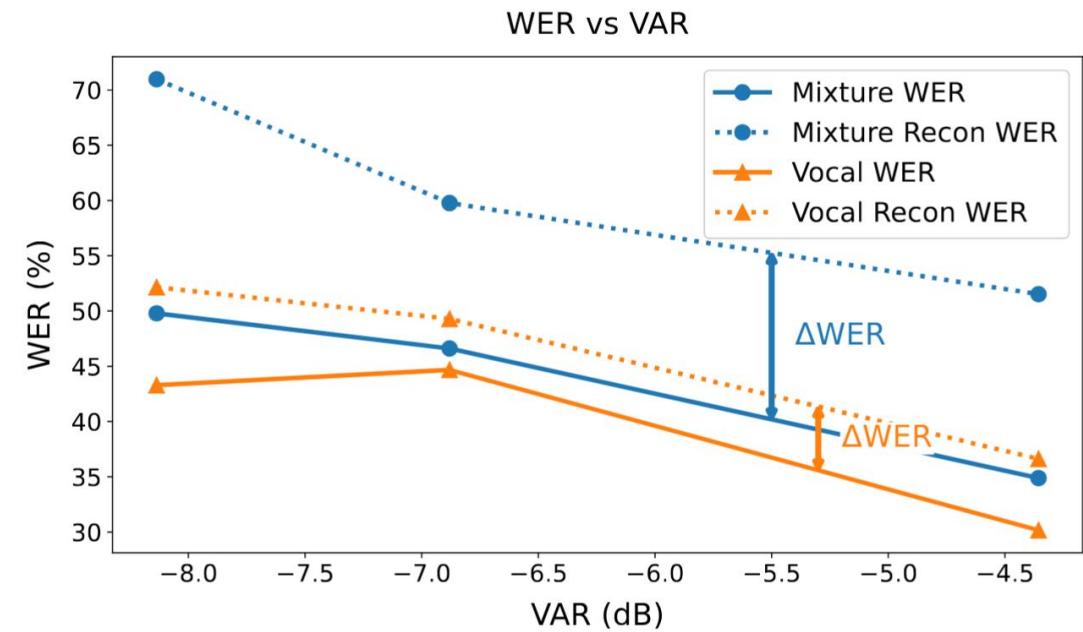


Figure 10. Comparison of WER and VAR for mixture and vocal models, including their reconstructions.

$$\Delta WER = WER_{recon} - WER_{ori}$$

# Challenge of Full-Song Generation

- long-term decay property of RoPE
- vanilla text prepend not working
- ABF to 100k not working
- curriculum learning not working (30s->60s->1min...)

Method	is_working
text prepend	✗
text prepend + ABF	✗
text prepend + cirriculum learning	✗
structural progressive conditioning	✓

# Structural Progressive Conditioning (CoT)

- Inst: Generate music from the given lyrics segment by segment.

$$\mathcal{D}_{\text{cot}} = \underbrace{\text{Inst} \circ \text{Tag} \circ \text{Lyrics}}_{\text{Prompt}} \circ \left( \bigcirc_{i=1}^N s_i \right) \circ \langle \text{EOD} \rangle,$$

Each segment  $s_i$  is structured as

$$s_i = [\text{START\_OF\_SEGMENT}] \circ \tau_i \circ \ell_i \circ \langle \text{SOA} \rangle \circ \psi_i \circ \langle \text{EOA} \rangle \circ [\text{END\_OF\_SEGMENT}],$$

with:

- $\tau_i \in \{[\text{intro}], [\text{verse}], [\text{chorus}], [\text{bridge}], [\text{outro}], \emptyset\}$  an optional segment type token.
- $\ell_i$  representing the segment's lyric content (which may be empty).
- $\psi_i$  denoting auxiliary markers (typically a placeholder such as ...) indicating the context for music generation.

# In-Context Learning

# Speech In-Context Learning

Issues:

- continuation based
- require ref text
- unidirectional
- entanglement

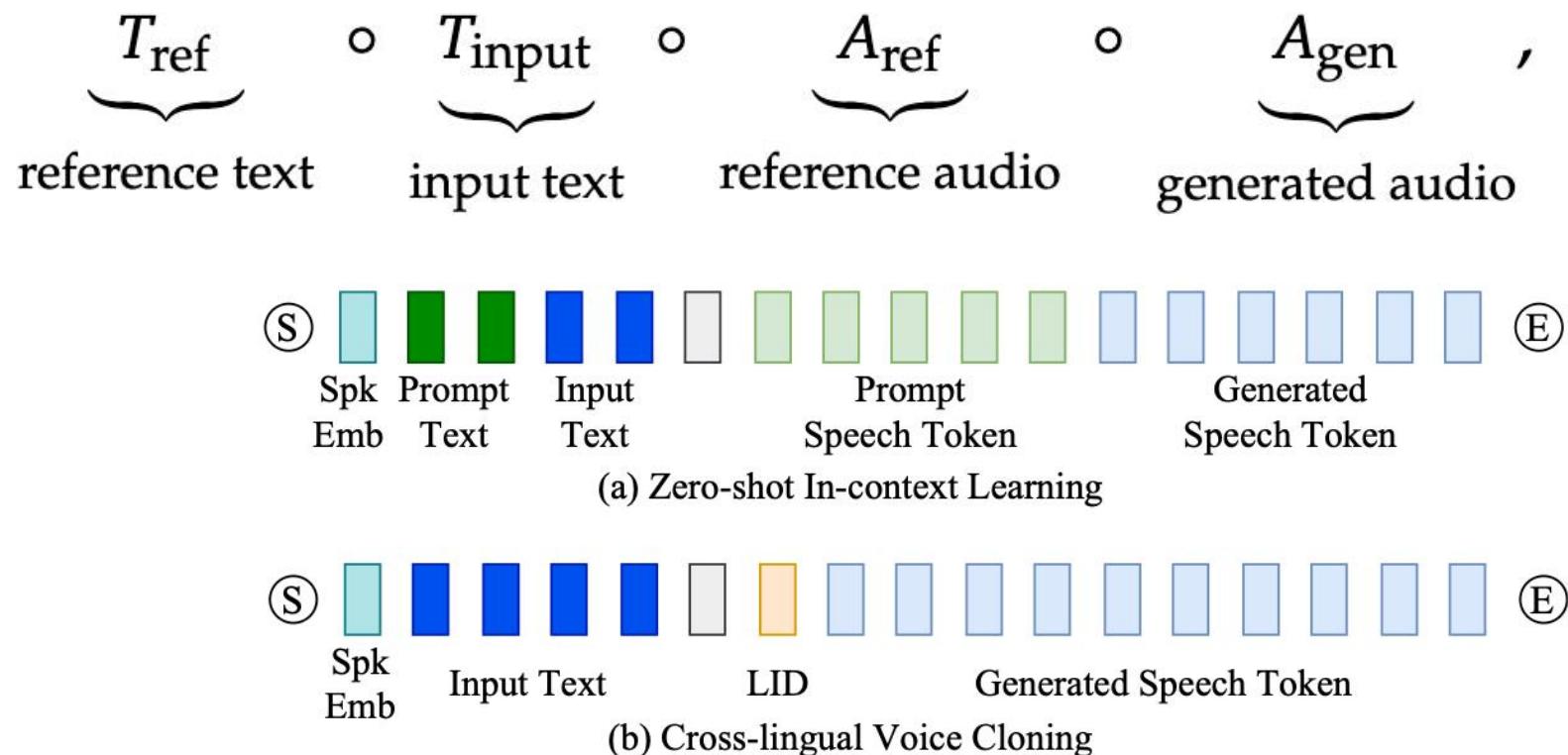
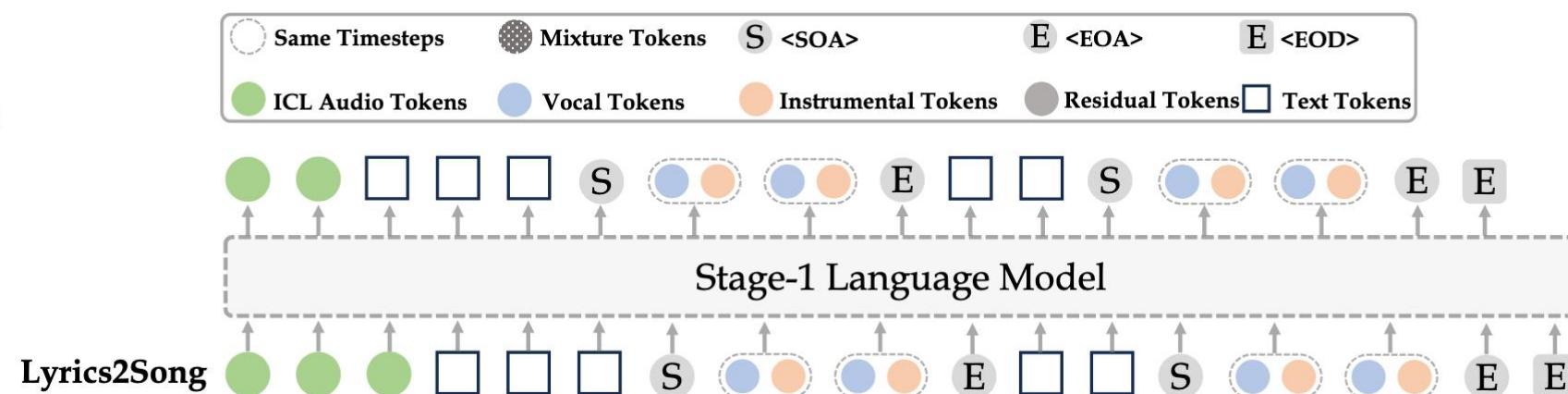


Figure 2: Sequence construction for (a) zero-shot in-context learning and (b) cross-lingual voice cloning. LID represents language identifier. (*cosyvoice*)

# Music In-Context Learning

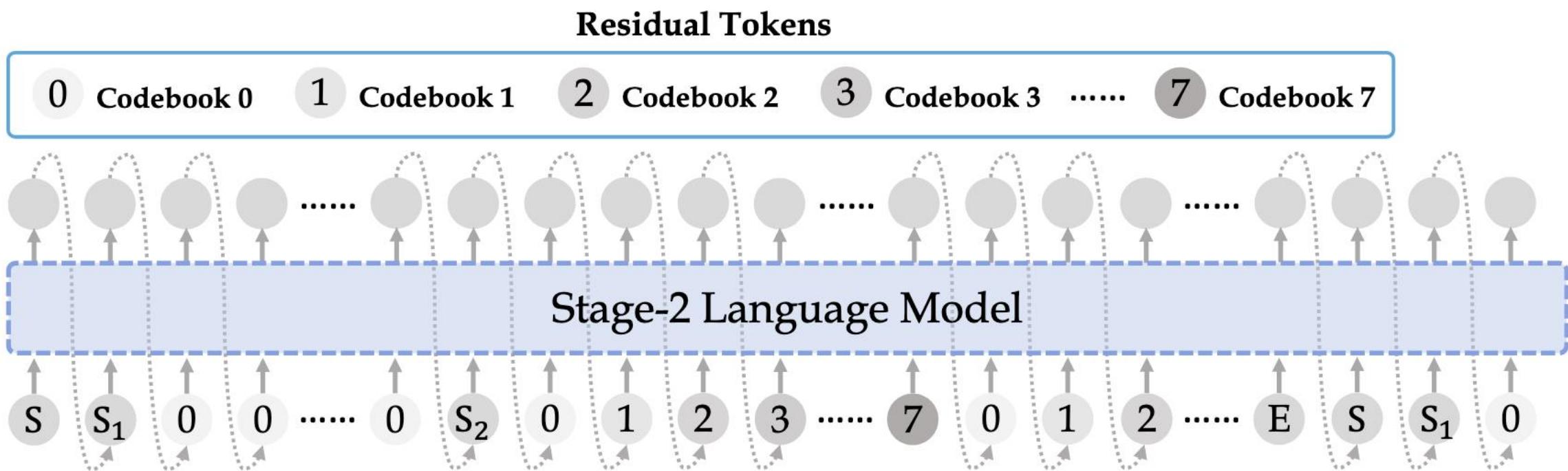
- delayed activation prevent shortcut learning & content entanglement
- ~2% of the total compute can enable this form of ICL
- bidirectional content creation (chorus -> full-song)
- enable voice cloning, music style cloning, style transfer
  - (jp female citypop -> en male rap citypop)

$$\mathcal{D}_{\text{icl}} = A_{\text{ref}} \circ \mathcal{D}_{\text{cot.}}$$



# Stage-2: Residual Modeling

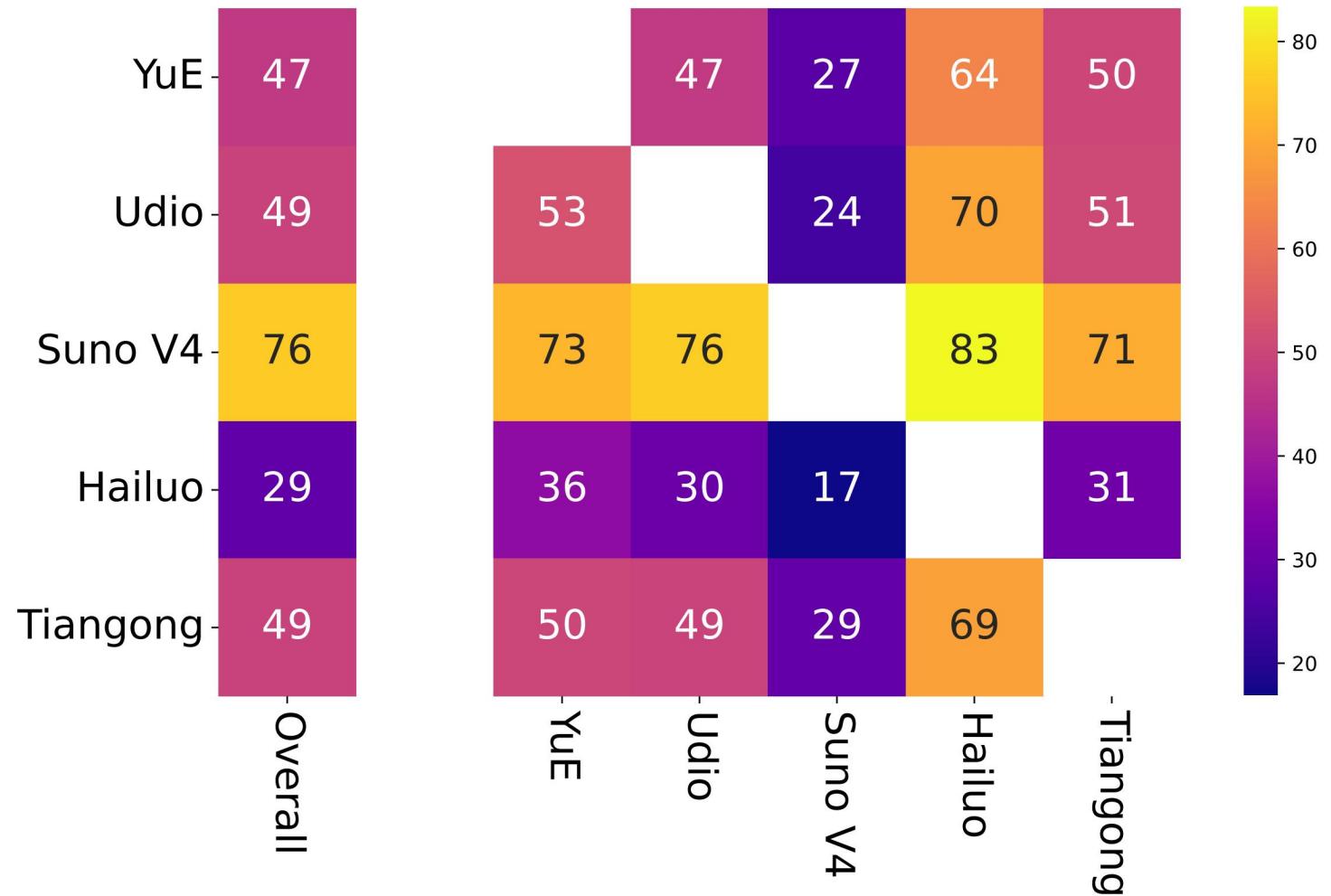
Figure 4. The Stage-2 Framework of YuE.



# Evaluation

# Human Evaluation (EN All Aspects Avg)

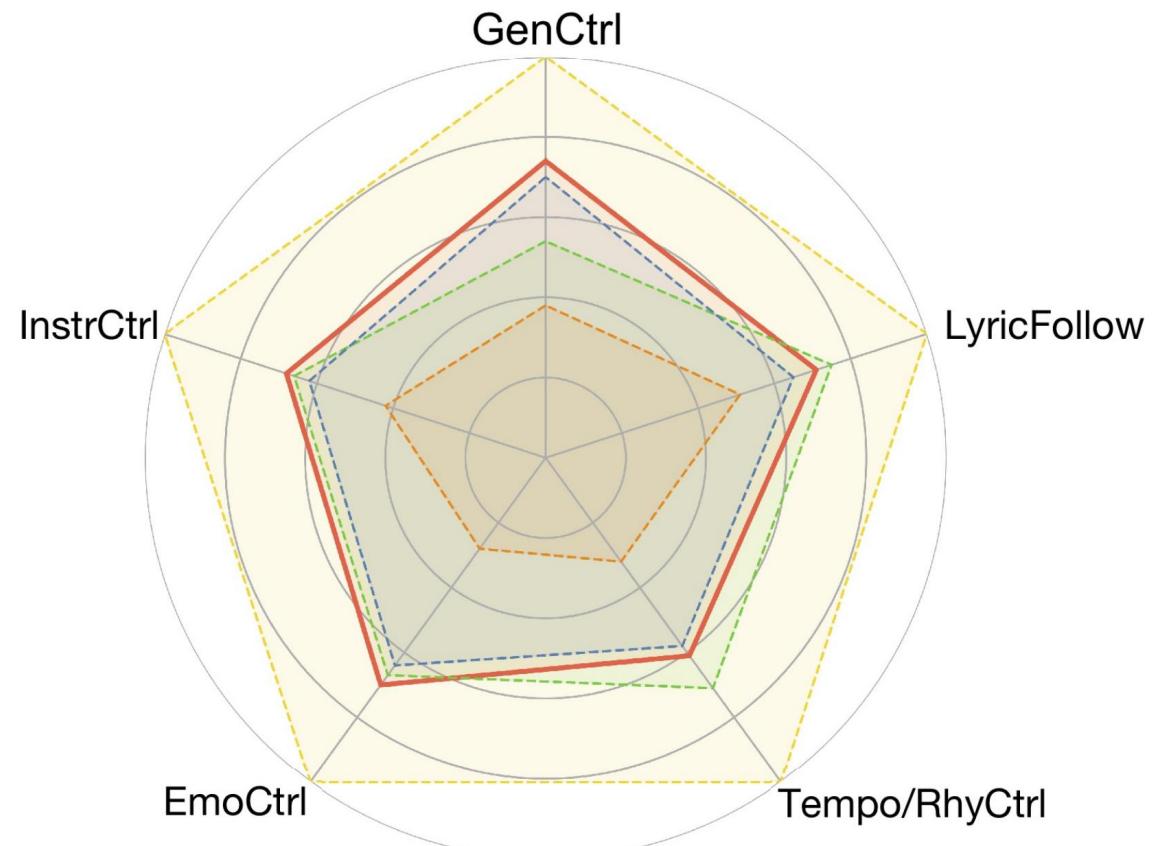
- **YuE** beats one of the closed-source sys, and matches the performance of 2 other sys
- ~40 people participated in the human evaluation, several experts included
- **Suno V4** is SOTA



# Human Evaluation (EN)

- **YuE** shows good control among top tier closed-source sys (ICL enabled)
- **Suno V4** is SOTA

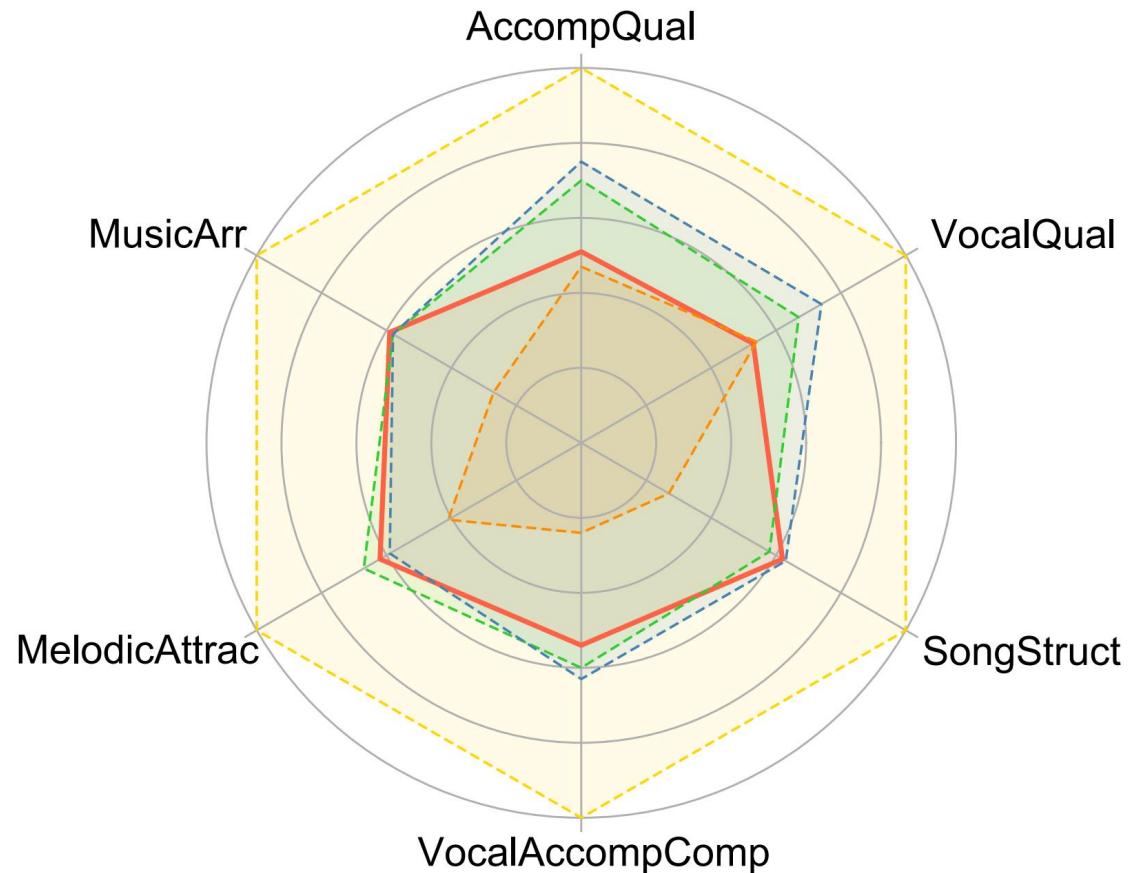
— YuE — Udio — Suno V4 — Hailuo — Tiangong



# Human Evaluation (EN)

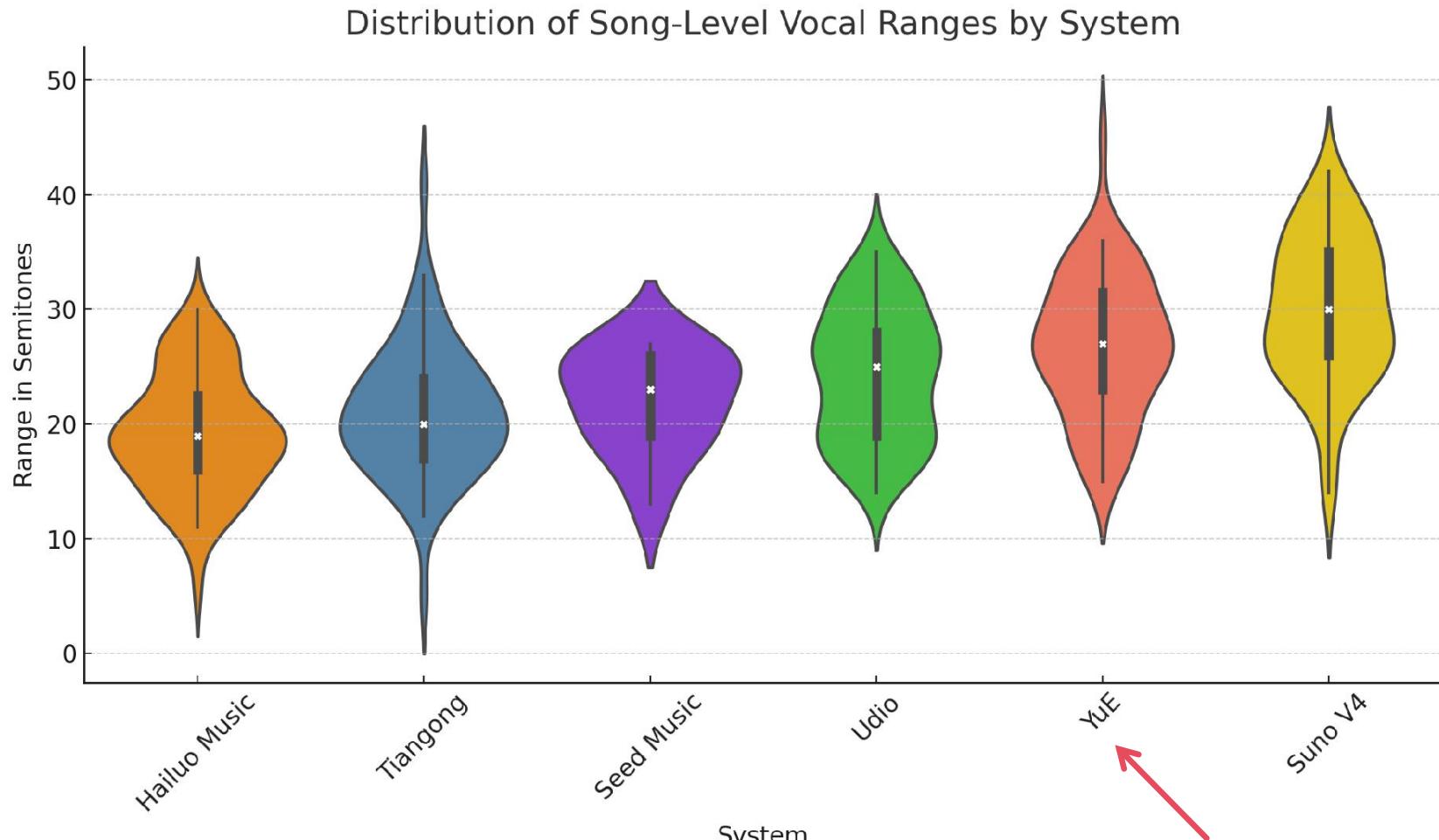
- **YuE** needs to work on audio quality (AccompQual & VocalQual)
- **YuE** has good musicality
- **Suno V4** is SOTA

— YuE — Udio — Suno V4 — Hailuo — Tiangong



# Vocal Agility

RMVPE + 40ms duration filter + manual inspection



`song_vocal_range = highest_note - lowest_note`

The higher the better. Lower value = more static vocal.

among the top tier

# Representation Quality

Table 3. Evaluation of YuE single-track unconditional mode on MARBLE. Including GTZAN genre classification, GS key recognition, MTG top 50 classes tagging, and EMO emotion regression.

Dataset	GTZAN	GS	MTG		EMO	
Task	Genre	Key	Top50		Emotion	
Metrics	Acc↑	Acc <sup>Refined</sup> ↑	AP↑	AUC↑	R2 <sup>V</sup> ↑	R2 <sup>A</sup> ↑
MERT	78.6	65.6	29.9	83.4	61.2	74.7
MusicFM	83.8	63.9	-	-	60.3	<b>76.3</b>
MuQiter	<b>85.6</b>	65.0	-	-	<b>62.8</b>	76.1
YuE	81.7	<b>67.0</b>	29.2	82.7	58.9	75.0

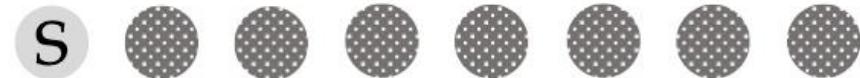
good sense of key

Single track unconditional mode YuE matches or surpasses SOTA SSL.

Note that:

- 1) YuE is not optimized for representation learning
- 2) YuE only uses discrete tokens as input
- 3) YuE is optimized for dual track input

Uncond.  
Mix Music



# Training Recipe

# Failed Scaling Attempts

- 1st: encodec32k -> 7B 1T token failed
- 2nd: unconditional pre-train -> 0.5B, 2B works, 7B failed
- 3rd: ICL pre-train -> copy machine, bad musicality

# The 4th Scaling: Final Recipe

For Stage-1 LM (7B)

- multitask learning
  - cond speech + uncond mix music + uncond demix music + cond music
- data mixture before annealing:
  - cond : uncond = 3:1, music : speech = 10:1
- multiphase training:
  1. warmup 240B token to 3e-4, use EN+ZH data only. 8k seqlen.
  2. constant lr until 1T token, add multilingual data. 8k seqlen.
  3. constant lr until 1.75T token, context length extension -> 16k seqlen
  4. cosine annealing to 3e-5 for another 20B token. Inject ICL(cot:icl=2:1), vocal, (bpm) control.

For Stage-2 LM (2B):

- default mixture, 8k seqlen train on all data for 2T tokens.

# Lessons Learned

# On Scaling & Convergence

- 0.5B LM, 70B token budget is enough for predicting convergence
  - if your model doesn't make valid sound at this point, then it will not converge even if you scale to 1T token 7B size
  - 16xH800 ~18hr budget
- scaling brings lower loss, better compression of data
- scaling  $\propto$  musicality & lyrics following

# On Scaling & Convergence

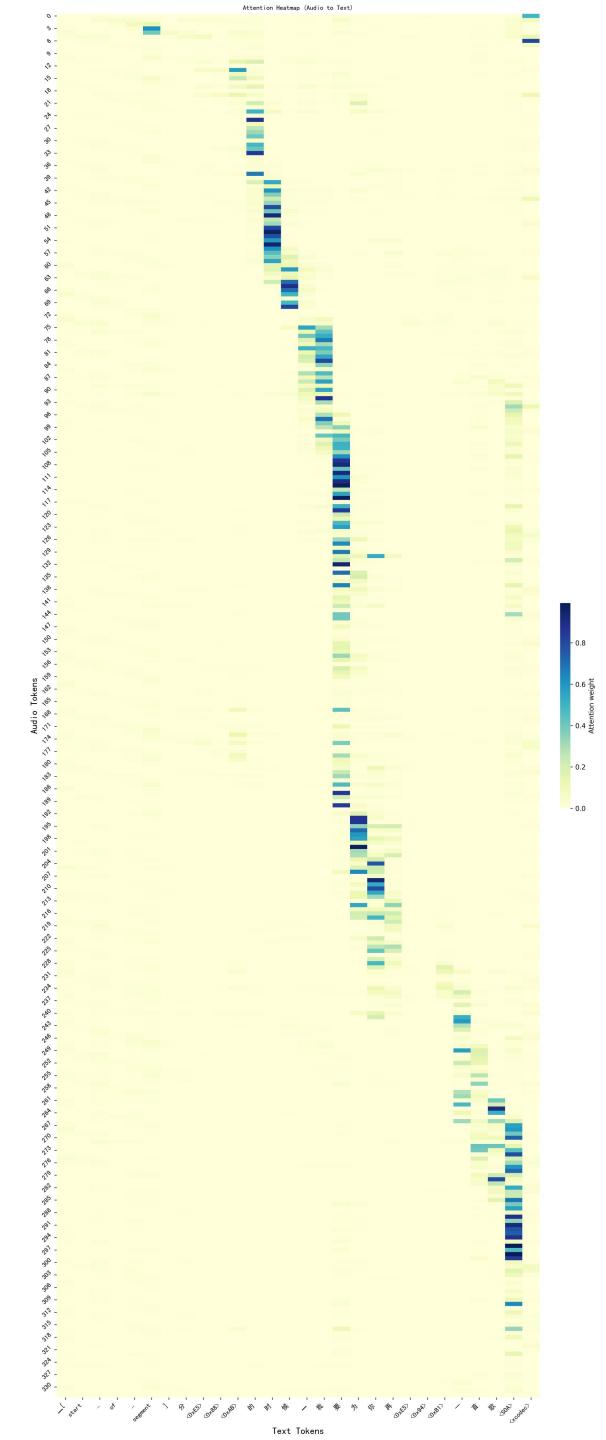
- token efficiency/predictability: semantic >> acoustic
  - linear separability / clusterability
  - consistent token mapping / mutual info between ID and sound events
- the typical loss under 8K seqlen
  - acoustic token LM loss > 4
  - semantic token LM loss < 3

# On Musicality

- lower loss != good music
- scaling  $\propto$  musicality
- unconditional data  $\propto$  musicality
- speech  $\propto$  **-musicality**
  - too much speech results in a rap machine, read instead of sing
- ICL harms musicality/musicality
  - introduce ICL too early = copy machine

# On Lyrics-Following

- Noisy in-the-wild music data also works
- TTS data is the key
- speech data  $\propto$  lyrics following
- unconditional data  $\propto$ -lyrics following
  - unconditional pretraining does not work for 7B model
  - larger model tends to overfit to dominant learning signal
- neural structure: learned lyrics head



# On Controllability

- CFG improves good case rate
- the strongest control signal = audio condition (ICL)
  - 5s audio = genre, emo, instrument, vocal, bpm, chords, melody, etc.
  - strong control = shortcut != creativity & skill
- using chorus as ICL ref audio = better musicality/stability

# On Automatic Evaluation

**NOT ALIGN WITH HUMAN PERCEPTION!**

Table 3. Comparison of various music generation models across multiple metrics.

Model	KL↓	FAD↓	CE↑	CU↑	PC↑	PQ↑	CLAP↑	CLaMP↑
Hailuo	0.756	2.080	7.350	7.737	6.793	8.132	0.181	0.106
Seed	0.734	4.245	7.586	7.921	5.461	8.309	0.160	0.122
SunoV4	0.960	1.64	6.958	7.646	6.001	7.950	-0.014	0.160
Tiangong	0.708	2.547	7.421	7.766	6.060	8.220	0.088	0.114
Udio	0.541	1.238	7.045	7.502	6.541	7.784	0.095	0.156
YuE	0.372	1.624	7.115	7.543	6.280	7.894	0.118	0.171

CE, CU, PC, PQ: from audiobox-aesthetic  
CLAP not working well with song?  
CLaMP3 seems better.

Thanks For All the Support/Collaboration Along the Way

- MAP, HKUST, HKGAI, moonshot.ai, 01.ai, bytedance, geely, 智源, Qwen, 昆仑万维...

# Discussion

