

---

# Music Arena: Live Evaluation for Text-to-Music

---

Yonghyun Kim<sup>##</sup>

Wayne Chi<sup>b</sup>

Anastasios Angelopoulos<sup>b</sup>

Wei-Lin Chiang<sup>b</sup>

Koichi Saito<sup>#</sup>

Shinji Watanabe<sup>b</sup>

Yuki Mitsufuji<sup>#</sup>

Chris Donahue<sup>b</sup>

<sup>b</sup>Carnegie Mellon University   <sup>b</sup>LMarena   <sup>#</sup>Sony AI   <sup>##</sup>Georgia Tech

Music Arena is available at: <https://music-arena.org>

Preference data is available at: <https://huggingface.co/music-arena>

## Abstract

We present Music Arena, an open platform for scalable human preference evaluation of text-to-music (TTM) models. Soliciting human preferences via listening studies is the gold standard for evaluation in TTM, but these studies are expensive to conduct and difficult to compare, as study protocols may differ across systems. Moreover, human preferences might help researchers align their TTM systems or improve automatic evaluation metrics, but an open and renewable source of preferences does not currently exist. We aim to fill these gaps by offering *live* evaluation for TTM. In Music Arena, real-world users input text prompts of their choosing and compare outputs from two TTM systems, and their preferences are used to compile a leaderboard. While Music Arena follows recent evaluation trends in other AI domains, we also design it with key features tailored to music: an LLM-based routing system to navigate the heterogeneous type signatures of TTM systems, and the collection of *detailed* preferences including listening data and natural language feedback. We also propose a rolling data release policy with user privacy guarantees, providing a renewable source of preference data and increasing platform transparency. Through its standardized evaluation protocol, transparent data access policies, and music-specific features, Music Arena not only addresses key challenges in the TTM ecosystem but also demonstrates how live evaluation can be thoughtfully adapted to unique characteristics of specific AI domains.

## 1 Introduction

Text-to-music (TTM) generation has advanced rapidly in recent years, with models demonstrating remarkable capabilities in creating high-fidelity music audio from text prompts [1–5]. This progress has highlighted two critical and intertwined challenges for the research community. Firstly, designing *rigorous TTM evaluation* protocols is essential for navigating tradeoffs in methodologies and training data, and also to track progress over time. Secondly, identifying a source of *open and renewable human TTM preference data* would help researchers to better align TTM systems with human intent [6], and aid in the development of more reliable automatic evaluation metrics [7, 8].

The current TTM landscape is unable to meet these challenges. Music is a human endeavor, and a rigorous evaluation metric should thus reflect human preferences. However, human preferences are difficult to capture as they may be influenced more by subjective assessments of creativity than objective, quantifiable phenomena. While numerous automatic evaluation metrics have been proposed [8–10], past work shows they correlate imperfectly with human preferences [7, 8] and may not capture all key musical desiderata [8]. Moreover, while some open preference datasets have been released [7, 8, 11], these one time efforts are not *renewable* and will remain fixed even as new models

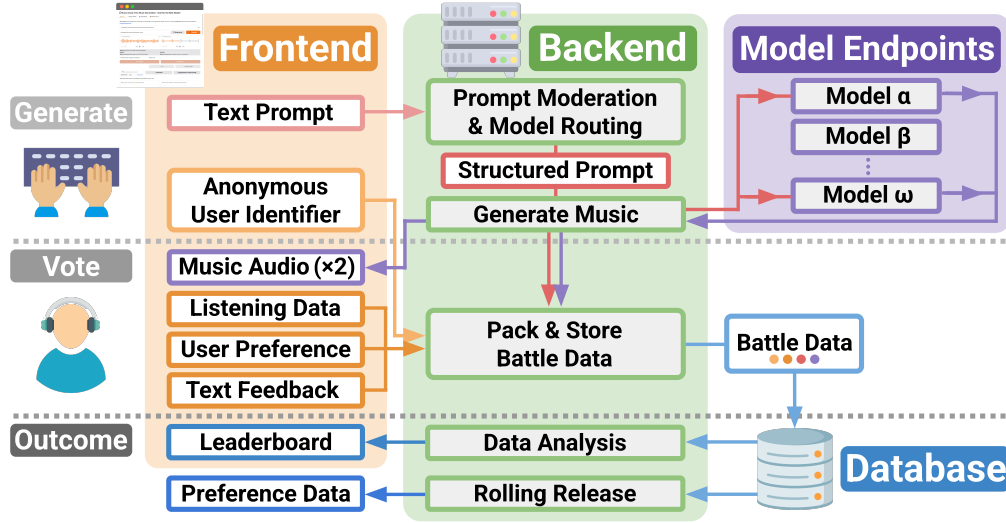


Figure 1: The Music Arena data lifecycle. On the **Frontend**, users engage in “battles”: they submit text prompts, listen to outputs from two music generation systems, and specify their preferences. The central **Backend** orchestrates the battles: it extracts structured information from text prompts using an LLM to determine model compatibility, routes prompts to appropriate **Model Endpoints** for generation, delivers music audio to users, and stores the resulting battle data in a **Database**. Collected data is used to compile a public leaderboard and publicly released on a recurring basis.

emerge or human preferences drift. Commercial providers may have access to renewable sources of preferences through their platform’s proprietary usage data, but this data is typically not *open*.

Human listening studies could potentially address both gaps, offering evaluation grounded in human preferences and a source of preference data, but current listening protocols lack *rigor*. Firstly, they are *inconsistent*—the meaning of metrics like win rates or mean opinion scores varies across numerous dimensions of ad hoc protocols including listening interfaces, models compared, and annotator distributions. Secondly, studies routinely cost hundreds or thousands of dollars on crowdsourcing platforms, making them *unscalable*. Finally, studies are *unrealistic* as users are presented with contrived listening scenarios that differ from real-world, self-motivated usage of TTM systems.

A new evaluation protocol, which we refer to here as *live evaluation*, has already helped navigate analogous challenges in other AI domains [12–16]. The key idea behind live evaluation is to align incentives by offering everyday users free access to generative AI systems in exchange for their preferences. On most live evaluation platforms, users first submit an input query, are presented with outputs from two different AI systems, and finally asked which of the two they prefer. These pairwise preferences are distilled into a global leaderboard ordered by Bradley-Terry coefficients [17] or related scores like Elo [18]. Compared to ad hoc human evaluation protocols, these scores are more *consistent* because they are calculated for all models from the same protocol and annotator distribution. Moreover, collecting preferences in this fashion is *scalable* because incentives are aligned, and the preference data reflects more *realistic* usage. Live evaluation was first proposed for language via Chatbot Arena [12], and subsequently explored for text-to-speech [13], image and video generation [14], 3D model generation [15], and coding assistance [16]. Here we propose to offer live evaluation for TTM.

Music presents unique challenges and opportunities for live evaluation relative to other AI domains. Firstly, music generation models have heterogeneous input and output type signatures: some models output vocals and may or may not accept user-specified lyrics, while other models output variable length audio and may or may not accept user-specified durations. We propose an LLM-based *prompt routing* system which adapts inputs on a simple unified user interface (a single input text box) to appropriate type signatures for different models. We also use this system to detect and reject malicious inputs from users, including references to copyrighted material and artists, or overtly inappropriate or harmful content. Secondly, unlike for other modalities like images, music must be consumed by users in real time, affording the opportunity to collect data on which portions of an output a user

observed before specifying their preferences. Accordingly, we collect *fine-grained* listening data consisting of timestamps for playback actions on both outputs, and also enforce that users listen to some non-trivial amount of each before submitting their preferences. These key features of Music Arena underscore the importance of tailoring live evaluation to the nuances of individual domains.

We also implement policies for Music Arena that aim to both increase platform trust and provide renewable access to preference data. Firstly, we anonymize private user information such as IP addresses by salting and hashing, ensuring *user privacy* while also facilitating *record linkage* for longitudinal preference research. Secondly, our entire platform is open source<sup>1</sup> aside from production configurations and secret keys. Finally, we propose to release data at *regular* (monthly) intervals, allowing the research community to access the latest data. By open sourcing our code and data, the entire lifecycle of our platform and leaderboard can be independently audited.

We release Music Arena to address key challenges in the landscape of TTM (rigorous evaluation, renewable preference data). Here, we present an overview of the Music Arena platform alongside an analysis of the preferences collected between July 28 and Aug 31, 2025, including 1,420 user-initiated battles from 373 unique users yielding 1,051 votes. From these votes, we build a preliminary **Music Arena leaderboard** (Appendix A), ranking contemporary TTM models via live evaluation.

## 2 Music Arena Platform Overview

Music Arena is a web-based live evaluation platform designed for the scalable collection of pairwise human preferences for TTM systems. The platform’s architecture is comprised of three core components (Figure 1): a user-facing **Frontend**, an intelligent **Backend** that orchestrates the entire generation and logging process, and **Model Endpoints** comprising the various TTM systems. These components are modular and communicate with one another via simple HTTP requests.

### 2.1 Frontend

Our frontend is a web-based interface built with Gradio [19] and serves as the primary means of user interaction. Upon their first visit, users are presented with a consent page detailing the IRB-approved research protocol and data handling policies. Once consent is provided, the main “Arena” interface allows users to engage in one or more “Battles” (pairwise comparisons). To initiate a battle, a user submits a text prompt of their choosing. Next, the user is presented with two audio tracks generated by two different TTM systems, and may listen to them in whatever order and for however long they like. To mitigate potential biases that could arise from differing audio lengths, the interface supports variable-length outputs but conceals the specific duration of each track from the user. After listening, users specify their preference (prefer A, prefer B, tie, both bad). Finally, after voting, the frontend reveals the identities of the competing models, along with other information such as generation speed. A download link for the preferred track is provided as an incentive for casting a decisive vote, and users are given the option to provide additional natural language feedback.

### 2.2 Backend

The backend is the central server-side component that acts as the main orchestrator for the platform. It receives all incoming requests from the frontend and is designed to handle numerous user sessions concurrently. Its core responsibility is to manage the entire lifecycle of each battle: it processes the user’s text prompt and dispatches generation tasks to two models in parallel. To mitigate bias from differing inference speeds, the backend waits for both models to complete and then delivers their audio tracks simultaneously—the actual generation time of each model is also logged on the backend. Finally, it collects the user’s preference data and ensures it is securely stored in the database.

### 2.3 Model Endpoints

To facilitate Music Arena, we aim to unify the heterogeneous type signatures of TTM systems. To this end, we implement a *model endpoint* for numerous TTM systems [2, 4, 5, 20–23]—code that adapts underlying type signatures and dependencies into a common interface. For open weights models, endpoints manage synchronous calls to GPU resources and batching for increased throughput. For

<sup>1</sup>Code is available at <https://github.com/gclef-cmu/music-arena>

API-based commercial systems, endpoints adapt inputs from our unified API to HTTP requests to proprietary APIs. To manage the varied (and sometimes conflicting) software dependencies across systems, each endpoint is paired with a bespoke Docker container. Each container exposes a simple API endpoint with a common type signature, allowing the backend to interact with all systems in a uniform manner. This modular architecture facilitates decentralized development—providers can contribute new TTM systems to Music Arena without disturbing other parts of the platform.

### 3 Key Features

Here we detail the methods and policies we propose in Music Arena, especially those distinct from other live evaluation platforms or ones that are specifically tailored for the music domain.

#### 3.1 LLM-based moderation and routing

The backend’s orchestration is powered by an LLM-based system that facilitates *moderation* of malicious input prompts, and *routing* of prompts on a unified interface (simple textbox) to relevant models. To moderate, the LLM is instructed to reject the user’s input prompt if it contains references to copyrighted musical material, culturally insensitive themes, or explicit themes, including profanity that would be atypical for the musical style (e.g., profanity okay for heavy metal, not okay for a nursery rhyme). For prompts that pass moderation, the LLM is instructed to extract structured information from the natural language input: the implied presence of vocals or lyrics (e.g., “folk song about a cat named Chamomile” implies lyrics), and explicitly-specified duration (e.g., “30 second lo-fi beat”). The backend uses this structured representation to seamlessly route prompts to a subset of models that support the user’s query (e.g., many models do not support vocal or lyrics generation). At time of writing, Music Arena uses OpenAI’s GPT-4o [24] for this component.

#### 3.2 Detailed preferences via listening data and language feedback

Most live evaluation platforms for other AI domains collect simple pairwise preferences. Here we propose to additionally collect more detailed preference signals including fine-grained listening data and natural language feedback. As the user listens to each generated audio during a battle, our system stores their listening behavior including the amount of time spent listening to each clip, and the wall clock time at which they played or paused each clip. To ensure meaningful user engagement, the voting interface is only unlocked after a user listens to each track for a predefined minimum duration (4 seconds at the time of writing). After a user specifies their preference between four options—*A is better*, *B is better*, *Tie*, or *Both are bad*—they are encouraged to provide additional natural language feedback, clarifying their rationale. We use listening data to better understand and model user behavior, and the language feedback to offer richer insights into preferences and desiderata than binary preferences alone can provide. An example of our detailed preference data per battle and an analysis of the initial data release appear in Appendix C and Appendix B, respectively.

#### 3.3 Reference TTM implementations

A key feature of Music Arena is the development of a unified Docker-based framework for managing inference from TTM systems as outlined in Section 2.3. In addition to supporting our core platform, we hope that this unified framework may benefit other research that requires comparing outputs from several TTM systems. At time of writing, we support the following open weights models: Meta’s MusicGen [4], Stability AI’s Stable Audio Open [5] and Stable Audio Open Small [25], SongGen [20], ACE Studio’s ACE-Step [21], and Google DeepMind’s Magenta RealTime [22]. We also support API-based commercial models including Producer.ai’s FUZZ models (1.0 & 1.1) [26], Stability AI’s Stable Audio 2.0 [27], and Google DeepMind’s Lyria RealTime [23]. Due to resource limitations, not all of these models will be available for live evaluation at a given time, however they can always be accessed by researchers running our code using their own resources.

Systems in this collection exhibit substantial heterogeneity in type signatures. Three support generating output vocals [20, 21, 26] while others are instrumental only. Commercial systems like Riffusion FUZZ [26] generate lyrics jointly with audio, while open weights systems like SongGen [20] and ACE-Step [21] require explicit lyrics conditioning— we use GPT-4o [24] to generate lyrics for these systems from a user’s input prompt. In addition to considerations around vocals, there is a long tail of

additional control signals across models, e.g., Stability AI models support explicit specification of output duration [5, 25, 27]. This heterogeneity is more pronounced in music, unlike the standardized signatures of chat [12] and image generation [14]. We design our unified framework to navigate this complex landscape, and aim to extend it in future work to support even broader music type signatures beyond TTM such as style transfer or symbolic music generation.

## 4 Key Policies

Along with the key *features* outlined previously, here we emphasize key *policies* of Music Arena designed to increase platform trust by surfacing considerations beyond user preferences, promoting user privacy and platform transparency, and providing renewable access to preference data.

### 4.1 Surfacing model considerations beyond Arena Score

The primary metric of interest across existing live evaluation platforms in other domains is an “Arena Score”, derived from pairwise comparisons. In music, holistic comparisons may require considerations of factors beyond preferences. Accordingly, our leaderboard (Appendix A) also surfaces information on **training data** and **generation speed**. Including information on training data acknowledges the legal, ethical, and quality implications of the “uneven playing field” for training data across model providers. Reporting generation speed (measured by median RTF) recognizes that some models may trade off quality for speed, e.g., to facilitate low-latency creative workflows.

### 4.2 Protecting user privacy while facilitating record linkage

The ability to perform *record linkage* across Music Arena sessions—identifying multiple battles from the same user—is critical for longitudinal preference analysis and detecting spam or malicious behavior. However, it is essential that we also protect user privacy by ensuring that personally identifying information is never exposed. Following established recommendations for user privacy in research [28], we implement a pseudonymization protocol using salted hashing [29]. When users interact with Music Arena, we transform linkable identifiers such as IP addresses by applying a server-side salt (a secret random string) followed by a one-way cryptographic hash. We only store these anonymized user identifiers, never the original identifiers. This approach provides strong privacy guarantees and protects against de-anonymization strategies such as rainbow table attacks. The resulting anonymized identifiers allow anyone to link battles from the same user across sessions for research purposes without ever exposing their private identifiers.

### 4.3 Maximizing platform transparency and data access

We are committed to making Music Arena as *transparent* as possible. To this end, all of our platform code is **open source**, aside from secret keys for private salting and API access. Additionally, we commit to a policy of **rolling, comprehensive data releases**. Unlike preferences from one-time data collection efforts [7, 8, 11], we aim to publish Music Arena data at regular monthly intervals. This rolling approach is critical in the rapidly evolving field of generative AI, addressing key sources of distribution shift like the development of new TTM systems, and changes in user preferences over time. Moreover, we aim for our data releases to be comprehensive,<sup>2</sup> including anonymized user identifiers, generated audio, and detailed preferences. This combination of open code and data allows the research community to audit our platform’s entire lifecycle and evaluation results.

## 5 Ethical Considerations and Safeguards

The design and operation of Music Arena are guided by principles of ethical research, user privacy, and responsible AI development. All research activities involving human subjects in this study were approved by the IRB at Carnegie Mellon University under Protocol ID STUDY2024\_00000489. We have implemented several key safeguards in Music Arena to uphold these principles.

---

<sup>2</sup>Minor exceptions to this policy may apply, e.g., some model licenses prevent the release of generated audio.

**Informed consent.** Before any interaction, users are presented with a consent page that transparently outlines the study’s objectives, data collection methods, and our commitment to public data release. Explicit, informed consent is required to participate.

**User privacy.** Music Arena does not store personally identifiable information, such as raw IP addresses. Instead, we collect anonymized identifiers through salting and hashing. Additionally, users consent that they will not upload private information in their text prompts or language feedback.

**Content moderation.** To mitigate risks of harmful or infringing content creation through Music Arena, every user-submitted prompt is first processed by our LLM-based moderation pipeline.

There remain ethical considerations for our work beyond these safeguards. Music Arena provides increased access to TTM systems for everyday users, which could have long term psychological effects. Our user distribution will likely be skewed to US users and AI enthusiasts, potentially promoting increased focus to the needs of those user populations by model providers. Music Arena’s current focus on text-to-music may inadvertently steer the research community’s attention away from other important tasks like style transfer or symbolic generation. Music Arena also inherits many ethical and societal considerations from music generation more broadly. Music generation may change the economic landscape of music labor, accelerate the commodification of music, or contribute to the homogenization of music cultures. Overall, we believe the benefits of Music Arena (more rigorous and transparent evaluation, open availability of preference data) outweigh the risks.

## 6 Limitations and Future Work

Our work has several limitations that present clear avenues for future improvement. On the frontend, our ability to precisely track which segment of the audio a user is listening to is constrained by the user interface—we track total listening duration but cannot capture seeking actions within audio clips. Our current backend system also selects pairs of TTM systems uniformly at random, rather than more principled strategies [16] that navigate tradeoffs around quality, speed, and coverage of relevant pairs. Furthermore, the scope of Music Arena is currently limited to text-to-music generation, excluding other important tasks like symbolic generation or style transfer. Finally, as a public web platform, our user base is not representative of the global population, and the long-term sustainability of providing free access to self-hosted open-weights models remains a challenge.

Our plans for future work aim to address these limitations, improve our understanding of human musical preferences, and contribute to the science of live evaluation. We aim to continue refining our frontend and backend—a particular direction of interest is improving the backend pair selection algorithm to better balance tradeoffs around leaderboard fidelity and user experience. We will leverage the growing preference dataset to better understand strengths and weaknesses of specific models and perform meta-evaluation against automatic evaluation metrics. Through analysis of natural language feedback and live evaluation of controlled degradations of systems (e.g., adding latency or noise to a system), we may better understand which attributes users consider most prominently when making preference decisions. As creative workflows mature, we hope to integrate live evaluation directly into user workflows [16]. Finally, we will continuously refine our evaluation methodology based on community feedback to ensure the long-term rigor and fairness of our platform.

## 7 Conclusion

We present Music Arena, a live evaluation platform that addresses critical gaps in text-to-music evaluation through scalable human preference collection and transparent data releases. Our platform introduces key innovations tailored specifically for music: an LLM-based system that enables content moderation and intelligent routing across heterogeneous model type signatures, a detailed preference collection methodology that captures fine-grained listening behaviors and natural language feedback, and a commitment to open science through comprehensive rolling data releases and full lifecycle auditability. By aligning user incentives with research needs, Music Arena provides the community with both a standardized human evaluation protocol and a renewable dataset of human musical preferences that reflects real-world usage patterns. As text-to-music generation continues to advance rapidly, Music Arena establishes a foundation for rigorous evaluation that can evolve alongside the field, supporting researchers in building more aligned systems while maintaining transparency and ethical standards that respect both user privacy and the broader implications of AI-generated music.

## Acknowledgments and Disclosure of Funding

The development of Music Arena was supported by Sony AI. We extend our sincere thanks to our commercial contacts at Producer.ai, Stability AI, Google DeepMind, and Suno for productive discussions that informed the key features and policies of Music Arena.

## References

- [1] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv:2005.00341*, 2020.
- [2] Seth Forsgren and Hayk Martiros. Riffusion - stable diffusion for real-time music generation. 2022. URL <https://riffusion.com/about>.
- [3] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. MusicLM: Generating music from text. *arXiv:2301.11325*, 2023.
- [4] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *NeurIPS*, 2023.
- [5] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable Audio Open. In *ICASSP*, 2025.
- [6] Geoffrey Cideron, Sertan Girgin, Mauro Verzetti, Damien Vincent, Matej Kastelic, Zalán Borsos, Brian McWilliams, Victor Ungureanu, Olivier Bachem, Olivier Pietquin, et al. MusicRL: Aligning music generation to human preferences. *arXiv:2402.04229*, 2024.
- [7] Florian Grötschla, Ahmet Solak, Luca A Lanzendörfer, and Roger Wattenhofer. Benchmarking music generation models and metrics via human preference studies. In *ICASSP*, 2025.
- [8] Yichen Huang, Zachary Novack, Koichi Saito, Jiatong Shi, Shinji Watanabe, Yuki Mitsufuji, John Thickstun, and Chris Donahue. Aligning text-to-music evaluation with human preferences. In *ISMIR*, 2025.
- [9] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, 2019.
- [10] Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting frechet audio distance for generative music evaluation. In *ICASSP*, 2024.
- [11] Cheng Liu, Hui Wang, Jinghua Zhao, Shiwan Zhao, Hui Bu, Xin Xu, Jiaming Zhou, Haoqin Sun, and Yong Qin. MusicEval: A generative music dataset with expert ratings for automatic text-to-music evaluation. In *ICASSP*, 2025.
- [12] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *ICML*, 2024.
- [13] mrfakename, Vaibhav Srivastav, Clémentine Fourier, Lucain Pouget, Yoach Lacombe, main, Sanchit Gandhi, Apolinário Passos, and Pedro Cuenca. TTS Arena 2.0: Benchmarking text-to-speech models in the wild. <https://huggingface.co/spaces/TTS-AGI/TTS-Arena-V2>, 2025.
- [14] Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhui Chen. GenAI Arena: An open evaluation platform for generative models. In *NeurIPS*, 2024.
- [15] Dylan Ebert. 3D Arena: An open platform for generative 3d evaluation. *arXiv:2506.18787*, 2025.
- [16] Wayne Chi, Valerie Chen, Anastasios Nikolas Angelopoulos, Wei-Lin Chiang, Aditya Mittal, Naman Jain, Tianjun Zhang, Ion Stoica, Chris Donahue, and Ameet Talwalkar. Copilot Arena: A platform for code LLM evaluation in the wild. In *ICML*, 2025.

- [17] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 1952.
- [18] Arpad E Elo. The proposed USCF rating system, its development, theory, and applications. *Chess Life*, 1967.
- [19] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ML models in the wild. *arXiv:1906.02569*, 2019.
- [20] Zihan Liu, Shuangrui Ding, Zhixiong Zhang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. SongGen: A single stage auto-regressive transformer for text-to-song generation. In *ICML*, 2025.
- [21] Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo. ACE-Step: A step towards music generation foundation model. *arXiv:2506.00045*, 2025.
- [22] Google DeepMind. Magenta RealTime. 2025. URL <https://magenta.withgoogle.com/magenta-realtime>.
- [23] Google DeepMind. Lyria RealTime. 2025. URL <https://magenta.withgoogle.com/lyria-realtime>.
- [24] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv:2410.21276*, 2024.
- [25] Zachary Novack, Zach Evans, Zack Zukowski, Josiah Taylor, CJ Carr, Julian Parker, Adnan Al-Sinan, Gian Marco Iodice, Julian McAuley, Taylor Berg-Kirkpatrick, et al. Fast text-to-audio generation with adversarial post-training. *arXiv:2505.08175*, 2025.
- [26] Riffusion Team. Riffusion FUZZ: State-of-the-art diffusion transformer for creating and editing music, 2025. URL <https://riffusion.com>.
- [27] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Long-form music generation with latent diffusion. *arXiv:2404.10301*, 2024.
- [28] Omer Tene and Jules Polonetsky. Privacy in the age of big data: a time for big decisions. *Stanford Law Review Online*, 2012.
- [29] Clete A Kushida, Deborah A Nichols, Rik Jadrnicek, Ric Miller, James K Walsh, and Kara Griffin. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, 2012.

Table 1: The first Music Arena leaderboard results (July 28 – Aug 31, 2025), segmented by Instrumental and Vocal models. The table includes Arena Scores (with 95% confidence intervals), vote counts, generation speed (RTF), and key metadata for each model.

### Instrumental Music Generation

| Rank | Model              | Arena Score | 95% CI        | # Votes | Generation Speed (RTF) | Organization    | License        | Training Data | Supports Lyrics | Access       |
|------|--------------------|-------------|---------------|---------|------------------------|-----------------|----------------|---------------|-----------------|--------------|
| 1    | riffusion-fuzz-1-1 | 1250.8      | +52.0 / -45.5 | 252     | 6.01                   | Producer.ai     | Closed         | Unspecified   | True            | Proprietary  |
| 2    | magenta-rt-large   | 1113.6      | +56.5 / -57.2 | 276     | 1.01                   | Google DeepMind | Apache 2.0     | Stock         | False           | Open weights |
| 3    | musicgen-small     | 928.5       | +40.4 / -46.7 | 278     | 0.86                   | Meta            | CC-BY-NC 4.0   | Stock         | False           | Open weights |
| 4    | sao                | 924.7       | +45.7 / -41.5 | 286     | 2.63                   | Stability AI    | STAI Community | Open          | False           | Open weights |
| 5    | sao-small          | 782.4       | +50.9 / -62.2 | 292     | 12.79                  | Stability AI    | STAI Community | Open          | False           | Open weights |

### Vocal Music Generation

| Rank | Model              | Arena Score | 95% CI         | # Votes | Generation Speed (RTF) | Organization | License    | Training Data | Supports Lyrics | Access       |
|------|--------------------|-------------|----------------|---------|------------------------|--------------|------------|---------------|-----------------|--------------|
| 1    | riffusion-fuzz-1-0 | 1172.5      | +99.1 / -62.7  | 144     | 5.60                   | Producer.ai  | Closed     | Unspecified   | True            | Proprietary  |
| 2    | riffusion-fuzz-1-1 | 1087.3      | +40.8 / -47.2  | 218     | 5.25                   | Producer.ai  | Closed     | Unspecified   | True            | Proprietary  |
| 3    | preview-ocelot     | 1045.7      | +75.9 / -82.9  | 90      | 5.42                   | Hidden       | Closed     | Unspecified   | True            | Proprietary  |
| 4    | preview-jerboa     | 1034.4      | +92.6 / -80.8  | 88      | 5.61                   | Hidden       | Closed     | Unspecified   | True            | Proprietary  |
| 5    | acestep            | 660.1       | +75.5 / -121.3 | 178     | 2.89                   | ACE Studio   | Apache 2.0 | Unspecified   | True            | Open weights |

## A Leaderboard structure

Music Arena was launched on July 28, 2025, and after collecting 1,051 user votes by the end of August, we released the first public leaderboard on September 19, 2025. Here, we provide an overview of the leaderboard’s structure, which is designed to address the evaluation challenges in the music domain. The detailed results of this initial leaderboard are presented in Table 1, and the key tradeoffs are visualized in Figure 2.

Public leaderboards for live evaluation platforms [12–14, 16] tend to contain similar attributes: an overall “Arena Score” (usually derived from the Bradley-Terry model [17]), the number of votes, the model provider, and the system license. There are a number of unique considerations in music that motivate presentation of additional attributes.

For LLM training, it can be broadly assumed that all providers are training on large-scale text data mined from the web. However, in music, we see significantly more diversity in training data across models. For example, some models are trained on licensed stock music [4] or publicly-available music under Creative Commons licenses [5], while some only specify the quantity (rather than the provenance) of their training data [21]. These differences induce an uneven playing field for training data, affecting not only a system’s performance in pairwise comparisons but also its standing within broader legal and ethical conversations. Accordingly, we include on our leaderboard a summary of available *training data information* for each model, including provenance and quantity.

In addition to training data, we include *generation speed* on our leaderboard. Tools that facilitate the creation of music are often designed to have low latency interactions. Accordingly, TTM providers may make different tradeoffs between quality and speed, depending on if they are targeting more consumption-oriented (prefer quality) or creative-oriented (prefer speed) applications. Speed is codified on our leaderboard by median real-time factor (RTF), where RTF is the ratio of seconds of music generated divided by seconds of wall clock time to generate. For example, a system that generates 30 seconds of audio in 3 seconds has an RTF of 10x. Measuring speed via RTF gracefully handles variable length outputs, ensuring that systems are not punished for taking more time to generate longer audio.

The leaderboard interface allows users to view results segmented by model capability (e.g., “Instrumental” vs. “Vocal” tabs), allowing viewers to make principled decisions about models based on attributes of keen importance to their specific music application goals. Moreover, we emphasize these music-specific tradeoffs through visualization: a 2D scatter plot with speed on the X axis, Arena Score on the Y axis, and colors and shapes to emphasize training data and licensing information.

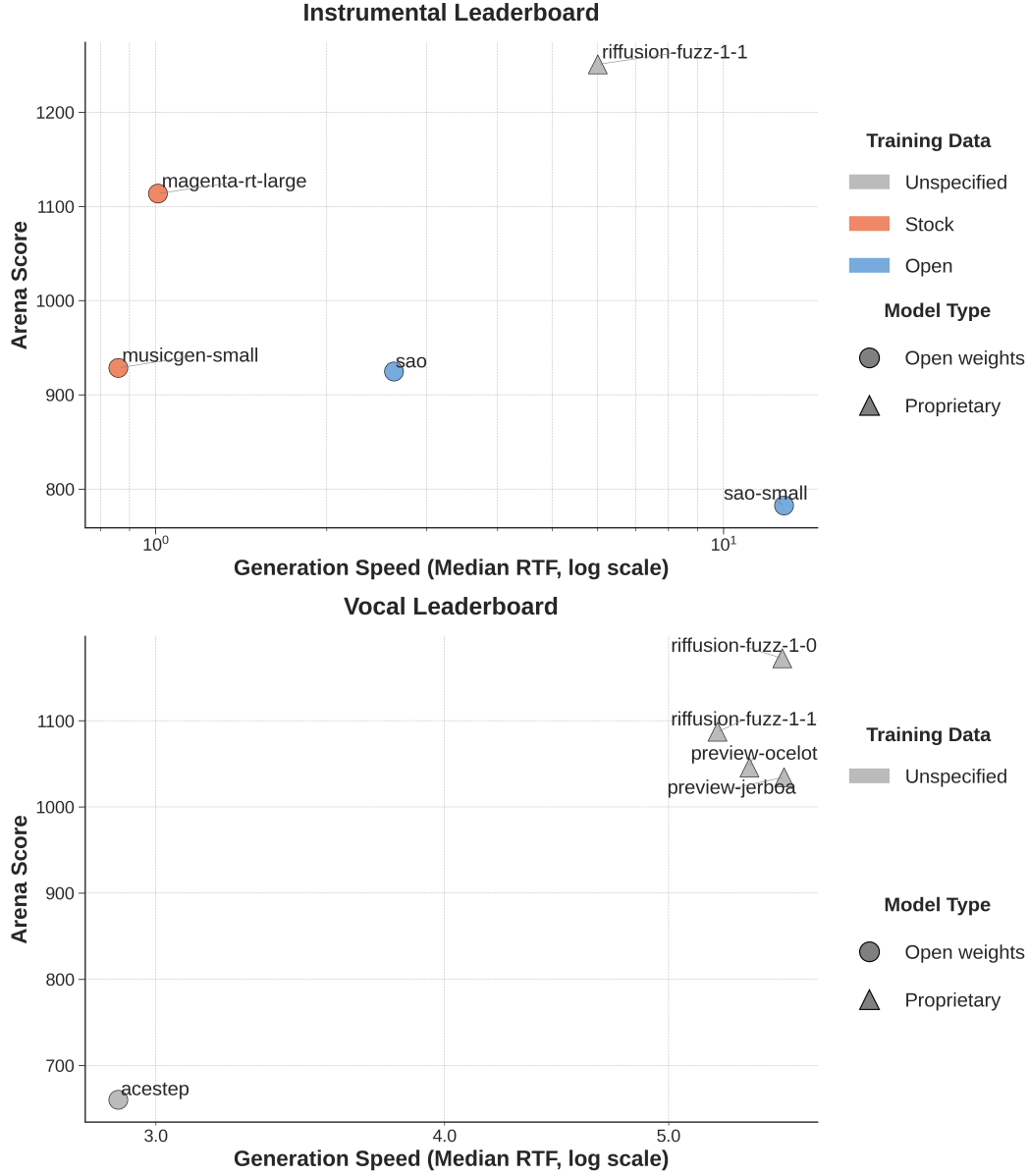


Figure 2: Music Arena leaderboard (July 28 – Aug 31, 2025), plotting Arena Score (Y-axis) against Generation Speed (Median RTF, X-axis, log scale). Colors and shapes distinguish models by their training data and access type (open weights/proprietary), respectively. This visualization emphasizes the key tradeoff between model quality (score) and interactive latency (speed), an important consideration for creative music applications.

## B Analysis of Initial Data Release

This section provides a more detailed look into user engagement patterns, prompt characteristics, and the types of music users create on the Music Arena platform.

### B.1 User Engagement Distribution

Our analysis shows a long-tail distribution of user engagement. While many users contribute a small number of votes, a dedicated group of “power users” is responsible for a significant portion of the data. Table 2 shows the distribution of votes submitted per user from our initial data release.

Table 2: Distribution of votes per user, based on 1,051 valid votes from 373 unique users.

| Number of Votes | Number of Users |
|-----------------|-----------------|
| 1               | 193             |
| 2               | 72              |
| 3               | 44              |
| 4               | 24              |
| 5               | 8               |
| 6-10            | 18              |
| 11-20           | 10              |
| 21-50           | 4               |

## B.2 Prompt Descriptiveness

Analyzing the 804 user-written prompts from valid, voted-on battles reveals they are typically concise. As shown in Table 3, the raw data shows a median prompt length of 7 words, but the mean (18.68) is heavily skewed by a long tail of very descriptive prompts (max 1000 words). To get a more accurate picture of typical behavior, we removed 82 extreme outliers using the IQR method (threshold at 33 words). Table 3 shows that the statistics for the remaining 722 prompts are much more focused, with a median length of 6 words and a mean of 8.27. Figure 3 visualizes this post-filtered distribution, confirming that the typical Music Arena user prefers to express their creative ideas in a few words.

Table 3: Descriptive statistics for user prompt lengths (in words), before and after Interquartile Range (IQR) outlier removal.

| Metric       | Raw Data (804 prompts) | Post-IQR (722 prompts) |
|--------------|------------------------|------------------------|
| count        | 804                    | 722                    |
| mean         | 18.68                  | 8.27                   |
| std          | 54.82                  | 6.87                   |
| min          | 1                      | 1                      |
| 50% (median) | 7                      | 6                      |
| max          | 1000                   | 33                     |

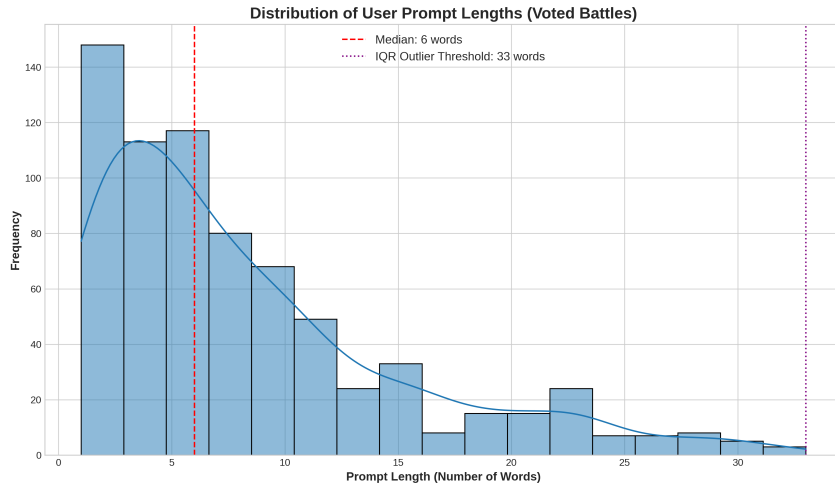


Figure 3: Distribution of user prompt lengths from voted battles (after removing outliers)

## B.3 Commonly Requested Musical Concepts

By analyzing the keywords in 804 user-written prompts, we identified the most common musical genres, instruments, and moods requested by users. Table 4 lists the most frequent keywords, and Figure 4 visualizes their prominence.



## 🎵 Music Arena: Free Music Generation - Vote for the Best Model!

[Arena](#) [Specific Model](#) [Leaderboard](#) [About & Terms](#)

Music Arena is a free platform for evaluating and comparing AI music generation models. We are currently in private beta. Please email any feedback to [musicarena@cmu.edu](mailto:musicarena@cmu.edu).

[Surprise me](#)[Generate](#)

Generated Music A



1x

⏮ ⏪ ⏩ ⏭

Generated Music B



1x

⏮ ⏪ ⏩ ⏭

**Battle ID:** dc8513ba-f75e-4762-bd2c-76d364495b15  
**Model A:** Riffusion FUZZ 1.0 ( riffusion-fuzz-1-0 )  
**Stats:** Generated 192.5s of music in 23.9s (8.0x real time 🐌)

**Model B:** ACE-Step ( acestep )  
**Stats:** Generated 30.0s of music in 9.1s (3.3x real time 🐌)

👍 Thank you for voting! You voted for 🐌 Riffusion FUZZ 1.0 in favor of 🐌 ACE-Step.

A is better 🐌

B is better 🐌

Tie 🐌

Both are bad 🐌

Download your preferred music!

original-ca4846e4267411137... .mp3 [4.0 MB ↓](#)

Start Over

Regenerate w/ Same Prompt

(Optional) Please share more feedback on this battle and your vote!

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.

Any additional feedback on this battle or Music Arena in general?  

I enjoyed listening to both clips! It was a close call.

Feedback submitted!

Figure 5: An example of a completed user battle in the Music Arena frontend.

```
"create_time": 1753572627.3779469,
"frontend_git_hash": "4138a182e618f2e7687e4d34bf039cff42275f1e:dirty",
"ack_tos": "c81b3d54ff3f196eaae354e5317dc6e7",
"new_battle_times": [
  1753572628.2408764,
  1753572653.6583097
],
},
"prompt_prebaked": false,
"prompt_routed": true,
"a_audio_url": "...",
"a_metadata": {
  "system_key": {
    "system_tag": "riffusion-fuzz-1-0",
    "variant_tag": "initial"
  },
  "system_git_hash": "1952210249ad28dad600013ccf0ce13130165ca5:dirty",
  "system_time_queued": 1753572662.6410472,
  "system_time_started": 1753572662.6510143,
  "system_time_completed": 1753572686.5897048,
  "gateway_time_started": 1753572661.6347294,
  "gateway_time_completed": 1753572696.3290386,
  "gateway_num_retries": 0,
  "size_bytes": 4167243,
```

13

```

    "lyrics": "[Verse 1]\nAnother blackboard on the wall, spitting names and all,
(la da da)\nI lace my boots but no one calls, just numbers falling,
crawling\nStale beer breath and hate mail\u2014louder than the praise
fails\nScored by some stranger\u2019s scale, (oh oh) chalk dust in the
veins\n[Pre-Chorus]\nWe shout and shake, never for their grades\u2014\nRunning
ragged, never break, chasing after the parade\n[Chorus]\nThis is the Music
Arena\u2014don\u2019t care if you rate us!\nWe bleed out our voices, tearing the
silence\nRaise up our riot, sing for the lost ones\nWhat\u2019s your arena? Ours
is defiance! (Oi! Oi!)\n[Verse 2]\nFour strings snapping, critics clapping
(ha!)\nNotes like knuckles, fists overlapping\nPushed through the carnage,
nobody clean\nYet we shout \u201cViva la m\u00fasica!\u201d\u2014lightning
in-between\n[Breakdown]\n(La la la, la la, hey!)\nNames on the scoreboard,
erased quick as drawn\nWe\u2019re not your numbers\u2014we live on and on!\n(Oh
oh oh, let the echoes run)\n[Chorus]\nThis is the Music Arena\u2014don\u2019t
care if you rate us!\nWe bleed out our voices, tearing the silence\nRaise up our
riot, sing for the lost ones\nWhat\u2019s your arena? Ours is defiance! (Oi!
Oi!)\n[Bridge]\nEntre gritos y cerveza, saltamos sin pena\nTu marca no pesa, en
mi condena\n(Whoa-oh, whoa-oh, vamos!)\n[Solo]\n[Electric guitar and tin
whistle]\n[Chorus]\nThis is the Music Arena\u2014don\u2019t care if you rate
us!\nWe bleed out our voices, tearing the silence\nRaise up our riot, sing for
the lost ones\nWhat\u2019s your arena? Ours is defiance! (Oi! Oi!)\n[Outro]\n(La
la la la, la la la)\nErase us, replace us\u2014we\u2019re loud \u2018til
we\u2019re gone.",
    "sample_rate": 44100,
    "num_channels": 2,
    "duration": 192.496327,
    "checksum": "0008ee5d14cc4bafb3f8ec2fa26f4784"
  },
  "b_audio_url": "...",
  "b_metadata": {
    "system_key": {
      "system_tag": "acestep",
      "variant_tag": "initial"
    },
    "system_git_hash": "1952210249ad28dad600013ccf0ce13130165ca5:dirty",
    "system_time_queued": 1753572655.0246835,
    "system_time_started": 1753572660.6987517,
    "system_time_completed": 1753572669.7986672,
    "gateway_time_started": 1753572655.0225708,
    "gateway_time_completed": 1753572671.424467,
    "gateway_num_retries": 0,
    "size_bytes": 735168,
    "lyrics": "Welcome to the Music Arena, where the notes collide, \nWith the
bagpipes wailin' and the fiddles in stride, \nThere are tunes to judge and
rhythms to compare, \nIn this grand ol' place where musics declare. \n\nRaise
a pint in the bar where the voices unite, \nFor in Music Arena, we measure the
heights, \nOf melodies and harmonies, bold and true, \nBringing joy to the
hearts of both me and you. \n\nJoin the chorus loud, in the thunder and the
cheers, \nMusic Arena's call will ring through the years. \nA place where the
music is set free, \nIn the heart of it all, where we want to be. ",
    "sample_rate": 48000,
    "num_channels": 2,
    "duration": 29.952,
    "checksum": "24b1af8031278bb85e126f9e0bb11028"
  },
  "vote": {
    "a_listen_data": [
      [
        "PLAY",

```

```

    1753572708.6986423
  ],
  [
    "TICK",
    1753572709.9190872
  ],
  ...,
  [
    "TICK",
    1753572729.919615
  ],
  [
    "PAUSE",
    1753572731.188438
  ],
  [
    "TICK",
    1753572731.2903912
  ],
  ...,
  [
    "TICK",
    1753572736.7407818
  ],
  [
    "PLAY",
    1753572763.5559134
  ],
  [
    "PAUSE",
    1753572789.6293015
  ]
],
"b_listen_data": [
  [
    "PLAY",
    1753572731.9962952
  ],
  [
    "TICK",
    1753572733.203671
  ],
  ...,
  [
    "TICK",
    1753572736.2387252
  ],
  [
    "PAUSE",
    1753572761.9931803
  ],
  [
    "PLAY",
    1753572762.144039
  ],
  [
    "PAUSE",
    1753572762.799093
  ]
]

```

```

    ],
    "preference": "A",
    "preference_time": 1753572791.0873723,
    "feedback": "I enjoyed listening to both clips! It was a close call.",
    "a_feedback": "The music quality was much higher overall, though the style
wasn't quite what I asked for",
    "b_feedback": "The music was lower quality and shorter, though I liked the
lyrics.",
    "feedback_time": 1753572842.6993084
  },
  "vote_user": {
    "ip": null,
    "salted_ip": "d15300d2f8f7a122a14793494c85057d",
    "fingerprint": null,
    "salted_fingerprint": null
  },
  "vote_session": {
    "uuid": "42a03157-e3dc-4f00-8a59-1cdc2c221527",
    "create_time": 1753572627.3779469,
    "frontend_git_hash": "4138a182e618f2e7687e4d34bf039cff42275f1e:dirty",
    "ack_tos": "c81b3d54ff3f196eaae354e5317dc6e7",
    "new_battle_times": [
      1753572628.2408764,
      1753572653.6583097
    ]
  },
  "timings": [
    [
      "parse",
      1753572653.815334
    ],
    [
      "generate",
      1753572653.815521
    ],
    [
      "route",
      1753572653.815524
    ],
    [
      "sample_pair",
      1753572655.0172503
    ],
    [
      "generate_parallel_start",
      1753572655.0173497
    ],
    [
      "health_check_riffusion-fuzz-1-0:initial_start",
      1753572655.0174298
    ],
    [
      "health_check_acestep:initial_start",
      1753572655.0176365
    ],
    [
      "health_check_acestep:initial_end",
      1753572655.0225341
    ]
  ],

```

```

[
  "generate_acestep:initial_start",
  1753572655.0225701
],
[
  "health_check_riffusion-fuzz-1-0:initial_end",
  1753572661.634688
],
[
  "generate_riffusion-fuzz-1-0:initial_start",
  1753572661.6347291
],
[
  "generate_acestep:initial_end",
  1753572671.4905503
],
[
  "generate_riffusion-fuzz-1-0:initial_end",
  1753572696.4137614
],
[
  "generate_parallel_end",
  1753572696.4139218
],
[
  "create_battle_obj",
  1753572696.4139223
],
[
  "upload_audio",
  1753572696.415068
],
[
  "upload_metadata",
  1753572697.0474696
],
[
  "vote",
  1753572791.2476099
],
[
  "vote",
  1753572842.7061708
]
]
}

```

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The claims in the abstract and Section 1 accurately reflect both the platform's contributions (detailed in Sections 2 and 3) and the initial data analysis results (presented in Appendices A and B).

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Limitations are discussed in Section 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: N/A

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our paper presents initial findings from our live evaluation platform. The data collection and analysis methodology are described in Section 2, Appendix B, and Appendix A. We commit to regular public data releases, and our platform code is open-source, allowing for full auditability of our results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide full access to our platform as open source code aside from production configurations, API keys, and secret keys. We commit to regular data releases for preferences collected on our platform. Code is here and also included in a footnote in our paper: <https://github.com/gclef-cmu/music-arena>.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The details of our data collection methodology (the “experimental setting”) are described in Section 2 and Section 3. The data analysis details (e.g., outlier removal, metrics) are presented in Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report 95% confidence intervals for the Arena Scores in our leaderboard results (Table 1).

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The compute resources used to host the models (which enables our experiment) are described: Open weights models on our platform are served by four dedicated GPUs, and proprietary models are served by the private infrastructure of those companies.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Participants are not paid, instead, they are provided with free access to text-to-music systems in exchange for their preferences. Participation is completely voluntary and users must indicate informed consent before participating. We release data under permissive licenses while respecting the license terms of models served on the platform. We take user privacy seriously and adopt policies that ensure that private or de-anonymizing information is never released, or even collected in our own database. We discuss limitations around the diversity of our users. We propose to indicate training data sources on our leaderboard, to emphasize the uneven playing field for training data across model providers. Our study is fully approved under the Institutional Review Board at Carnegie Mellon University. All of this information is available in the main paper in Sections 4.2 and 5.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[Yes\]](#)

Justification: We discuss the positive impacts to TTM research in Section 1. We discuss potential negative ramifications in Section 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [\[Yes\]](#)

Justification: We take measures to preserve user privacy in our regular data releases, outlined in Section 4.2. We discuss ethical safeguards in Section 5

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

#### 12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We discuss all models we evaluate in Section 3.3.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Our paper documents the proposed features of our new platform, leaderboard, and data releases throughout. We release our entire platform code as open source.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Our full instructions and evaluation interface are publicly available on our platform website: <https://music-arena.org>. A screenshot is in Appendix C. We do not pay participants—instead their participation is entirely voluntary and they receive access to models in exchange.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

**15. Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: Risks to subject participants are discussed in Section 5. Our platform is approved by the Institutional Review Board at Carnegie Mellon University under identifier STUDY2024\_00000489.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

**16. Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We describe our usage of LLMs on our platform in Section 3.1.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.