NeurIPS 2025 Creative AI Panel #2

Art Content Creation: When demands are met by pipelines (or not)

Music Arena: Live Evaluation for Text-to-Music

Yonghyun Kim[♯♯]

Wayne Chi[♯]

Yuki Mitsufuji[♭]

Chris Donahue[♯]

🌐Platform

Anastasios N. Angelopoulos[♮]

Wei-Lin Chiang[♮]

Koichi Saito[♭]

Shinji Watanabe[♯]

More recently:
Nathan Pruyne[♯]

[♯] CMU [♭] Sony AI [♮] LMArena [♯♯] Georgia Institute of Technology

NEURAL INFORMATION PROCESSING SYSTEMS

Emergence of Music Generative AI

Rapid Democratization of Text-To-Music (Audio) Creation

Suno

Producer AI

udio

Sonauto

ElevenLabs

ACE STUDIO

MusicGen

Magenta RealTime

Problem: Eval on Music Generative AI

How can we Evaluate Generated Music?

Objective

FD (FAD): overall quality & **distributional similarity** (Set-level).
IS: intrinsic **fidelity** and **diversity** (Unpaired).
KLD: content consistency via paired comparisons (Instance-level).
CLAP Score: semantic alignment between **text and audio** (Instance-level).

Model	AA-Train.	TA-Train.	FD_{pann} ↓	FD_{vgg} ↓	Inception Score ↑	KL Div. ↓
Riffusion [8]	✗	✓	68.95	10.77	1.34	5.00
MuBERT [26]	—	—	31.70	19.04	1.51	4.69
AudioLDM	✓	✗	38.92	3.08	1.67	3.65
MusicLDM	✓	✗	26.67	2.40	**1.81**	3.80
MusicLDM (Only TA-Training)	✗	✓	32.40	2.51	1.49	3.96
MusicLDM w/. mixup	✓	✗	30.15	2.84	1.51	3.74
MusicLDM w/. BAM	✓	✗	28.54	**2.26**	1.56	3.50
MusicLDM w/. BLM	✓	✗	**24.95**	2.31	1.79	**3.40**
MusicLDM w/. Text-Finetune	✓	✓	27.81	1.75	1.76	3.60
MusicLDM w/. BAM & Text-Finetune	✓	✓	28.22	1.81	1.61	3.61
MusicLDM w/. BLM & Text-Finetune	✓	✓	**26.34**	**1.68**	**1.82**	**3.47**

Following evaluation techniques used in past work on audio generation [24], we use frechet distance (FD), inception score (IS), and kullback-leibler (KL) divergence to evaluate the quality of generated musical audio outputs. Frechet distance evaluates the audio quality by using an audio embedding model to measure the similarity between the embedding space of generations and that of targets. In this paper, we use two standard audio embedding models: VGGish [12] and PANN [20]. The resulting distances we denote as FD_{vgg} and FD_{pann}, respectively. Inception score measures the diversity and the quality of the full set of audio outputs, while KL divergence is measured on individual pairs of generated and groundtruth audio samples and averaged. We use the `audioldm_eval` library[1] to

	channels/sr	output length	FD_{openl3} ↓	KL_{passt} ↓	$CLAP_{score}$ ↑
AudioLDM2-48kHz [4]	1/48kHz	10 sec	101.11	2.04	0.37
AudioLDM2-large [4]	1/16kHz	10 sec	170.31	1.57	0.41
AudioGen-medium [20]	1/16kHz	10 sec	186.53	1.42	0.45
Stable Audio 1.0 [5]	2/44.1kHz	95 sec †	103.66	2.89	0.24
Stable Audio 2.0 [6]	2/44.1kHz	190 sec †	116.14	2.67	0.24
Stable Audio 2.0 [6]	2/44.1kHz	285 sec †	110.62	2.70	0.23
Stable Audio Open	2/44.1kHz	47 sec †	**78.24**	**2.14**	**0.29**

We employ established quality metrics[3] that include FD_{openl3} [26], KL_{passt} [27] and $CLAP_{score}$ [10, 28]. A low FD_{openl3} implies that the generated audio is plausible and closely matches the reference [29, 8]. A low KL_{passt} indicates semantic correspondence between the generated and the reference audio [8]. A high $CLAP_{score}$ denotes that the generated audio adheres to the given text prompt [10, 28]. We use two evaluation sets: AudioCaps Dataset [30] for sound generation, and Song Describer Dataset [31] for music generation.

Chen, Ke, et al. "**MusicLDM**: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies." *ICASSP 2024.*
Evans, Zach, et al. "**Stable audio open**." *ICASSP 2025.*

Problem: Eval on Music Generative AI

Why are current Objective Metrics Imperfect?

Objective

FD (FAD): overall quality & **distributional similarity** (Set-level).
IS: intrinsic **fidelity** and **diversity** (Unpaired).
KLD: **content consistency** via paired comparisons (Instance-level).
CLAP Score: **semantic alignment** between **text and audio** (Instance-level).

Absence of Musicality & Structure	Fail to capture temporal coherence, rhythm, and melody.
Perceptual Artifacts	Statistical averages dilute metallic noises or unnatural glitches.
Semantic Nuance (CLAP)	High alignment scores do not guarantee accurate mood or style.
Misalignment with Human Preference	Metric scores do not always correlate with human perception.

Subjective

Gold Standard: Indispensable for more accurate evaluation.

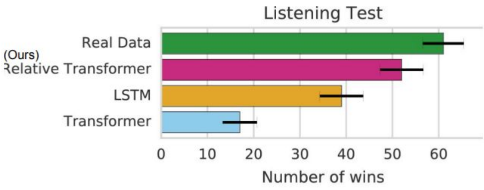
Problem: Human Eval on Music Generative AI

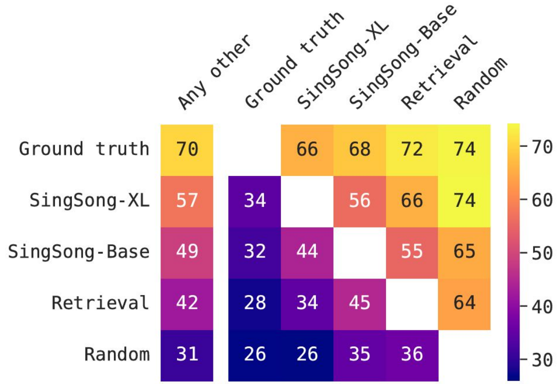
Why is the “Gold Standard” still failing us?

Inconsistent Lack of standardized questions or rating scales across papers.

Subjective Listening Test		
Quality↑	Relevance↑	Musicality↑
2.02	1.50	**2.33**
1.98	2.17	2.19
—	—	—
2.04	2.21	2.01
2.13	**2.31**	2.07

MEL. ↑	OVL. ↑	REL. ↑
64.44±0.83	82.18±1.21	81.54±1.22
61.89±0.96	81.65±1.13	**82.50**±0.98
72.87±0.93	**83.94**±1.99	80.28±1.06





Chen, Ke, et al. "**MusicLDM**: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies." *ICASSP 2024.*
Copet, Jade, et al. "**Simple and controllable music generation**." *NeurIPS 2023.*
Huang, Cheng-Zhi Anna, et al. "**Music Transformer**: Generating Music with Long-Term Structure." *ICLR 2019.*
Donahue, Chris, et al. "Singsong: Generating musical accompaniments from singing." *arXiv preprint arXiv:2301.12662 (2023).*

Platform

NEURAL INFORMATION PROCESSING SYSTEMS

Problem: Human Eval on Music Generative AI

Why is the "Gold Standard" still failing us?

Inconsistent Lack of standardized questions or rating scales across papers.

Unscalable Collecting high-quality human feedback is slow and costly.

musical. We recruited crowd workers on the Amazon Mechanical Turk platform to perform these tasks. We paid workers \$0.75 US dollars for each pairwise evaluation. Assuming that workers listen to each clip twice—and spend an additional 40 seconds to make their decision and overhead time between tasks—this amounts to two minutes of time per task, or a \$22.50 hourly rate. We pre-qualified workers for by asking them to distinguish between five pairs of human compositions versus melodies accompanied by the random retrieval baseline (described below for the accompaniment task).

between 3% and 16% of the total preferences (standard deviation of 5%). We paid annotators \$0.40 USD per pair of audio examples and the median time spent per pair was 68 seconds, equating to a wage of around \$21 USD per hour. The total cost of the data collection procedure was roughly \$1200 USD.

Thickstun, John, et al. "**Anticipatory Music Transformer**." *Transactions on Machine Learning Research (April 2024).*
Huang, Yichen, et al. "**Aligning Text-to-Music Evaluation with Human Preferences**." *ISMIR 2025.*

Platform

NEURAL INFORMATION PROCESSING SYSTEMS

Problem: Human Eval on Music Generative AI

Why is the "Gold Standard" still failing us?

Inconsistent	Lack of standardized questions or rating scales across papers.
Unscalable	Collecting high-quality human feedback is slow and costly.
Unrealistic	Controlled listening tests != Real-world music enjoyment.

Please use headphones in a quiet environment if possible.

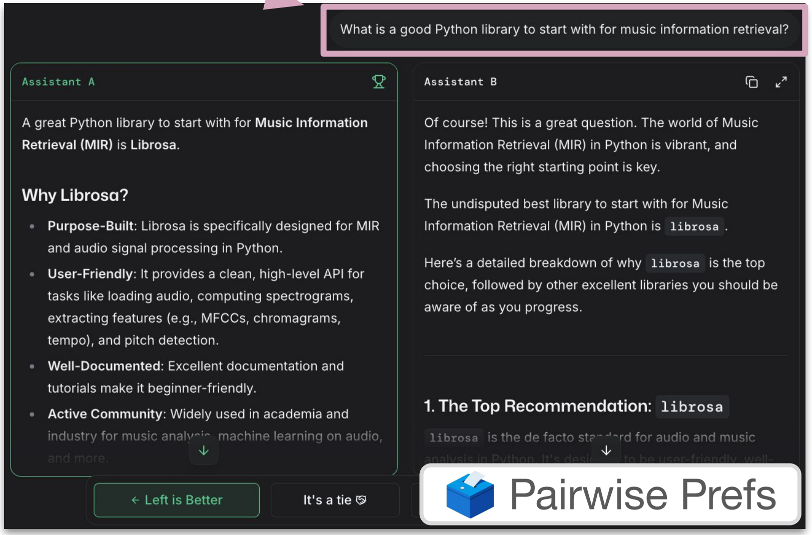
You will be presented two recordings of computer-generated music. Please compare them in audio fidelity and musicality.

Fidelity: How clear is the audio? Does it sound like it's coming from a walkie-talkie (bad fidelity) or a studio-quality sound system (excellent fidelity)?

Musicality: How conventionally musical is the recording? Does it feel like a well-composed song (excellent musicality) or sound more abstract, noisy, or chaotic, with minimal resemblance to typical music patterns (bad musicality).

Inspiration: Success of Live Evaluation

Realistic: real queries, real users, real goals



LMArena

Scalable (access 🤝 preferences)

Rank	Model	Score	95% CI (±)	Votes	Organization
1	gemini-3-pro	1498 Preliminary	±11	3,768	Google
2	grok-4.1-thinking	1483 Preliminary	±11	3,467	xAI
3	grok-4.1	1464 Preliminary	±10	3,588	xAI
4	gpt-5.1-high	1454	±11	3,796	OpenAI
5	gemini-2.5-pro	1451	±4	66,734	Google
6	claude-sonnet-4-5-20250929-thinking-32k	1450			

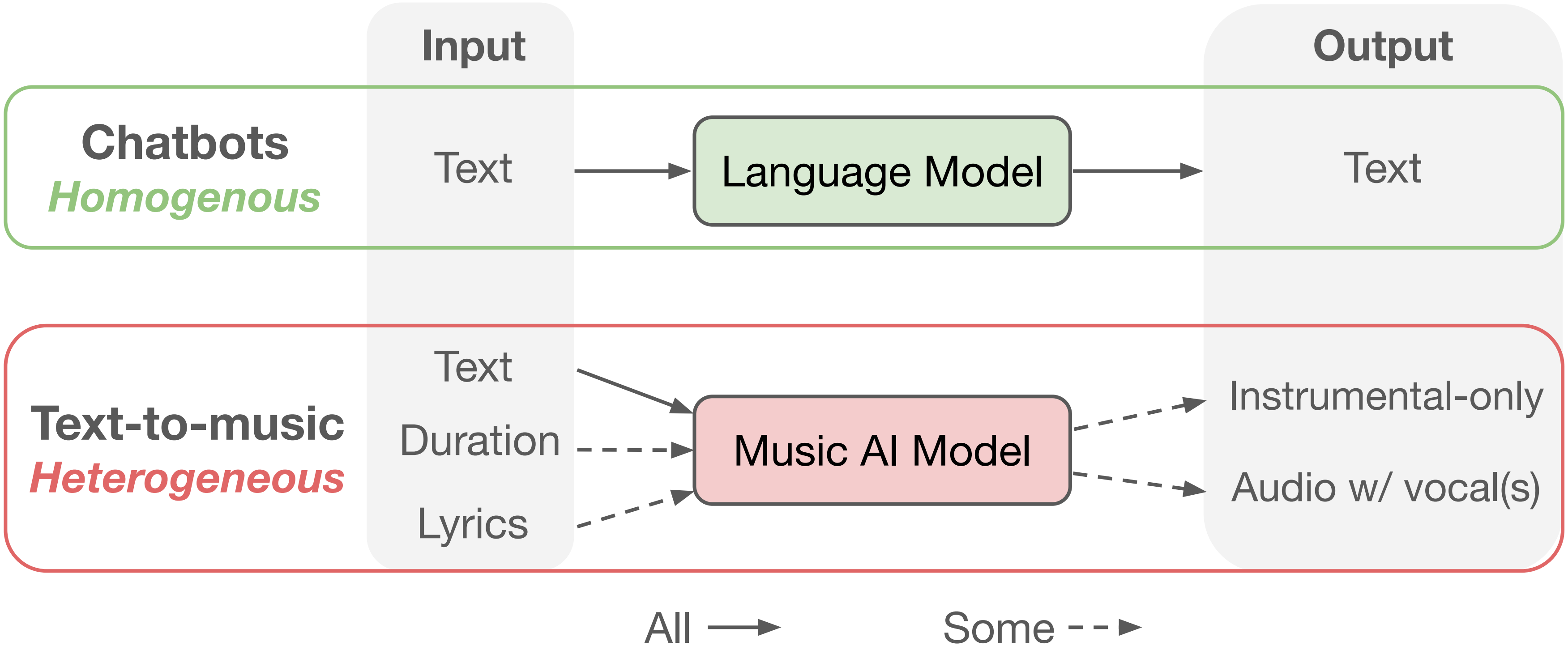
🥇 Leaderboard

Consistent: "scores" directly comparable across models

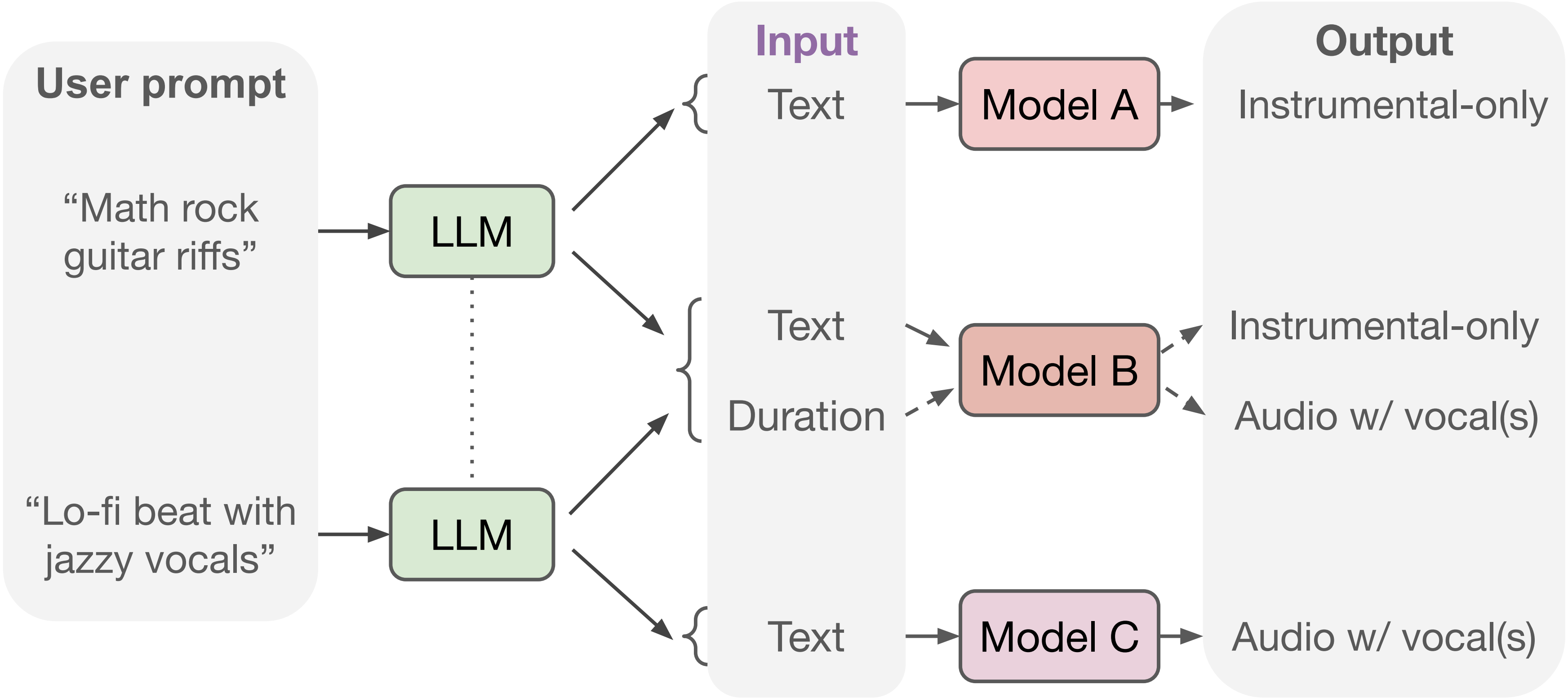
Platform

NEURAL INFORMATION PROCESSING SYSTEMS

Challenge: Adapting Live Evaluation to Music

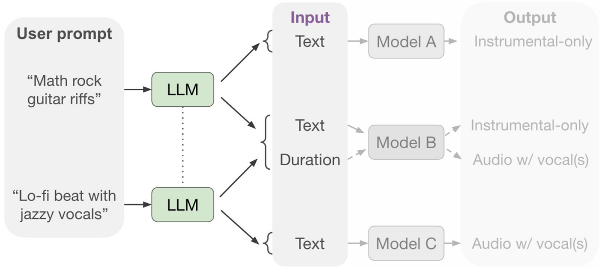


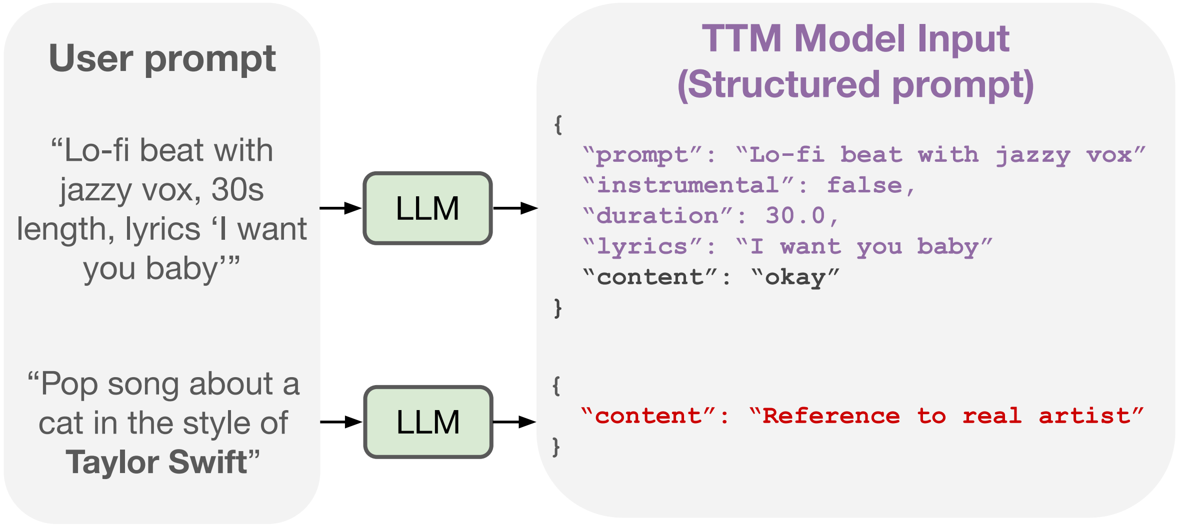
Solution: Unified text input via LLM-based routing



Solution: Unified text input via LLM-based routing

User prompt → Structured prompt w/ Content moderation





Platform

NEURAL INFORMATION PROCESSING SYSTEMS

Opportunity: Going beyond binary preferences

We collect …

🗳️ Votes

🫶 Thank you for voting! You voted for 🥇 **Riffusion FUZZ 1.0** in favor of 🥈 **ACE-Step**.

👈 A is better | B is better 👉

👂 Listening Data

✍️ Language Feedback

What did you *like* about 🥇 Generated Music A (from Riffusion FUZZ 1.0)?

The music quality was much higher overall, though the style wasn't quite what I asked for

What did you *dislike* about 🥈 Generated Music B (from ACE-Step)?

The music was lower quality and shorter, though I liked the lyrics.

Opportunity: Going beyond binary preferences

We col

E.g., listening behavior differs greatly by track order.
Median listening time for first track 3x that of second

👂 **Listening Data**

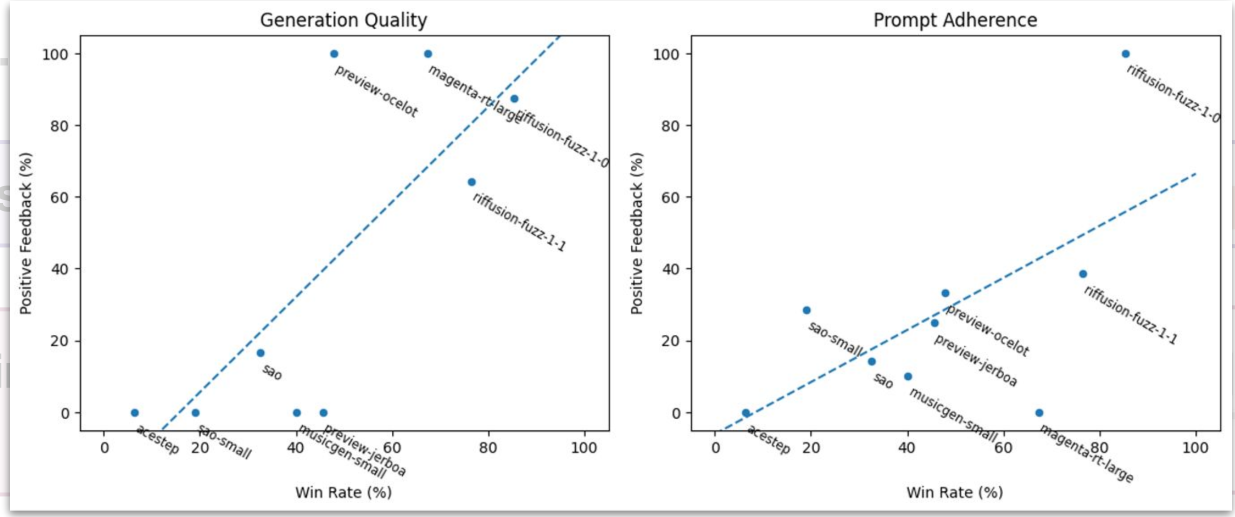
✍️ **Language Feedback**

Opportunity: Going beyond binary preferences

We collect ...

🗳️ Votes

👂 Listeni
Data



✍️ **Language Feedback**

What did you *like* about 🥇 Generated Music A (from Riffusion FUZZ 1.0)?

The music quality was much higher overall, though the style wasn't quite what I asked for

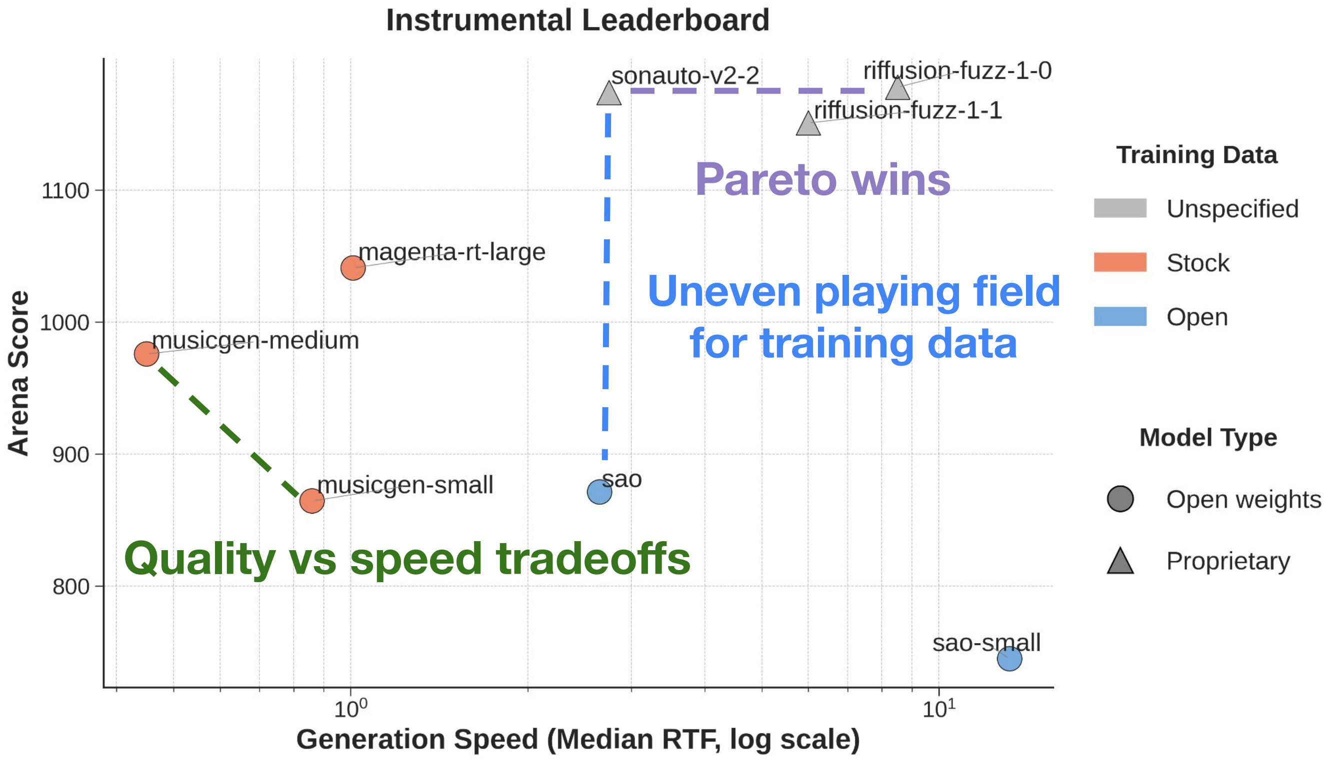
What did you *dislike* about 🥈 Generated Music B (from ACE-Step)?

The music was lower quality and shorter, though I liked the lyrics.

🌐Platform

NEURAL INFORMATION PROCESSING SYSTEMS

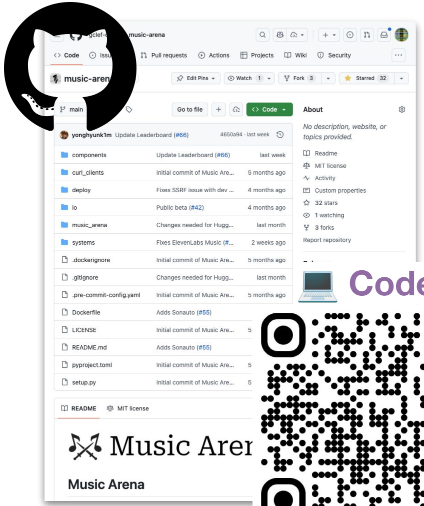
Surfacing holistic considerations beyond preference



Full Leaderboard

Increasing trust, privacy, and transparency

Open source



💻 Code

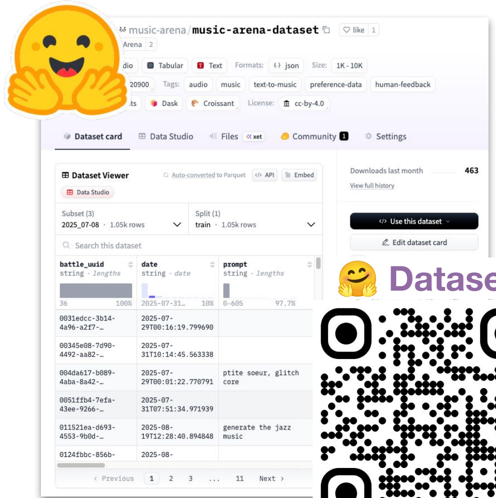
Pseudonymization

Hash (ID + Secret Salt)

f9b5 e8a8 ac2b

✅ Longitudinal analysis
✅ User privacy

Rolling data releases



🤗 Dataset

NeurIPS 2025 Creative AI Panel #2

Art Content Creation: When demands are met by pipelines (or not)

Music Arena: Live Evaluation for Text-to-Music

Yonghyun Kim[##]

Wayne Chi[#]

Yuki Mitsufuji[♭]

Chris Donahue[#]

🌐Platform

Anastasios N. Angelopoulos[♮]

Wei-Lin Chiang[♮]

Koichi Saito[♭]

Shinji Watanabe[#]

More recently:
Nathan Pruyne[#]

[#] CMU [♭] Sony AI [♮] LMArena [##] Georgia Institute of Technology

NEURAL INFORMATION PROCESSING SYSTEMS