



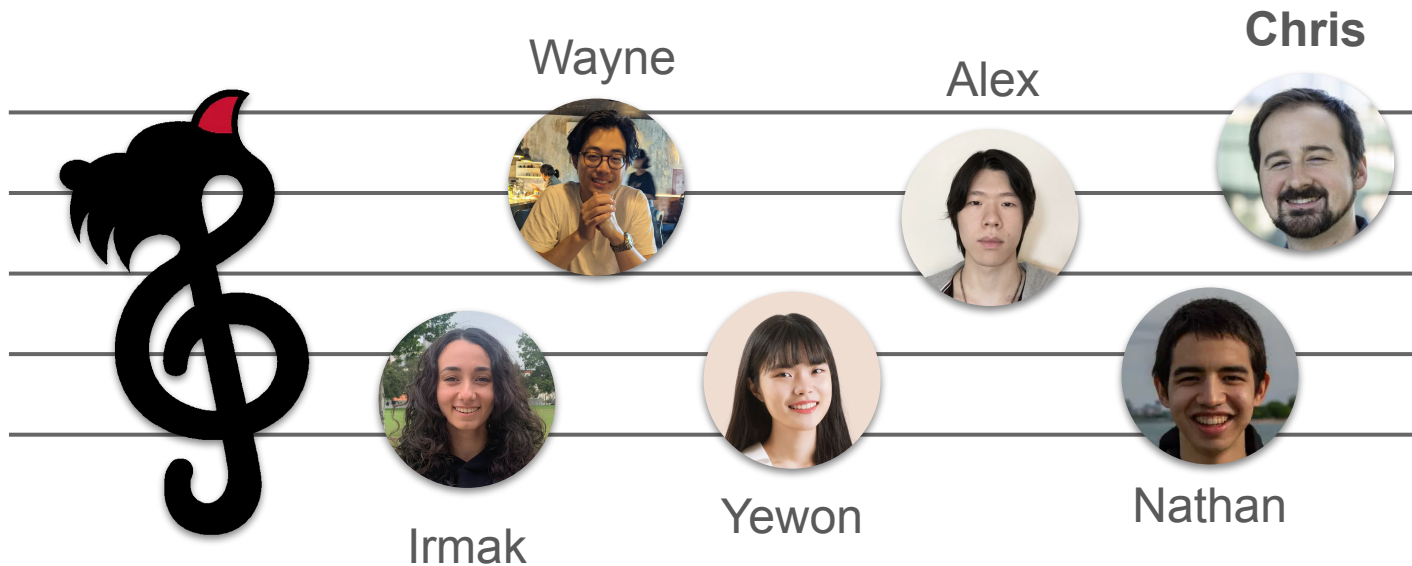
Music 🤝 LLMs

Chris Donahue

Dannenberg Assistant Professor, CMU CSD
Research scientist, Google DeepMind / Magenta
CMU 11-667 LLMs, 26 Apr 02



Generative Creativity Lab (G-CLef), CMU CSD





Understand and shape the role of ✨ AI
in 🎵 *human creativity and productivity*



Creativity is a *process*,
not just a *product*

We often focus on music,
but the goals are broader

Modern demos / outline

What can music LMs do in 2026?



Text-to-music

"lo fi beat, chill vocals about
natural language processing"



**As you listen: have you
ever deliberately
listened to AI-generated
music? What do you
feel listening to this
example?**

What can music LMs do in 2026?



Text-to-music

"lo fi beat, chill vocals about
natural language processing"



**As you listen: have you
ever deliberately
listened to AI-generated
music? What do you
feel listening to this
example?**

[Instrumental intro]

[Verse]

Data whispers like a gentle stream
Words unweave into a lucid dream
Syntax bends where the thoughts collide
Meaning flows on the other side

[Chorus]

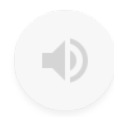
Speak to me in code / In light, in fire
Every word a spark / A line, a wire
Syntax and soul / They intertwine
Natural language / Yours and mine

What can music LMs do in 2026?



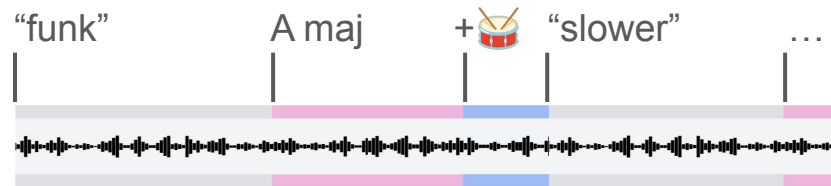
Text-to-music

"lo fi beat, chill vocals about
natural language processing"



Real-time gen.

LM

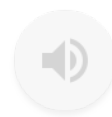


What can music LMs do in 2026?



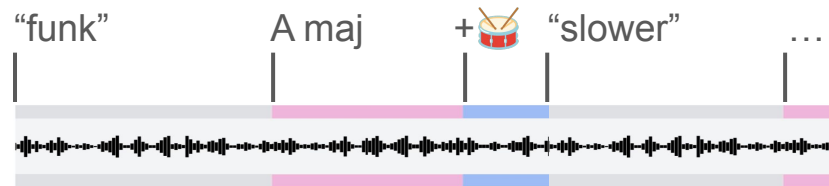
Text-to-music

"lo fi beat, chill vocals about
natural language processing"



Real-time gen.

LM

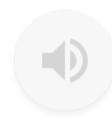


What can music LMs do in 2026?



Text-to-music

“lo fi beat, chill vocals about natural language processing”



Real-time gen.

LM



Creative co-pilots,
music programs



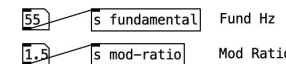
LM



“FM synth in Pd”



LM



What can music LMs do in 2026?

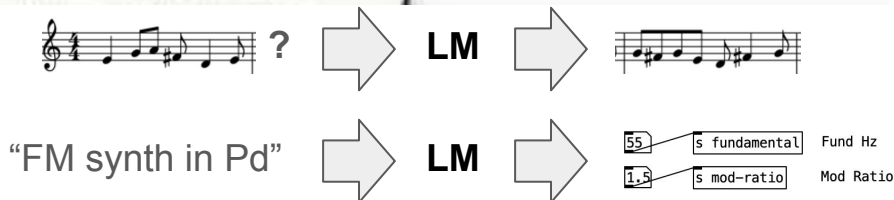
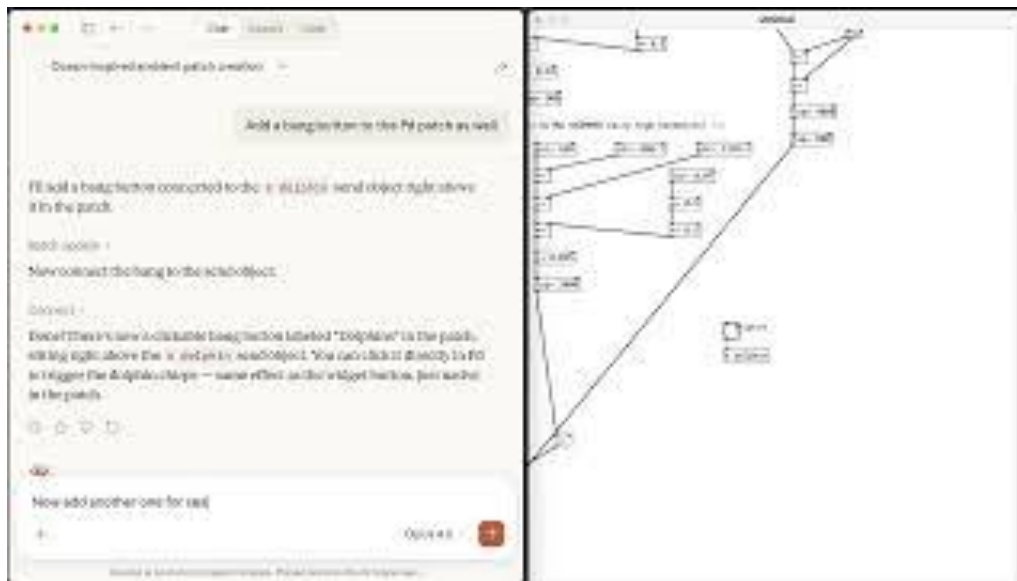
Text-to-music



Real-time gen.



Creative co-pilots, music programs



What can music LMs do in 2026?



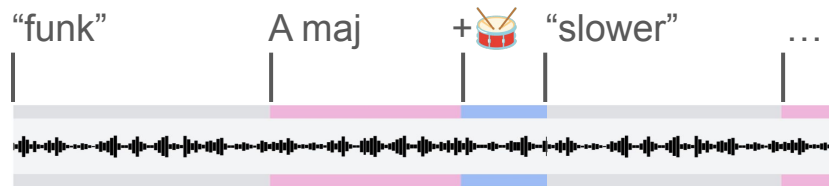
Text-to-music

“lo fi beat, chill vocals about natural language processing”



Real-time gen.

LM



Creative co-pilots,
music programs



LM



“FM synth in Pd”



LM



55
1.5

s fundamental
s mod-ratio

Fund Hz
Mod Ratio

Overview

- Symbolic music LMs methods and history
- Modern music audio LMs
- Research frontiers in music LMs
- Today is a crash course! If you're interested:
 - Take 15-798 Generative AI for Music and Audio, planned Spring 2027
 - Join our music AI reading group (send me an email)
 - Learn more at gclef-cmu.org
 - Email me chrisdonahue@cmu.edu

Basics, history, symbolic

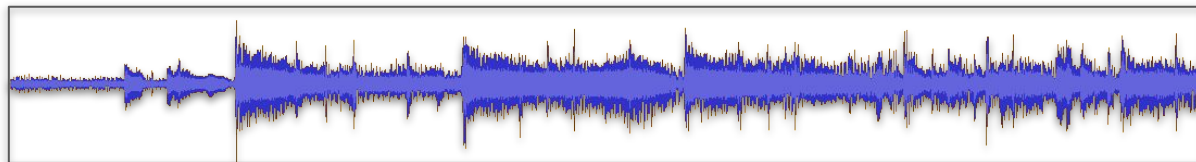
Overview

- **Symbolic music LMs methods and history**
- Modern music audio LMs
- Research frontiers in music LMs
- Today is a crash course! If you're interested:
 - Take 15-798 Generative AI for Music and Audio, planned Spring 2027
 - Join our music AI reading group (send me an email)
 - Learn more at gclef-cmu.org
 - Email me chrisdonahue@cmu.edu

Music is inherently multimodal

Musical *sound* (acoustic, continuous, dense)

 **Audio**
(10 MB)



may depict an underlying **musical *score*** (discrete, symbolic, sparse)

 **Score**
(1 kB)

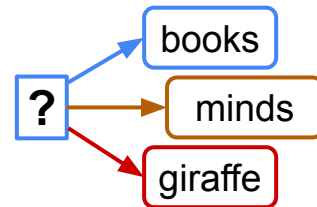


Do you know how to read sheet music? Y/N

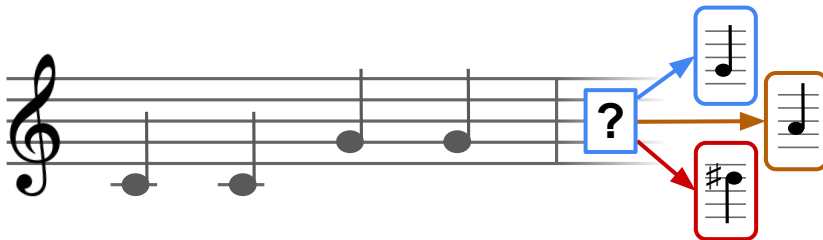
Language models for music

From *language*

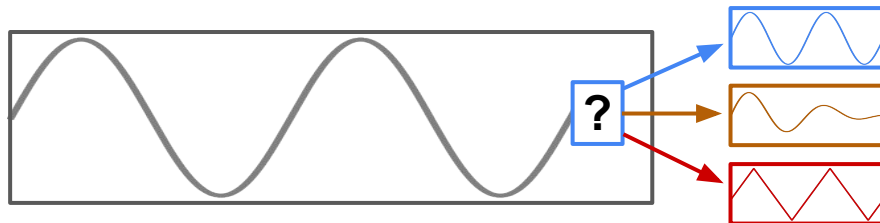
The students opened their



to music scores

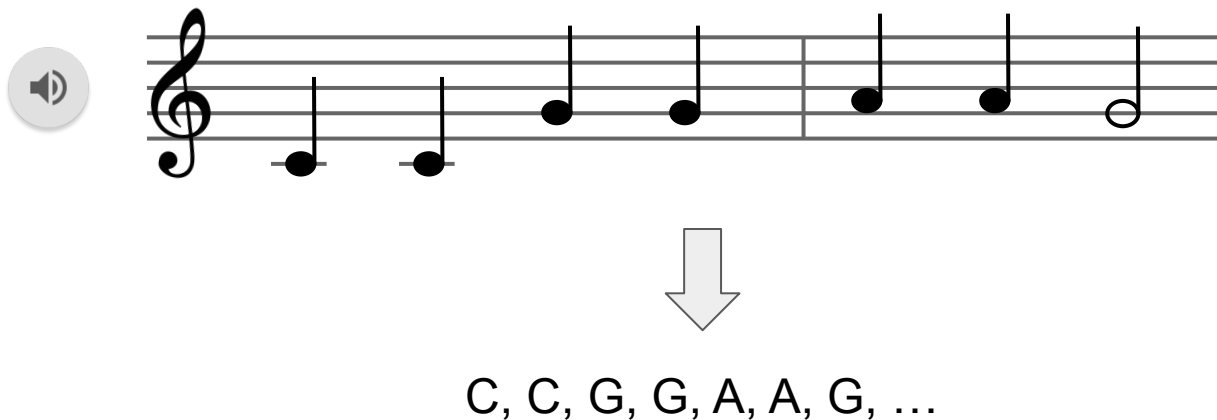


and music *audio*!



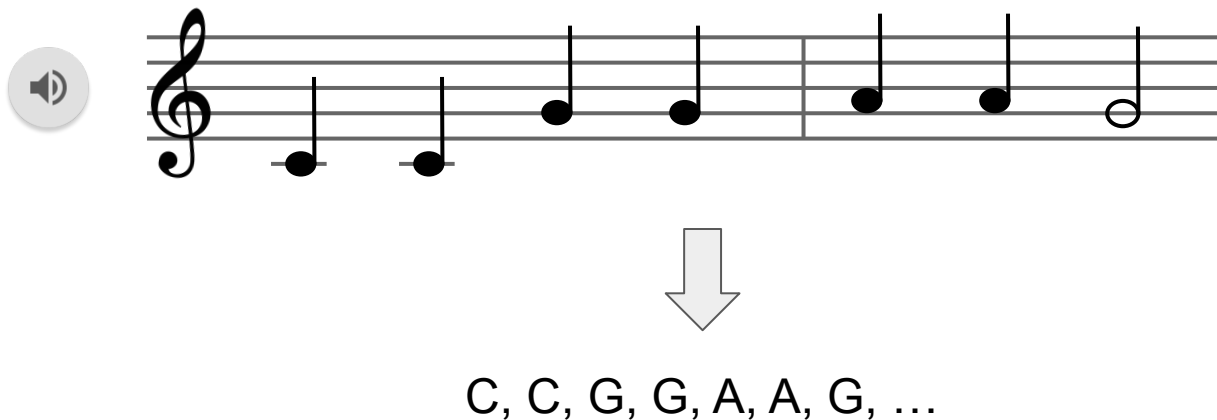
🤔 Scores look a lot like language!

We can *tokenize* in a manner similar to language



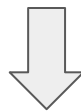
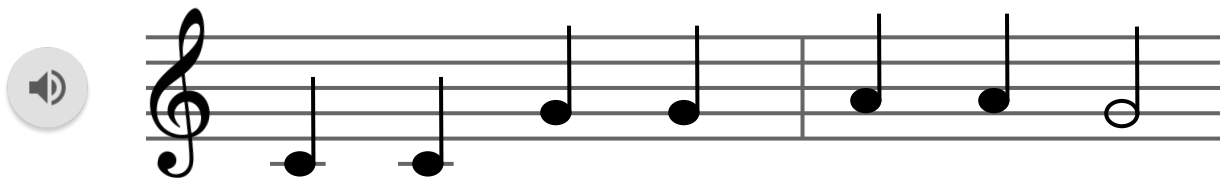
🤔 Scores look a lot like language!

We can *tokenize* in a manner similar to language



Designing tokenizers to emphasize musical attributes

Just like with text, there are different ways to tokenize music



Less faithful



More faithful

Pitch classes: C, C, G, G, A, A, G

With *time*: ♩C, ♩C, ♩G, ♩G, ♩A, ♩A, ♩G

With octaves: ♩C4, ♩C4, ♩G4, ♩G4, ♩A4, ♩A4, ♩G4

Small vocab



Large vocab

Tokenization schemes can preserve different musical info

Just like with text, there are different ways to tokenize music

We can also manipulate both the vocabulary $\mathbb{V}$.
Trade off sequence length and vocab size.

$$\mathbb{V} = \{\text{♩C}, \text{♩G}, \dots\} \qquad \mathbb{V} \neq \{\text{♩}, \text{♩}, \text{♩C}, \text{♩G}, \dots\}$$

More simple

Pitch classes: C, C, G, G, A, A, G

Small vocab



More complex

With *time*: ♩C, ♩C, ♩G, ♩G, ♩A, ♩A, ♩G ♩

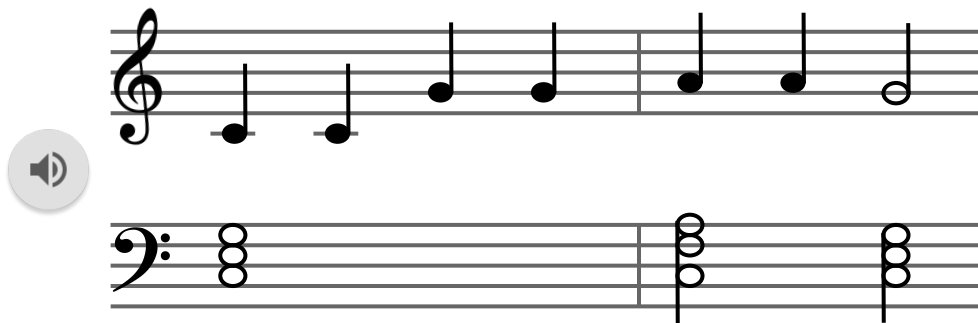


Large vocab

With octaves: ♩C4, ♩C4, ♩G4, ♩G4, ♩A4, ♩A4, ♩G4 ♩

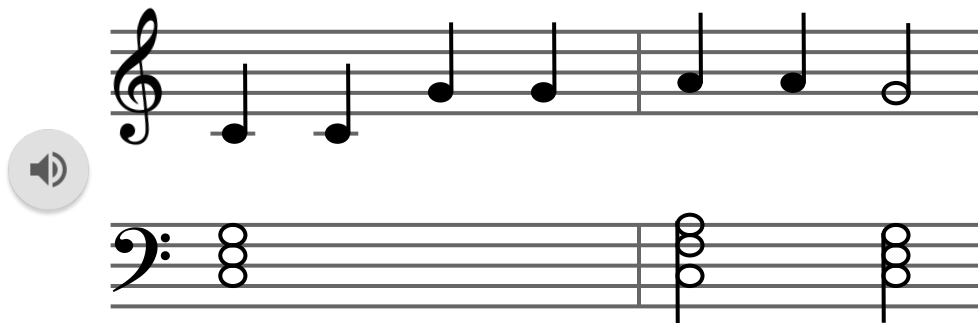
Domain-specific challenge: *simultaneity*

Unlike in text, multiple symbolic music tokens can coincide in time



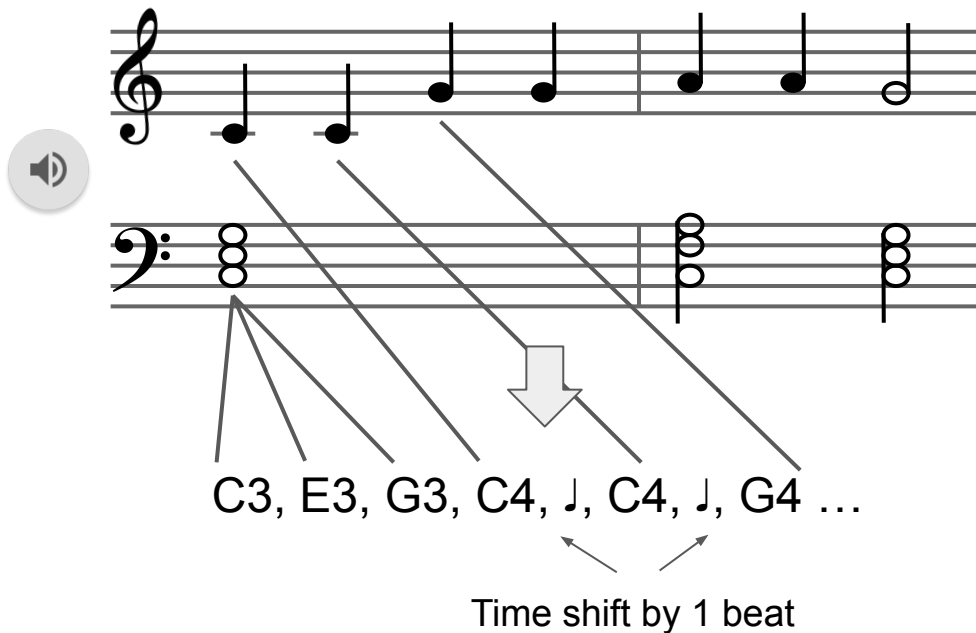
Domain-specific challenge: *simultaneity*

Unlike in text, multiple symbolic music tokens can coincide in time



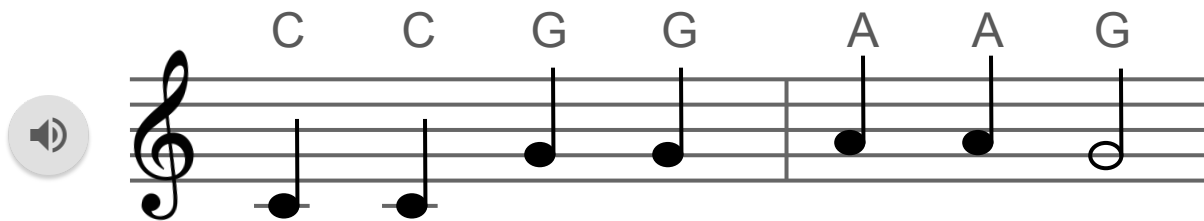
Domain-specific challenge: *simultaneity*

Unlike in text, multiple symbolic music tokens can coincide in time. One approach is to insert “time shift” tokens to encode changes in time [Oore and Simon 17]



A probabilistic view on music

Music can be interpreted as the result of a stochastic process. Example:



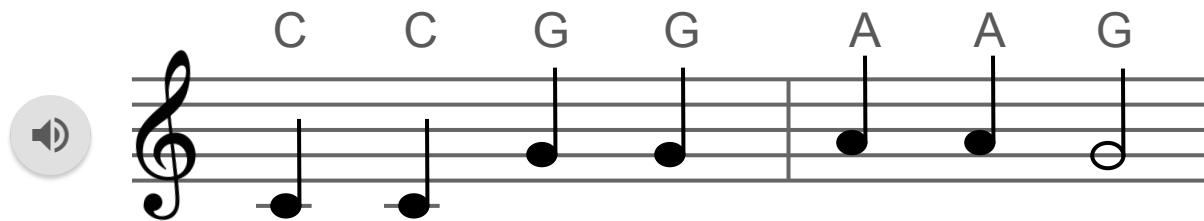
Roll a 🎲 seven times. Map outcomes $\{1, 2, 3, 4, 5, 6\}$ to pitches $\{C, D, E, F, G, A\}$



What is the likelihood that we get Twinkle Twinkle here?

A *probabilistic* view on music

Music can be interpreted as the result of a stochastic process. Example:



Roll a 🎲 seven times. Map outcomes $\{1, 2, 3, 4, 5, 6\}$ to pitches $\{C, D, E, F, G, A\}$

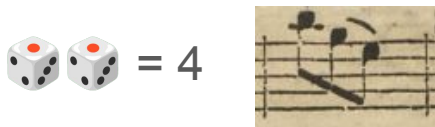
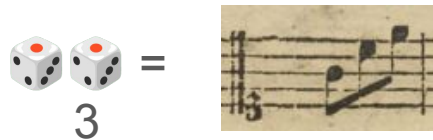
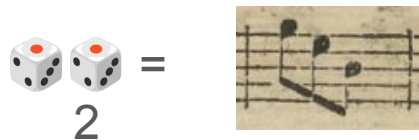
🤔 What is the *likelihood* that we get Twinkle Twinkle Little Star?

Extremely low. $\frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6}$, or 1 in 279936

History: Probability in classical music ([demo](#))

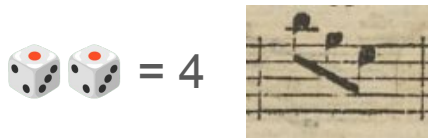
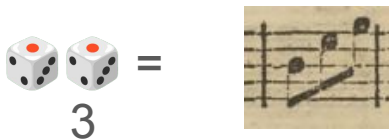
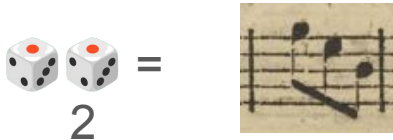
Classical example: *Musikalisches Würfelspiel* (“Musical dice game”), 18th century

Measure 1



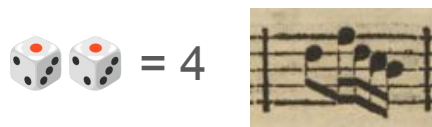
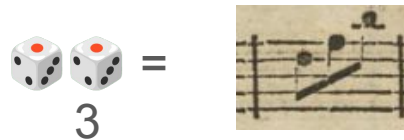
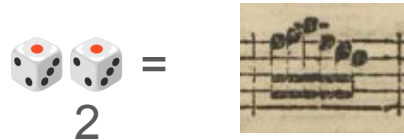
...

Measure 2



...

Measure 3

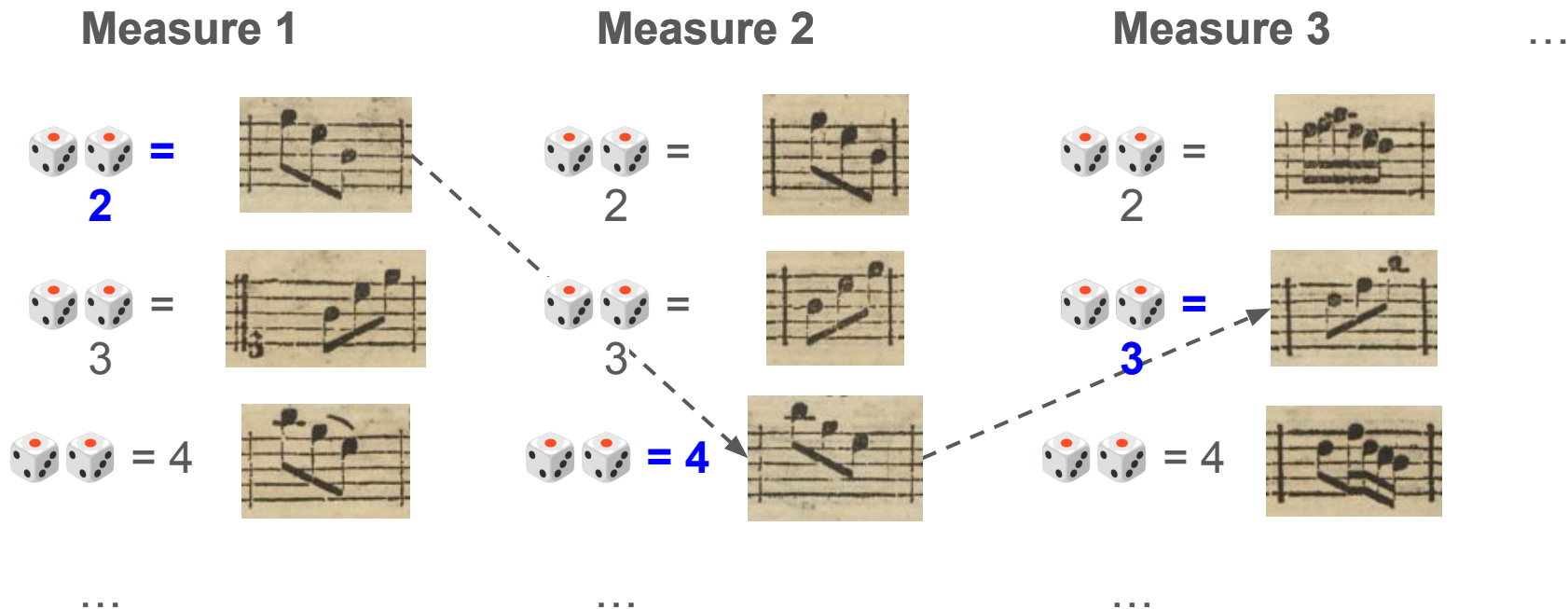


...

...

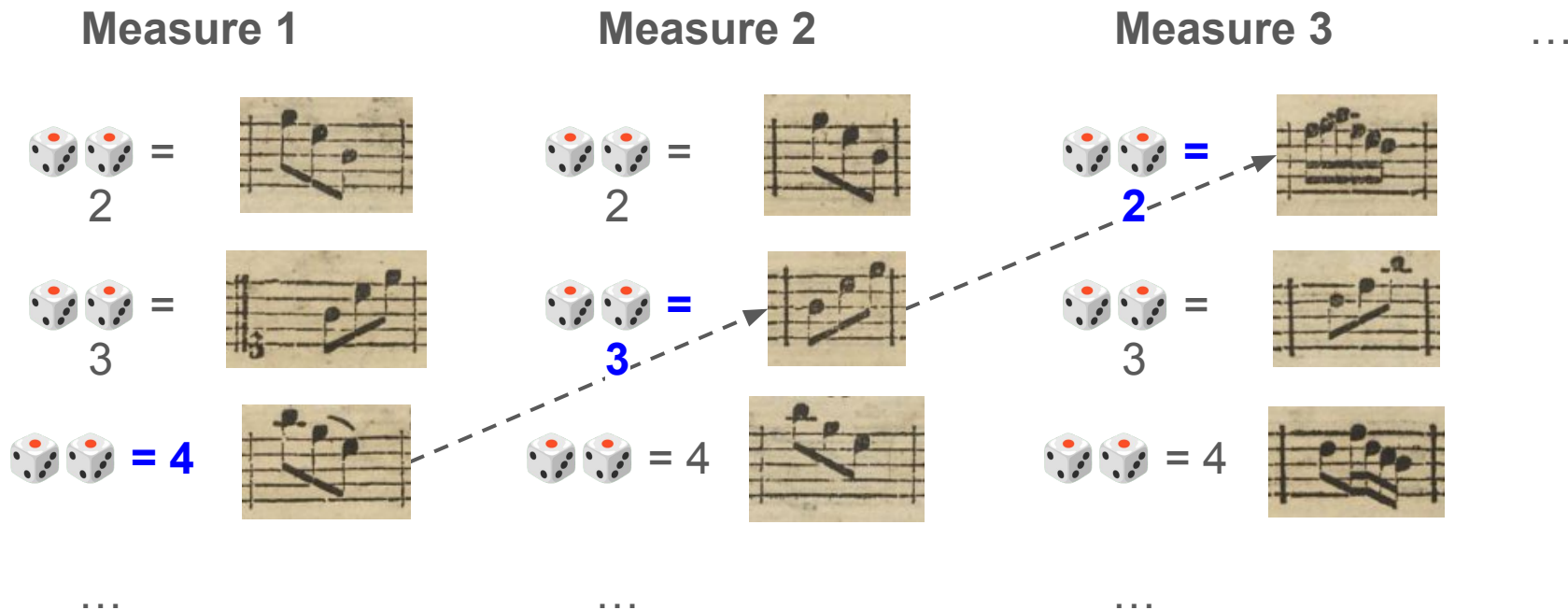
History: Probability in classical music ([demo](#))

Classical example: *Musikalisches Würfelspiel* (“Musical dice game”), 18th century



History: Probability in classical music ([demo](#))

Classical example: *Musikalisches Würfelspiel* (“Musical dice game”), 18th century



History: Probability in 20th century music

Algorithmic composition methods emphasized Markov chains (bigram models)

$$P(n_i | n_{i-1}) =$$

	C	D	E	F	G	A	B
C							
D							
E							
F							
G							
A							
B							

History: Probability in 20th century music

Earlier on, Markov chain probabilities were hand-tuned to encode expertise

$$P(n_i | n_{i-1}) =$$

	C	D	E	F	G	A	B
C	.00	.35	.20	.15	.15	.10	.05
D	.30	.00	.30	.15	.10	.10	.05
E	.15	.30	.00	.30	.10	.10	.05
F	.10	.10	.30	.00	.30	.10	.10
G	.20	.05	.10	.25	.00	.30	.10
A	.10	.05	.05	.15	.30	.00	.35
B	.45	.05	.05	.05	.05	.35	.00

ILLIAC Suite (1957)
Lejaren Hiller /
Leonard Isaacson

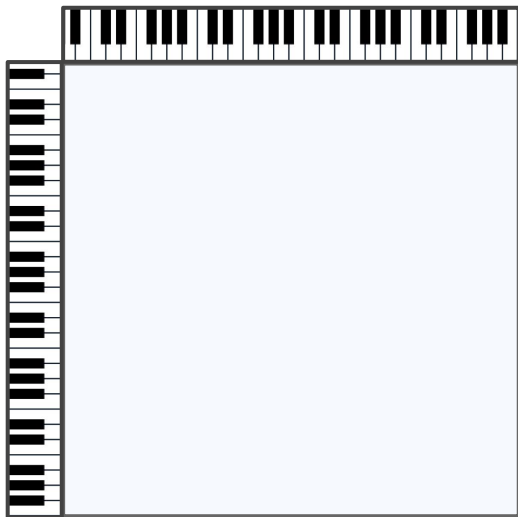


IV. Experiment Four: This movement consists of examples of "Markoff Chain music" in which successive note selection depends upon probability functions computed from tables derived from the overtone series and from consideration of vertical interval lengths. These probability functions also depend upon the relationship of successive melodic intervals both to one another and to the tonal center of the movement, which is C. We have attempted here to generalize in a simple way the dependence of melodic profile upon successive note relationships and upon longer range functions which define a species of tonality. Finally, the coda is included as an example of a simple closed structure generated by the computer. This coda is a generalized I - IV - V - I cadence developed from abstract probability formulations.

History: Probability in 20th century music

Later on, probabilities were “learned” (via n-gram statistics) from examples

$$P(n_i | n_{i-1}) =$$



Pluton (1988)
Philippe Manoury /
Miller Puckette

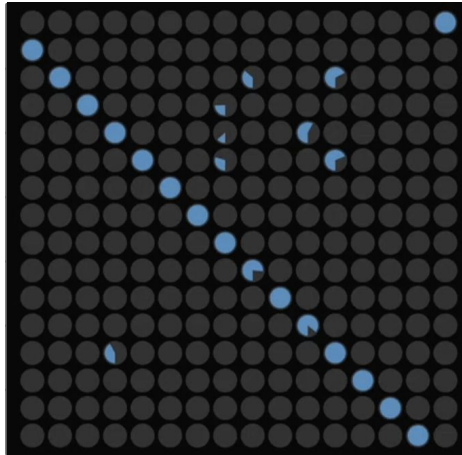


History: Probability in 20th century music

These simpler probabilistic techniques are still relevant today!

$$P(n_i | n_{i-1}) =$$

Untitled (2022)
Autechre



Modern deep learning / “gen AI” era

Key differences in modern approach:

$$P(n_i | n_{i-1}) =$$

	C	D	E	F	G	A	B
C	.00	.35	.20	.15	.15	.10	.05
D	.30	.00	.30	.15	.10	.10	.05
E	.15	.30	.00	.30	.10	.10	.05
F	.10	.10	.30	.00	.30	.10	.10
G	.20	.05	.10	.25	.00	.30	.10
A	.10	.05	.05	.15	.30	.00	.35
B	.45	.05	.05	.05	.05	.35	.00

Longer history (no
Markov assumption)

$$P(n_i | n_{<i}) =$$

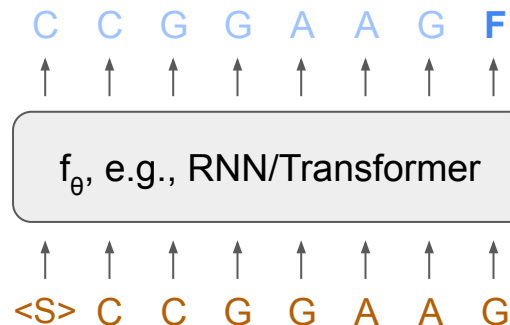
Same “output”
type signature →

C	D	E	F	G	A	B
.25	.05	.05	.50	.15	.05	.05

Neural network to fight
curse of dimensionality →

f_θ , e.g., RNN/Transformer

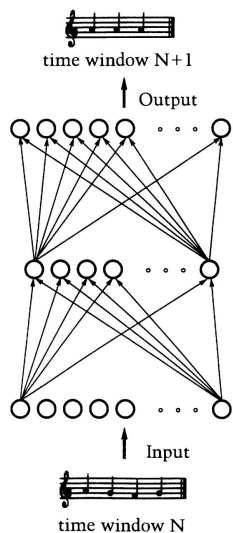
More training data →



Early experiments with neural networks

Todd CMJ'89

Autoregressive prediction



Eck/Schmidhuber NNSP'02

Early work with recurrent neural networks



Generations



Early experiments with neural networks

Todd CMJ'89

Eck/Schmidhuber NNSP'02

Doug Eck would later start the
Magenta team at Google

Early work with recurrent neural networks



MelodyRNN (Waite 16)

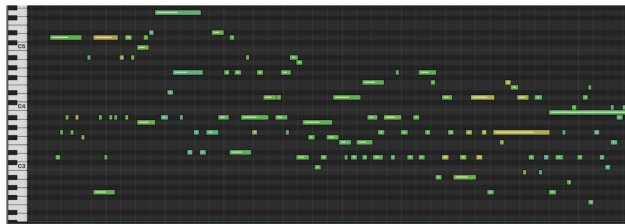


Generations



Example: training on *expressive* piano performance

PerformanceRNN ([Simon/Oore 17](#))



```
SET_VELOCITY<80>, NOTE_ON<60>  
TIME_SHIFT<500>, NOTE_ON<64>  
TIME_SHIFT<500>, NOTE_ON<67>  
TIME_SHIFT<1000>, NOTE_OFF<60>, NOTE_OFF<64>,  
NOTE_OFF<67>  
TIME_SHIFT<500>, SET_VELOCITY<100>, NOTE_ON<65>  
TIME_SHIFT<500>, NOTE_OFF<65>
```



Music Transformer (Huang+ ICLR'19)



Example: real-time generation conditioned on user control

Piano Genie (Donahue+ IUI'19)

 8 buttons ->  88 keys



<https://chrisdonahue.com/piano-genie>

The Flaming Lips play *Piano Genie*
using a plate of  fruit



<https://magenta.withgoogle.com/fruitgenie>

Performers

**Expertise
divide**

Audience

Piano Genie

1

2

3

4

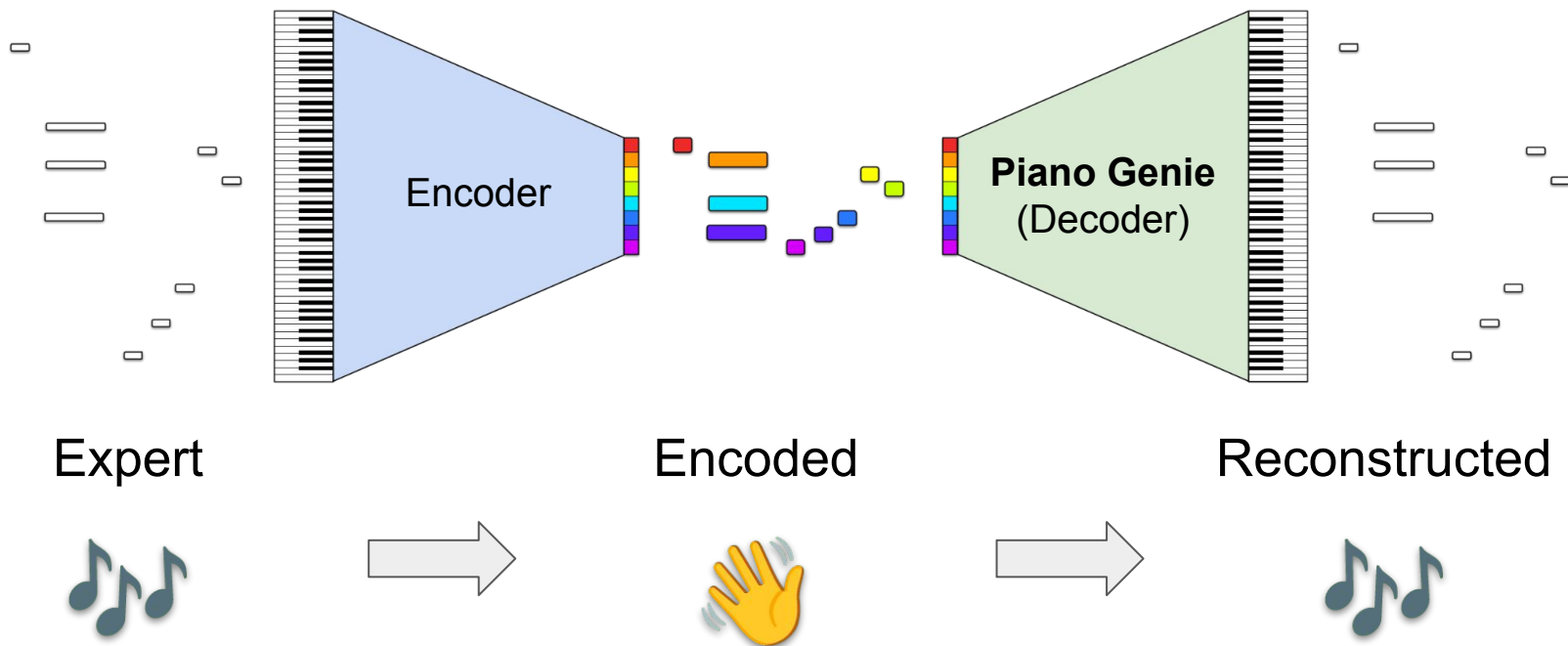
5

6

7

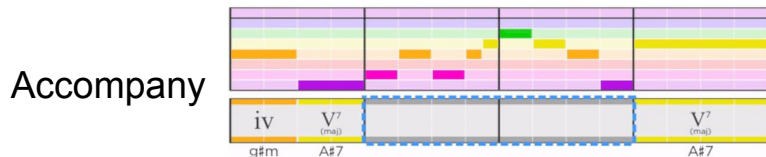
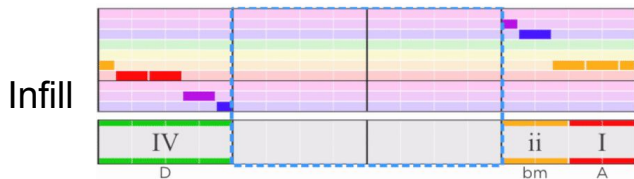
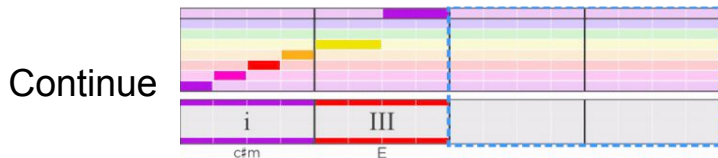
8

Autoencoder-based approach



Example: symbolic generation as music “co-pilot”

Anticipatory music transformer
(Thickstun+ TMLR'24)

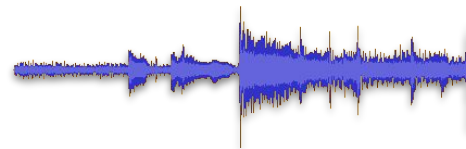


Powers *Hookpad Aria* co-pilot
(Donahue+ ISMIR'24)



<https://www.hooktheory.com/hookpad/aria>

Symbolic is great, but what about **audio**?



Audio

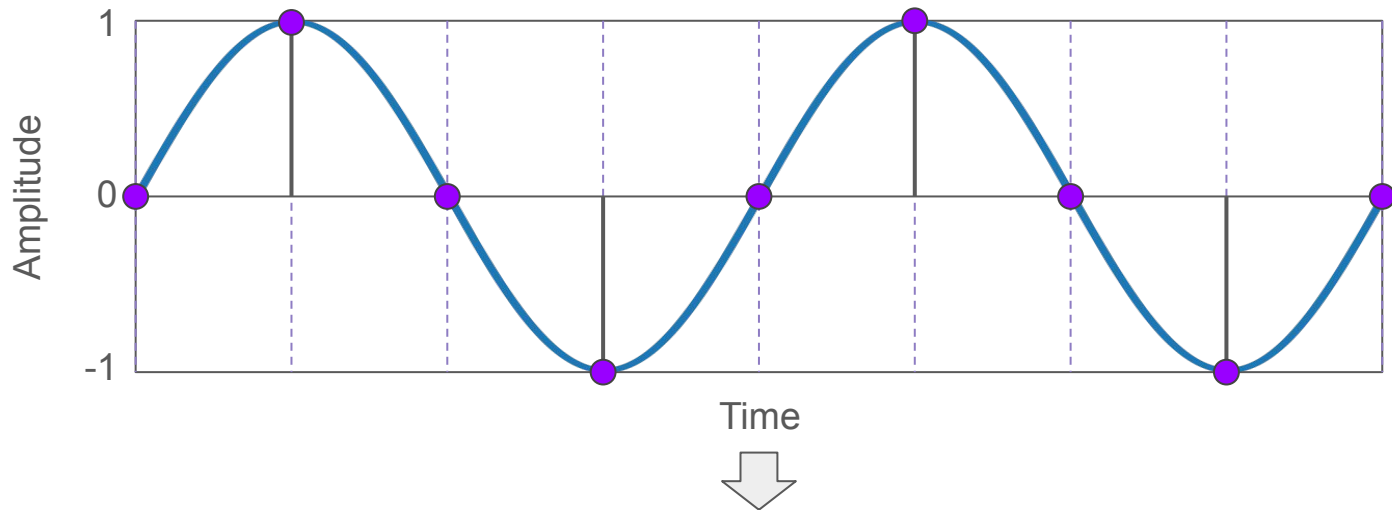
Overview

- Symbolic music LMs methods and history
- **Modern music audio LMs**
- Research frontiers in music LMs



Samples of audio can also be “tokens”!

Raw audio waveforms are also discretized through *sampling* at some uniform rate

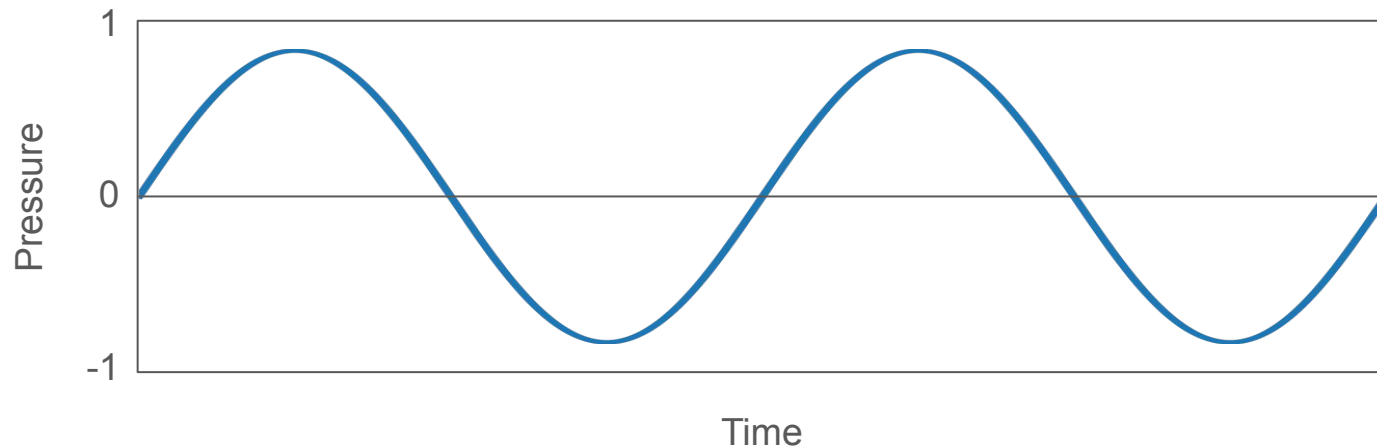


“Tokens” (samples): 0, 1, 0, -1, 0, 1, 0, -1, ...

A ⚡ primer on digital audio

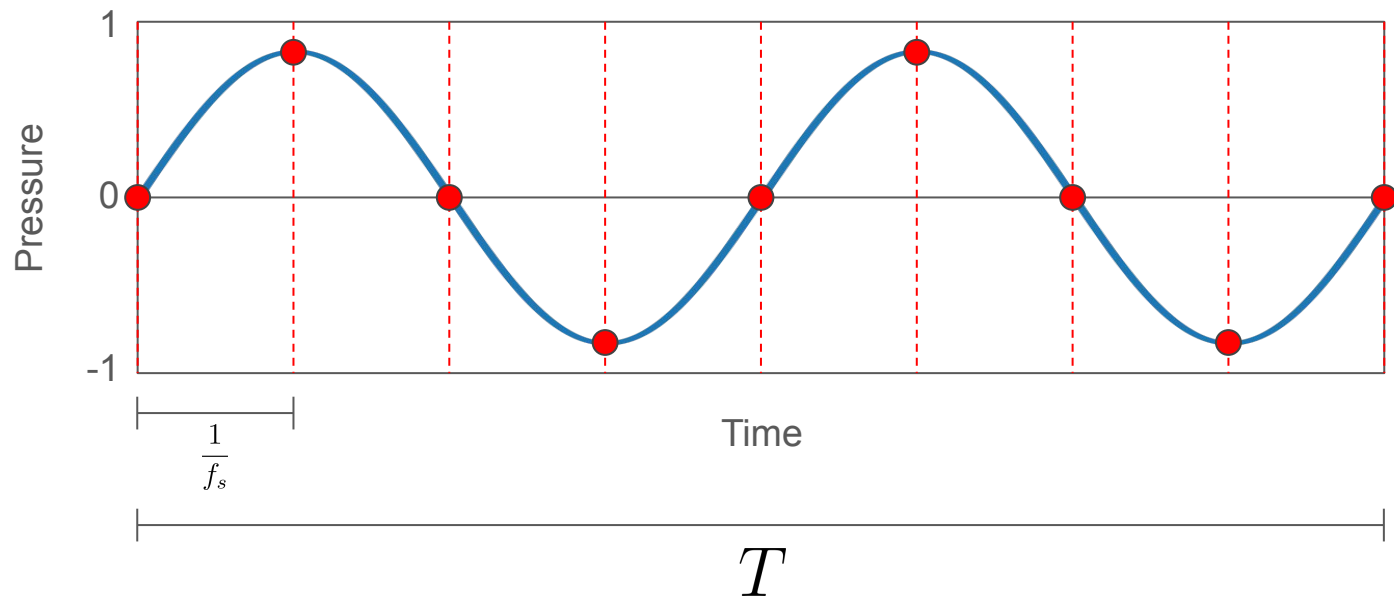
Audio is a *continuous* measurement of fluctuating air pressure caused by sound

But continuous signals cannot be natively stored on digital media



A ⚡ primer on digital audio

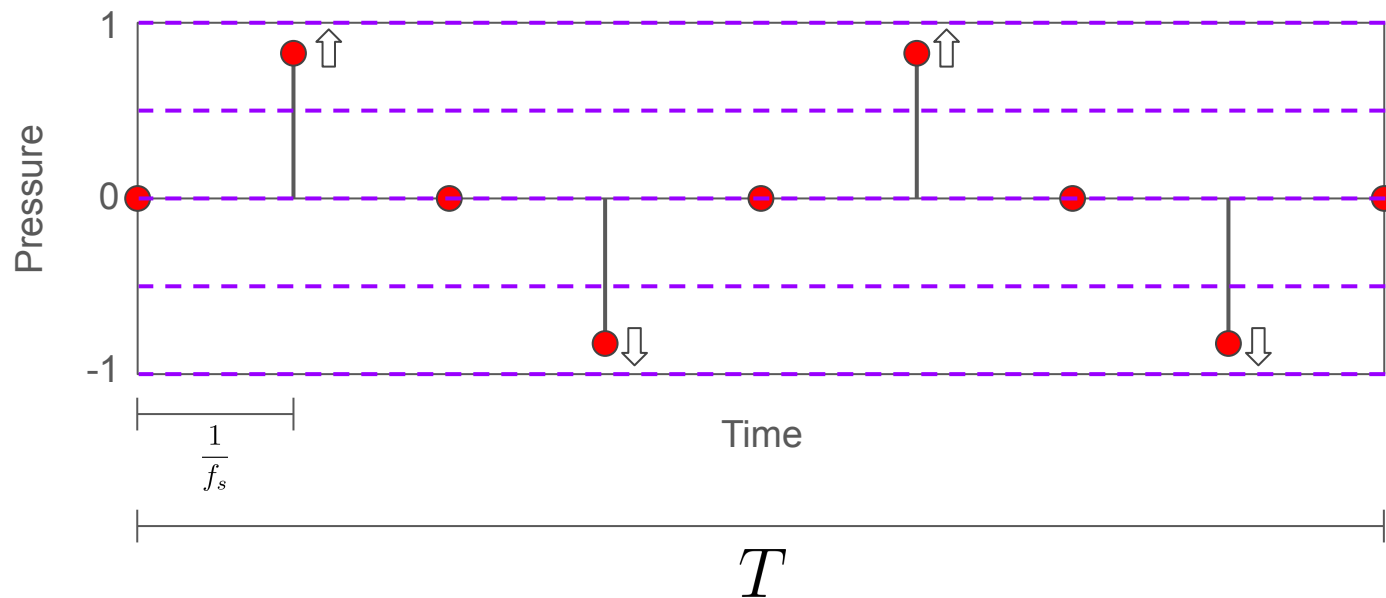
Digital audio involves *sampling* the signal at uniform intervals: $\mathbf{w} \in \mathbb{R}^{Tf_s}$



A ⚡ primer on digital audio

Digital audio further *quantizes* signals to set of discrete amplitudes: $\mathbf{w} \in \mathbb{V}^{Tf_s}$

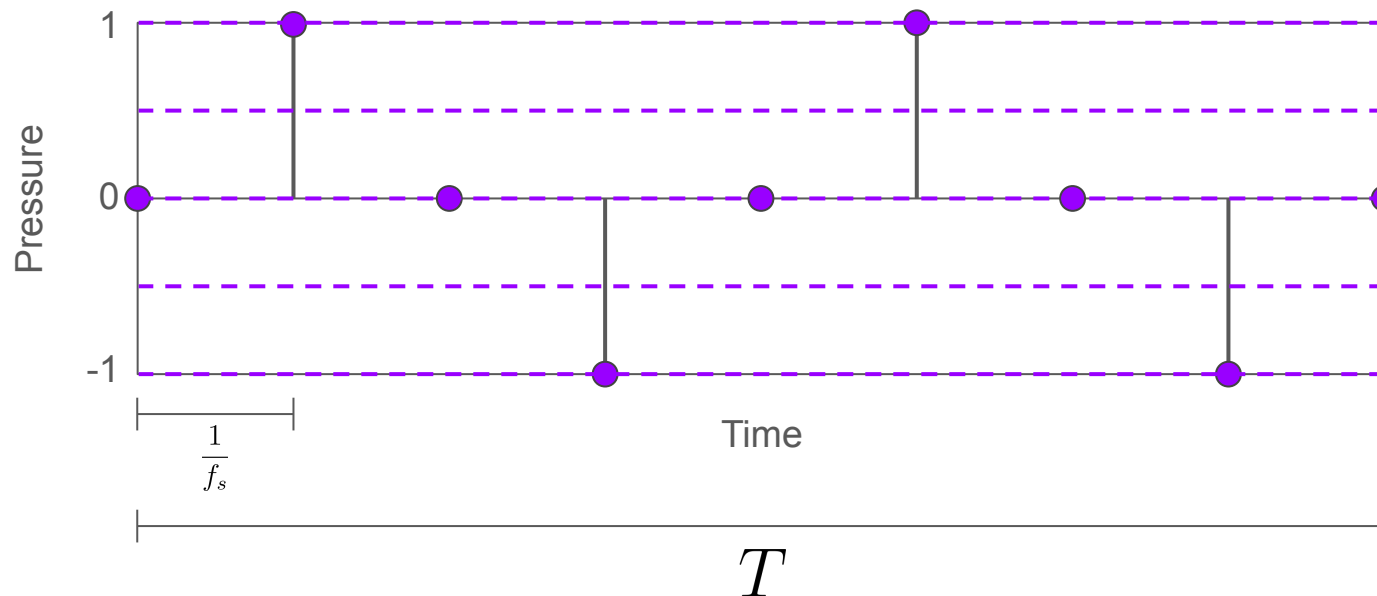
$$\mathbb{V} = \{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$$



A ⚡ primer on digital audio

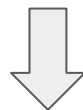
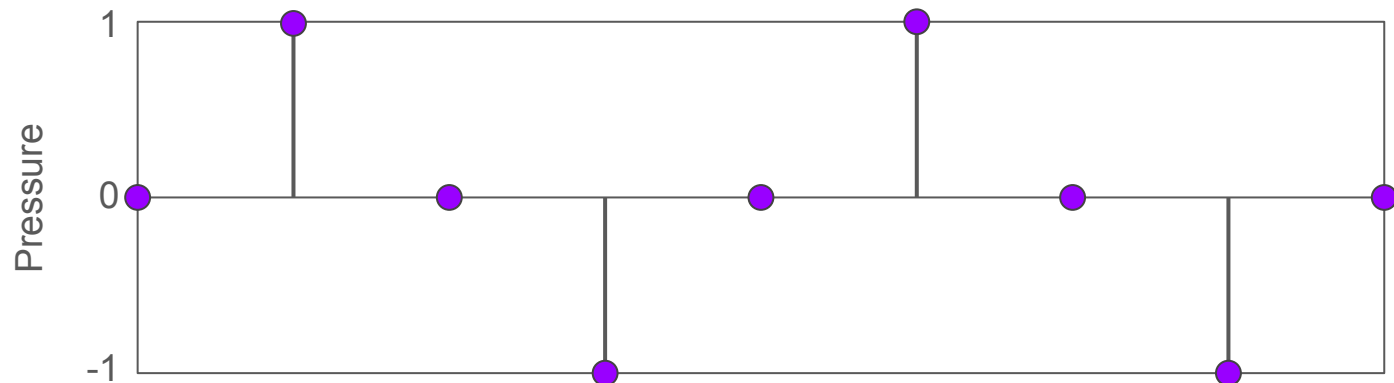
Digital audio further *quantizes* signals to set of discrete amplitudes: $\mathbf{w} \in \mathbb{V}^{Tf_s}$

$$\mathbb{V} = \{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$$



Waveforms can be “tokens”!

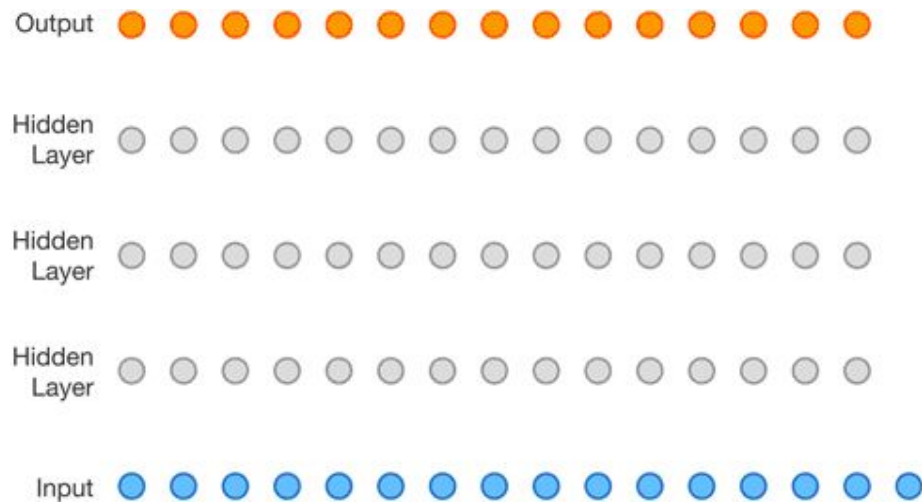
Hence, we could simply treat each *sample* of a digital audio *waveform* as a token



0, 1, 0, -1, 0, 1, 0, -1, ...

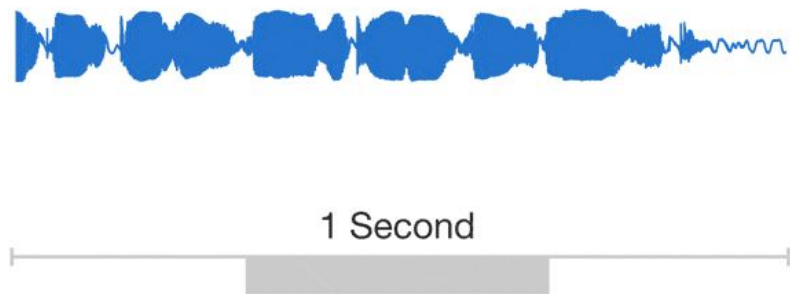
Modeling raw audio waveforms

WaveNet (van den Oord+ 16), a “language model” on raw audio time series

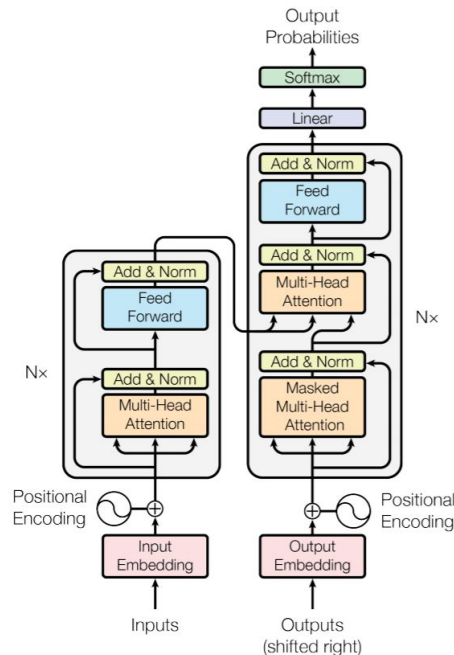


But there's a catch...

Tens of thousands of samples per second
Structure at many different timescales

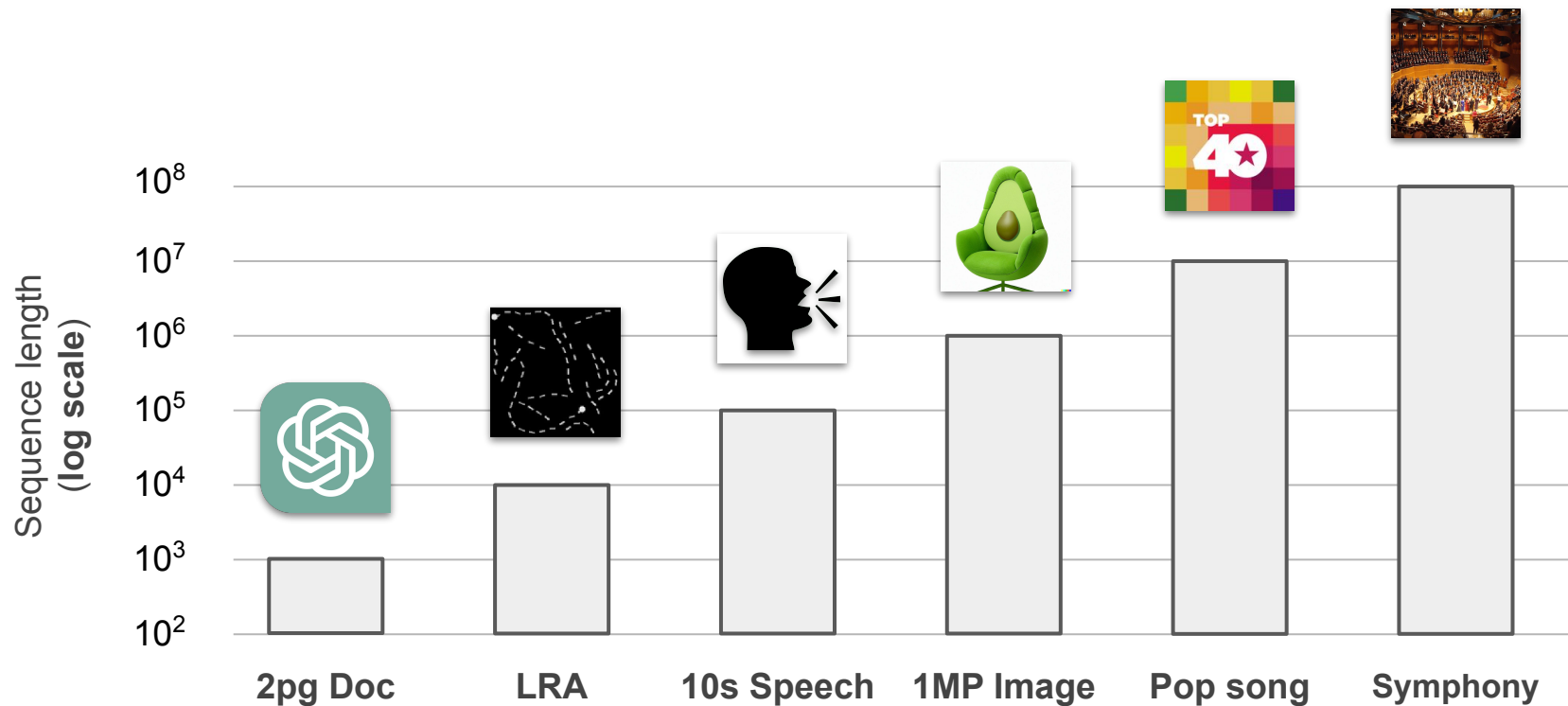


Waveforms are long



Achilles heel for Transformers

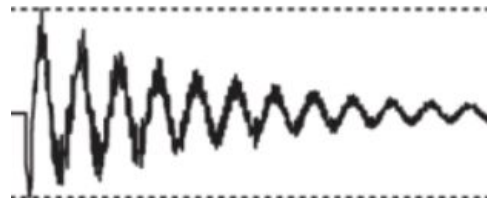
Waveform lengths in perspective



Structure across 6+ orders of magnitude

- ~2ms: Period of 440 Hz sinusoid
- ~10ms: Length of musical *attack*

Acoustic



- 500ms: Length of one beat at 120 BPM
- 2s: Length of one measure at 120 BPM
- ~20s: Length of a pop chorus
- ~180s: Length of a pop song
- ~3600s: Length of a symphony

Symbolic



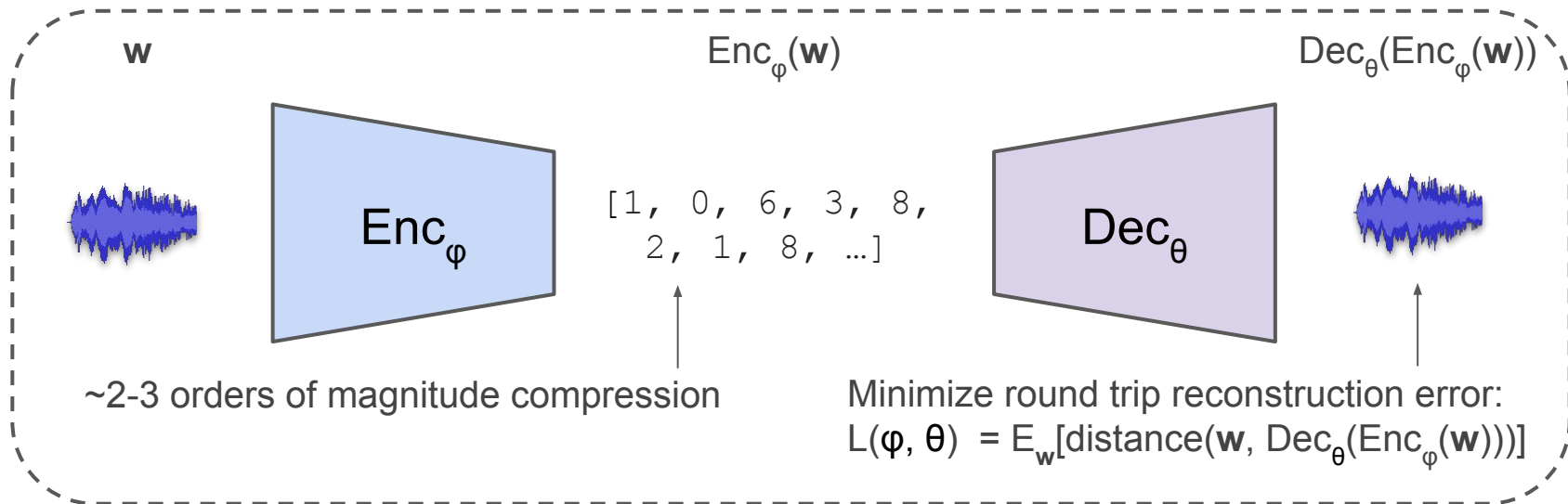
How can we tractably model audio with LMs?



Maybe we can compress it first!

Codec language models

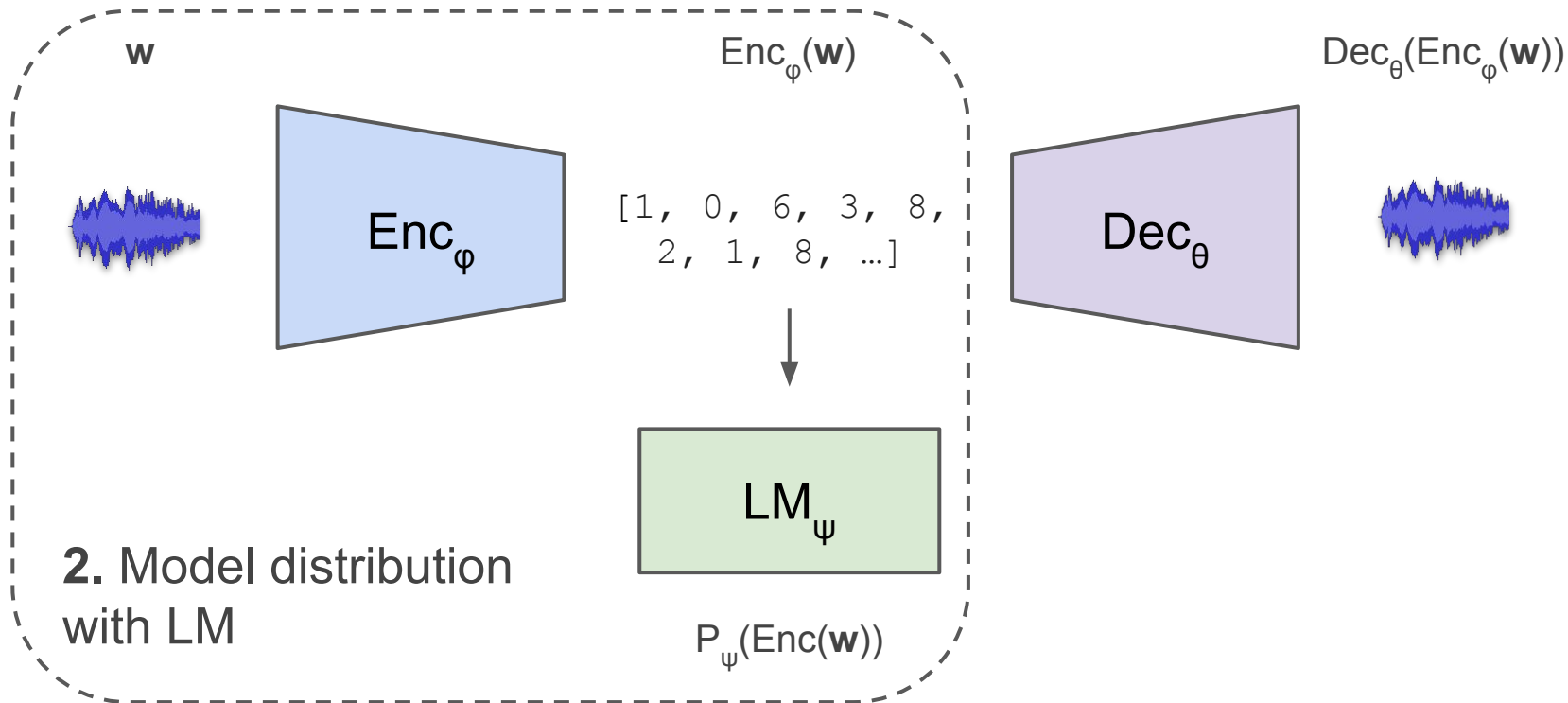
van den Oord+ NeurIPS'17, Dieleman+ NeurIPS'18, Dhariwal/Jun/Payne+ 20



1. Learn a discrete *codec*

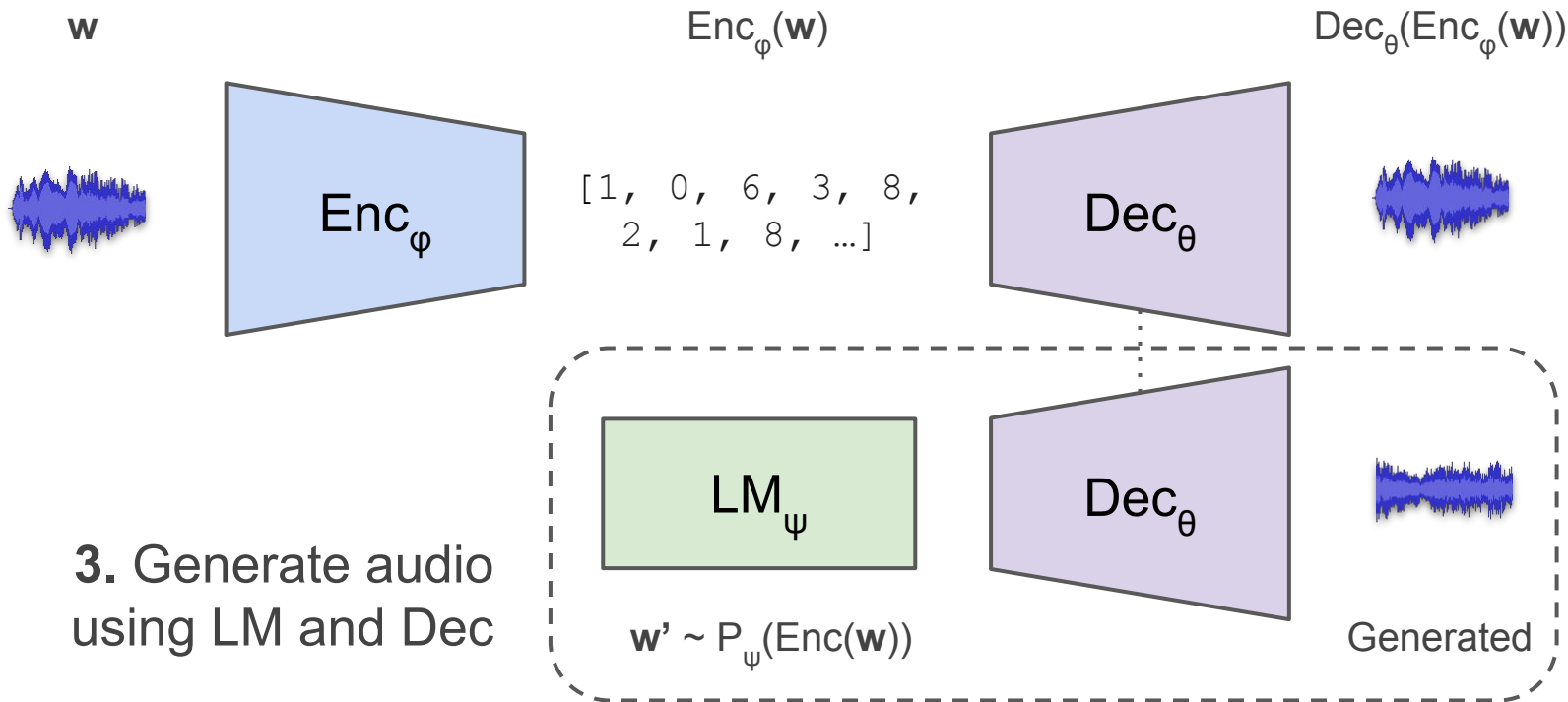
Codec language models

van den Oord+ NeurIPS'17, Dieleman+ NeurIPS'18, Dhariwal/Jun/Payne+ 20



Codec language models

van den Oord+ NeurIPS'17, Dieleman+ NeurIPS'18, Dhariwal/Jun/Payne+ 20



Codec LMs for music audio

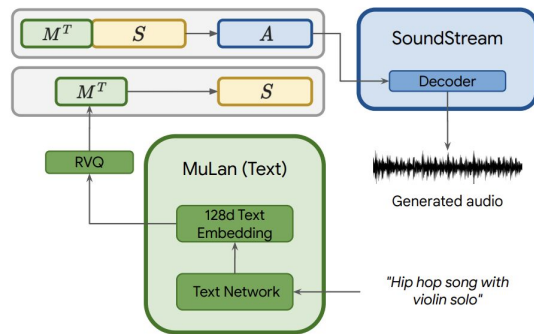
Jukebox ([Dhariwal/Jun/Payne+ 20](#))

Generation in broad acoustic styles with lyrics

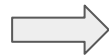


MusicLM ([Agostinelli+ 23](#))

Generation with open domain text control

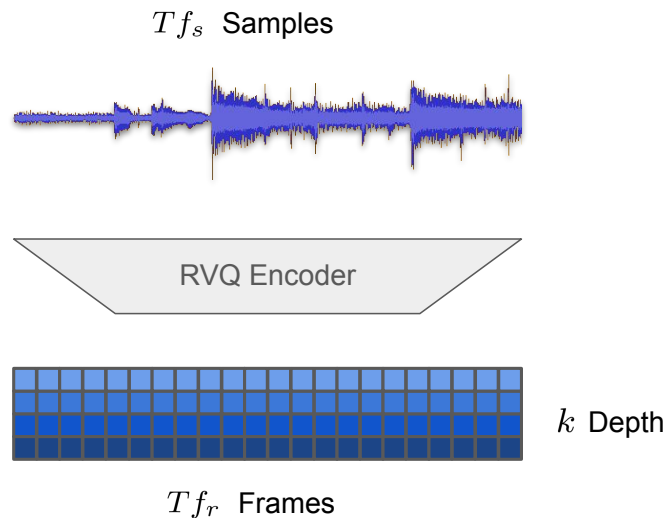


"Funky piece that grows gradually towards a strong, danceable beat and a prominent bassline."



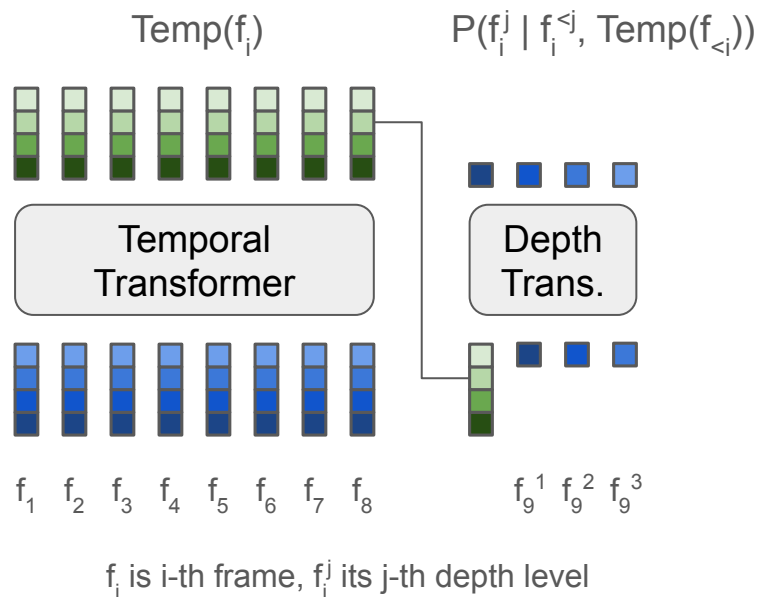
Modern “2D” codec LMs

Residual vector quant autoencoders (Zeghidour+ TASLP'21)



Where else might this
2D structure arise?

2D nested Transformers (Yu/Simig/Flaherty+ 23, Défossez+ 24)

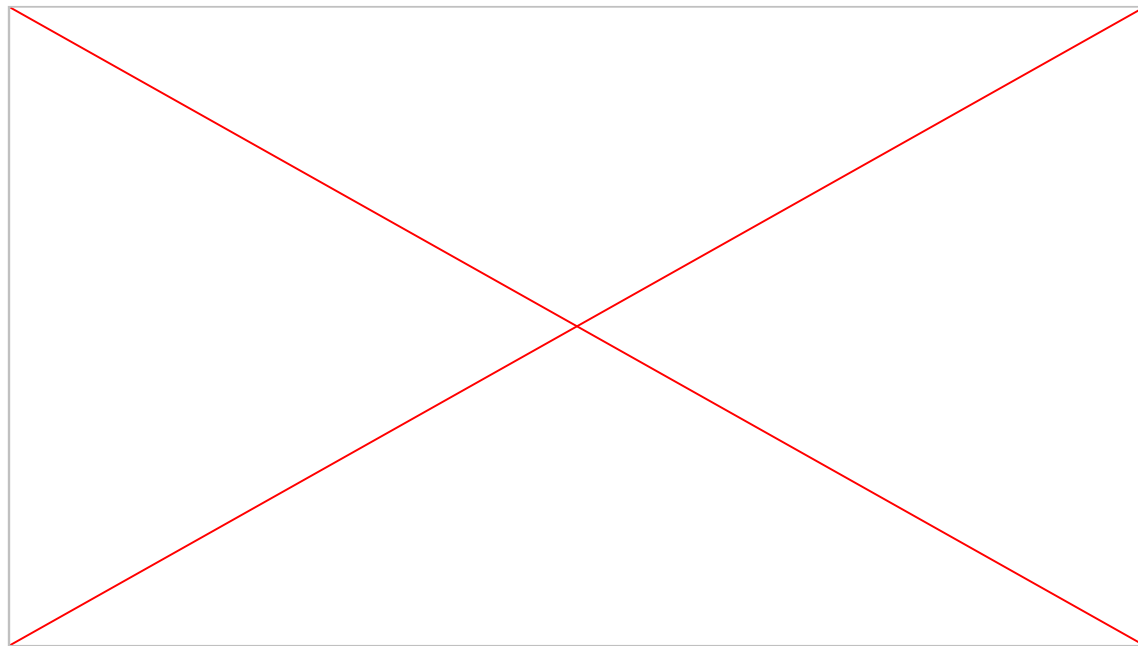


Example: Magenta RealTime

(GDM Lyria Team, NeurIPS'25)

Limited to **10s of audio history** for efficiency

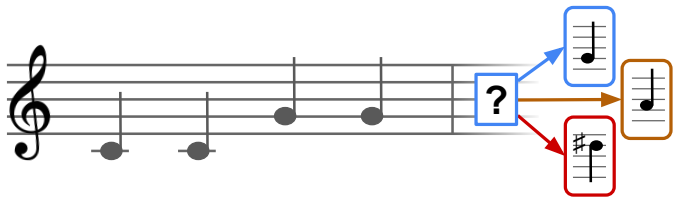
Text control
via CLIP-style
contrastive
text/audio
encoders



Generates in **2s chunks** to achieve real-time

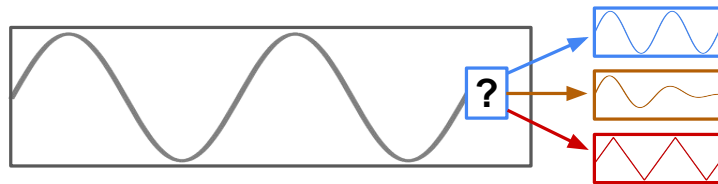
Understanding music requires both scores and audio!

Score LMs



- Scores are editable
- Cheaper to model
- Lacking detail
- Not universal

Audio LMs



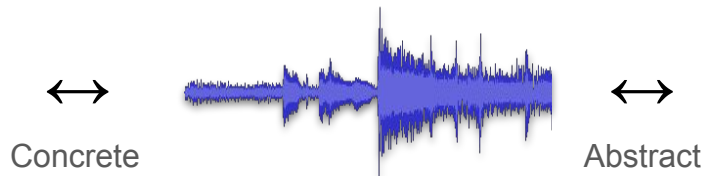
- Captivating, full detail
- “Universal”
- Expensive to train/generate
- Audio is not easily editable

Frontiers

Overview

- Symbolic music LMs methods and history
- Modern music audio LMs
- **Research frontiers in music LMs**

Music is inherently *multimodal*

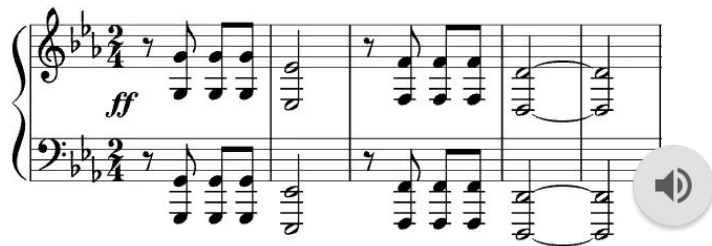


Frontier models fail to understand *concrete* multimodal relationships in music



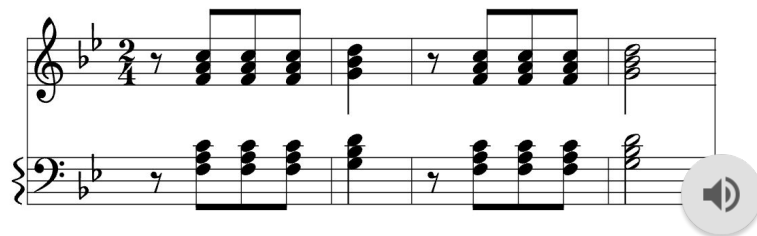
Translate this into ABC notation

[Input  image]



```
M:2/4
L:1/8
K:Bb
V:1 treble
%%staves {1 2}
V:2 bass
[V:1] z [FAc][FAc][FAc] | [G2B2d2] | z
[FAc][FAc][FAc] | [G4B4d4] |
[V:2] z [F,A,C][F,A,C][F,A,C] | [G,2B,2D2]
| z [F,A,C][F,A,C][F,A,C] | [G,4B,4D4] |
```

[Render of Claude's ABC output]



Frontier models fail to understand *concrete* multimodal relationships in music



Translate this into ABC notation

[Input  audio]



```
X:1
M:2/4
L:1/16
Q:1/4=240
K:Cmaj
|: C E G c G E C z | G B D g D B G z | C E
G c G E C z | F A C f C A F z :|
```

[Synthesis of Gemini's ABC output]



Our recent work: LM-based $\rightarrow$ **music *translation***

Jung+ TASLP'25

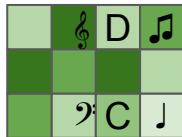
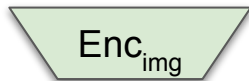


Irmak Bukey
CSD PhD Student



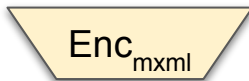
Method: modality-specific, locality-sensitive tokenization

Score Image

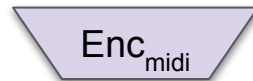
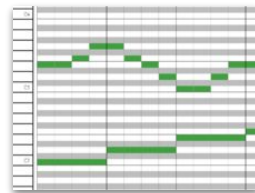


MusicXML

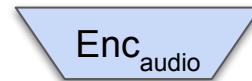
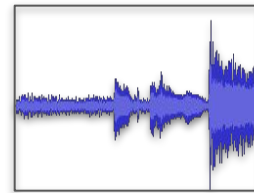
```
<note>  
  <pitch>  
    <step>E</step>  
    <alter>-1</alter>  
    <octave>4</octave>  
  </pitch>  
  <duration>2</duration>  
  <type>half</type>  
</note>
```



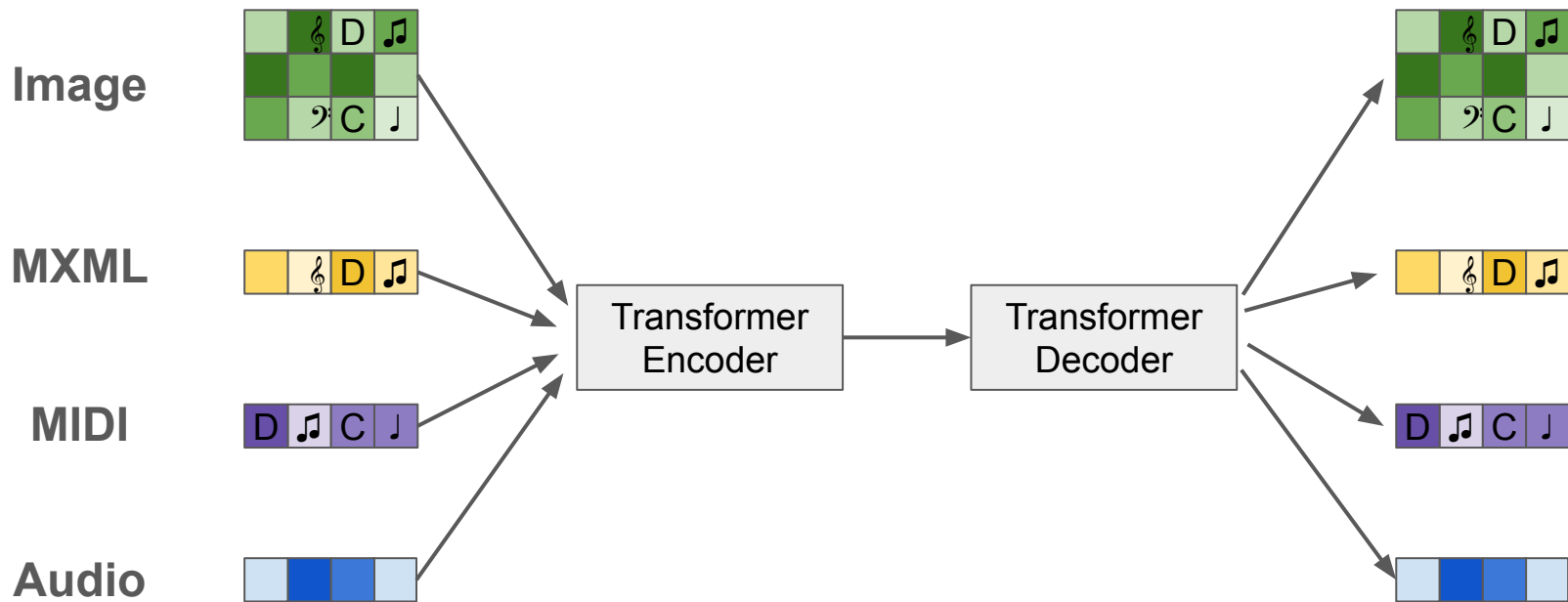
MIDI



Music Audio



Method: unified LM-based approach to music *translation*



Contextual *and* multimodal control with Aria and AMUSE

Kim+ CHI'25 🏆 Best Paper



Yewon Kim
CSD PhD Student

The poster for the AMUSE project features the title "AMUSE" in large, bold letters, with a small illustration of three people sitting around a table. Below the title is the subtitle "Human-AI Collaborative Songwriting with Multimodal Inspirations". A link to "Learn more at: yewon-kim.com/amuse" is provided. At the bottom, there are three circular portraits of the team members: Yewon Kim, Sang-Ju Lee, and Chris D'Amico. To the right of the text is a large, colorful visualization of a songwriting process, showing multiple tracks of music and lyrics over time.

Context-aware **music** adaptation

Wang+ CHI'24, UIS'24



Alex Wang
HCII PhD Student

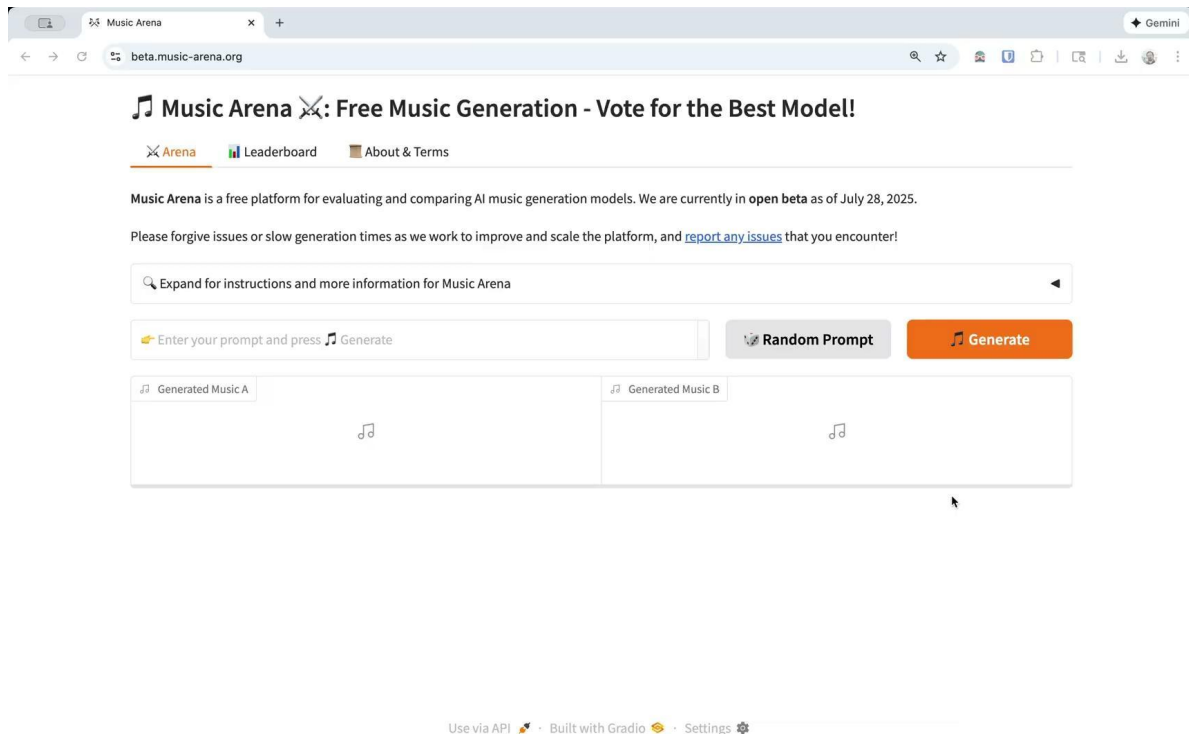
Towards Music-Aware Virtual Assistants



Video contains audio

In-the-wild evaluation music-arena.org

Kim+ NeurIPS'25



 Vote

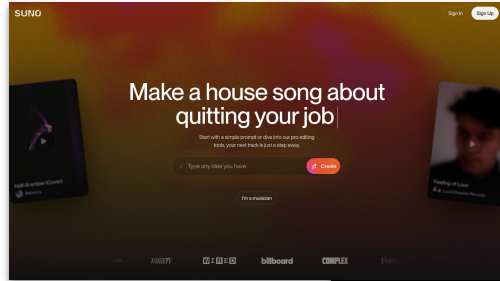


~4k votes collected
since July 2025



Wayne Chi
CSD PhD Student

AI is actively transforming music



Rolling Stone

**AI's Most Ambitious Music Generators
Accused of 'Massive' Infringement In
New RIAA Lawsuit**

THE VERGE

**YouTube is trying to make AI music deals
with major record labels** / Google is taking a
'permission first' approach as labels and artists sue
other AI music companies.

Introducing
Dream Track
featuring Charlie Puth



**North Carolina Musician Charged With Music
Streaming Fraud Aided By Artificial Intelligence**

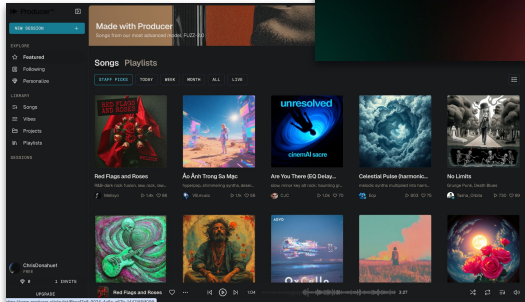
The New York Times

**An A.I. Hit of Fake 'Drake' and 'The
Weeknd' Rattles the Music World**

A track like "Heart on My Sleeve," which went viral before being
taken down by streaming services this week, may be a novelty for
now. But the legal and creative questions it raises are here to stay.

Pitchfork

**Grimes Unveils Software to Mimic Her
Voice, Offering 50-50 Royalties for
Commercial Use**



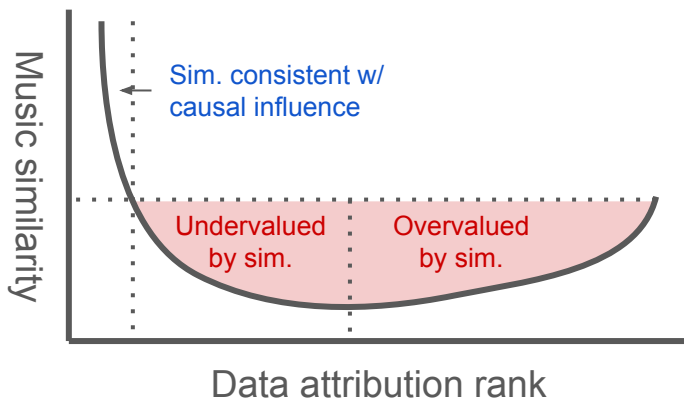
Training data attribution for music

Deng+ '25, Under review

For *policymakers*

Similarity $\neq$ causal influence

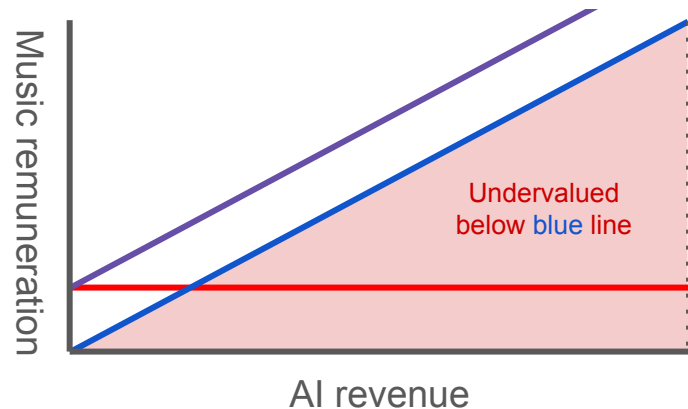
Causal may require regulation



For *rightsholders* (artists, labels)

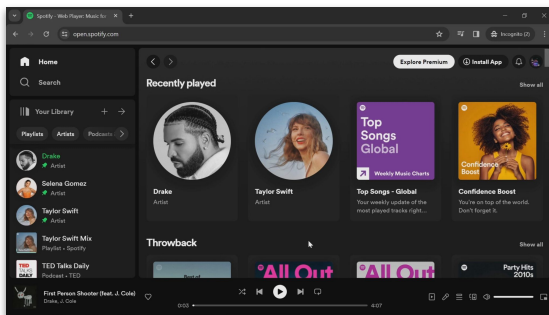
Push for “fair deployment”!

■ Fairly trained ■ Fairly deployed ■ Both

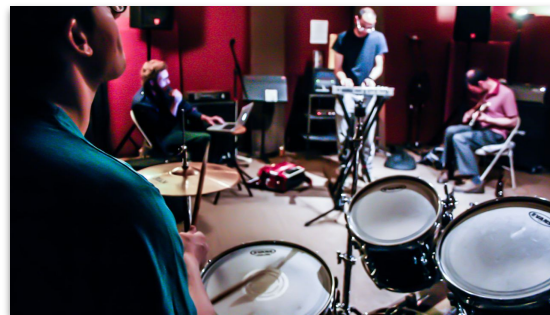


Music is not just *a product*, but a *process*

Product



Process



AI has tremendous potential to both:

automate the product 🌹

enrich the process 🌹



Overview

- Symbolic music LMs methods and history
- Modern music audio LMs
- Research frontiers in music LMs
- Today was a crash course! If you're interested:
 - Take 15-798 Generative AI for Music and Audio, planned Spring 2027
 - Learn more at gclef-cmu.org
 - Email me chrisdonahue@cmu.edu
- Thanks! Questions?



Would you rather go to a great concert by an AI performer or a mediocre concert by a human performer?