Benchmarking Music Generation Models and Metrics via Human Preference Studies

Ahmet Solak[*]
ETH Zurich
asolak@ethz.ch

Florian Grötschla[*]
ETH Zurich
fgroetschla@ethz.ch

Luca A. Lanzendörfer
ETH Zurich
lanzendoerfer@ethz.ch

Roger Wattenhofer
ETH Zurich
wattenhofer@ethz.ch

Abstract

Recent advancements have brought generated music closer to human-created compositions, yet evaluating these models remains challenging. While human preference is the gold standard for assessing quality, translating these subjective judgments into objective metrics, particularly for text-audio alignment and music quality, has proven difficult. In this work, we generate 6k songs using 12 state-of-the-art models and conduct a survey of 15k pairwise audio comparisons with 2.5k human participants to evaluate the correlation between human preferences and widely used metrics. To the best of our knowledge, this work is the first to rank current state-of-the-art music generation models and metrics based on human preference. To further the field of subjective metric evaluation, we provide open access to our dataset of generated music and human evaluations.[1]

1 Introduction

The field of AI-generated music has witnessed unprecedented progress, with recent models producing compositions that are becoming increasingly indistinguishable from those created by humans. As these advancements continue, the evaluation of AI-generated music becomes even more relevant. While human preference remains the gold standard for assessing the quality and effectiveness of these models, translating these subjective judgments into reliable, objective metrics remains an open challenge. Efforts to bridge this gap have largely focused on two key aspects: (1) The quality of alignment between the text prompt and the audio and (2) the overall quality of the generated music. Despite the development of various objective metrics to evaluate these aspects, their effectiveness in capturing human preferences remains uncertain. In this

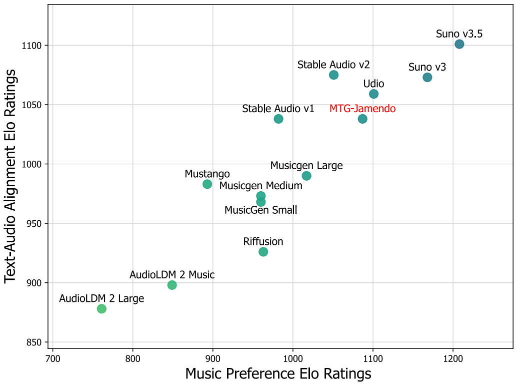


Figure 1: Elo ratings for all music generation models in the music preference and text-audio alignment human evaluation experiments. The reference dataset is shown in red.

[*] Equal contribution

[1] https://huggingface.co/datasets/disco-eth/AIME

Neural Information Processing Systems (NeurIPS) 2024 Workshop on AI-Driven Speech, Music, and Sound Generation.

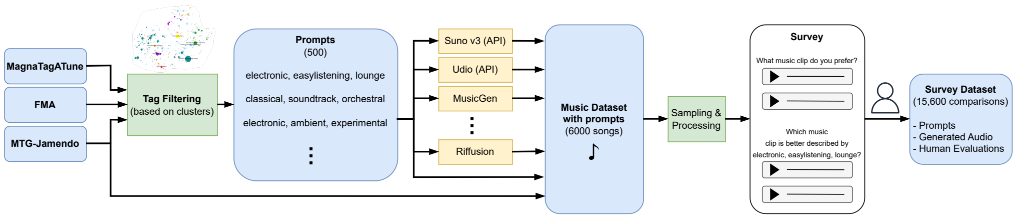


Figure 2: Overview of the tag selection, music generation, and study. We extract common and diverse tag combinations from MTG-Jamendo and use them to generate 6k music snippets with a series of models. We sample from this corpus to generate a survey dataset and let human evaluators compare the samples based on text-audio alignment and music quality.

context, our study seeks to assess the correlation between human evaluations and widely used objective metrics in music generation.

We generate 6,000 songs using 12 state-of-the-art music generative models and conduct a large-scale survey involving 15,600 pairwise audio comparisons with more than 2,500 human participants. These comparisons were designed to evaluate text-audio alignment and music preference from a human perspective and to compare these evaluations with existing objective metrics. Our results show that certain metrics align better with human judgment than others, allowing us to create a comprehensive ranking.

To facilitate further research in this area, we make both the generated music dataset and the human evaluation dataset publicly available as the **AI** **M**usic **E**valuation (**AIME**) dataset. Our contributions provide a foundation for future work aimed at improving the evaluation of AI-generated music and offer a testing ground and benchmark for new metrics that align more closely with human perception.

Our contributions can be summarized as follows:

- We conduct a large-scale survey of over 15k audio comparisons with 2.5k human participants to understand human preference for text-audio alignment and overall music preference. To this end, we create a dataset of 6k generated songs using 12 state-of-the-art music generation models.
- We analyze the relationship between human preference and existing metrics used in the domain of music generation. We find significant differences in their alignment and identify the metrics that reflect human perception the best.
- We open-source the tags, prompts, and generated songs, as well as the human responses to the survey, to enable further testing with new metrics and models.

2 Related Work

The Frechet Audio Distance (FAD) [Kilgour et al., 2018] metric serves as a quantitative measure of the perceptual quality of generated audio. FAD is computed on audio embeddings and measures the distribution similarity between a set of generated audio and ground-truth audio. FAD is commonly used on audio embeddings generated with VGGish [Hershey et al., 2017]. MusicLM [Agostinelli et al., 2023], utilized the FAD metric to demonstrate superior audio and music quality in comparison with previous models [Forsgren and Martiros, 2022]. Although FAD is a key metric for evaluating the perceptual quality of AI-generated music, the CLAP model [Elizalde et al., 2023, 2024, Wu et al., 2023] is widely used to measure the alignment between generated audio and text prompt.

Given the variety of objective metrics available, there have been efforts to assess their reliability and effectiveness, particularly in evaluating perceptual qualities such as audio and music quality. There has also been work done comparing the scores produced by objective metrics with those obtained from listening studies involving neural network-generated audio [Vinay and Lerch, 2022]. Those findings suggested that current objective metrics might not fully capture the perceptual quality of audio. More recent research has explored variations of FAD, proposing that certain embedding models provide results that correlate well on a per-song basis with subjective evaluation criteria for audio and music quality [Gui et al., 2024].

In addition to the existing objective metrics, previous approaches use a range of subjective metrics for evaluation, two prominent metrics being Mean Opinion Score (MOS) and head-to-head comparisons (HTH). MOS is a commonly used metric in which listeners rate audio samples on a scale of 1 to 5 based on specific criteria. HTH comparisons of models are used to determine the winner based on specific criteria such as text alignment or music quality [Agostinelli et al., 2023, Huang et al., 2023].

3 Dataset

A high-quality reference dataset is essential for evaluating music preference and text-audio alignment. This dataset should provide realistic music descriptions paired with corresponding audio tracks, enabling music generation with various models and serving as a reliable benchmark for both human and objective evaluation. We selected the MTG-Jamendo dataset [Bogdanov et al., 2019], which contains 55k tracks. The dataset is annotated with 195 distinct tags across genre, instrument, and mood/theme categories.

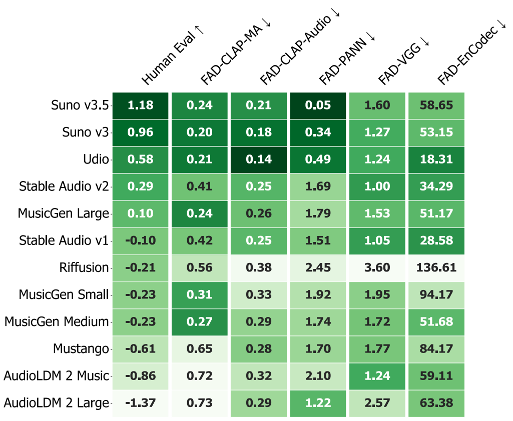
Tag Selection. To ensure effective tag-based music descriptions based on the reference dataset, we select tags that are commonly used in practice and combinations of tags that have an appropriate length for describing the generated music. Additionally, we ensure that the tags offer sufficient diversity to allow for a meaningful comparison of the models' capabilities, particularly in terms of text-audio alignment. We remove tags that are not commonly used in practice by filtering out all tags that do not appear in the FMA [Defferrard et al., 2016] or MagnaTagATune [Law et al., 2009] datasets. Additionally, we choose tag combinations of length three to ensure a consistent length that is descriptive enough and has enough unique tag combinations of that exact length in the MTG-Jamendo dataset. After these steps we are left with 1,248 unique tag combinations with at least one track in the reference dataset. To ensure a diverse set of tag-based music descriptions, we ensure that no two tag-combinations have a CLAP embedding [Elizalde et al., 2024] with a cosine similarity value above a threshold of 0.1382. The threshold is selected such that the final set of tag-based music descriptions is 500.

Music Generation. For each one of 12 music generation models we generate 500 music tracks with the selected prompts to create a dataset of 6,000 AI-generated music tracks. The models cover a diverse range of capabilities, enabling a comprehensive comparison between human and objective evaluations. For MusicGen [Copet et al., 2024], a transformer-based music generation model, we generate clips from the three checkpoints "musicgen-small", "musicgen-medium", and "musicgen-large". Additionally, we generate music with diffusion-based models Riffusion [Forsgren and Martiros, 2022], AudioLDM 2 [Liu et al., 2023] ("audioldm2-music" and "audioldm2-large" checkpoints), Mustango [Melechovsky et al., 2023] and Stable Audio [Evans et al., 2024] ("Stable Audio AudioSparx 1.0" and "Stable Audio AudioSparx 2.0"). Furthermore, we evaluate two state-of-the-art commercial music generation models, Suno [Suno] (Suno v3 and Suno v3.5) and Udio [Udio], which have recently gained attention for their ability to generate high-quality audio across a wide range of music styles and genres.

Given that many of these models were designed to generate shorter music clips without vocals or lyrics, we limit the generated track duration to 10 seconds and instrumental versions only. For models such as Suno-v3, Suno-v3.5, and Udio, as well as tracks from the MTG-Jamendo dataset that tend to exceed 10 seconds, we select a 10-second segment that contains the highest average energy. This was done to ensure that we do not randomly select sections with silence.

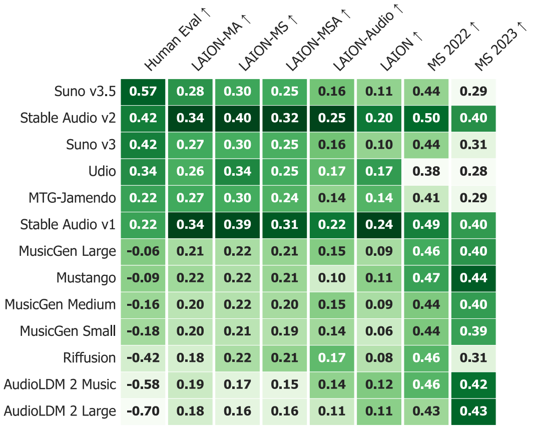
4 Human Evaluation

We focus on two key metrics: text-audio alignment and human music preference. Text-audio alignment assesses how accurately a model can generate music that abides by a given textual input. In addition, we measure human music preference to determine which music generation methods yield the subjectively best results. We take inspiration from human evaluations of LLM chatbots [Chiang et al., 2024] and use a similar methodology and evaluation technique in our study. To evaluate music preference and text-audio alignment, we design a survey using pairwise comparisons with binary preference choices between two music clips. Each survey question presents participants with two music tracks, each clipped to 10 seconds. For the evaluation of music preferences, participants were asked "What music clip do you prefer?". For text-audio alignment, participants were asked "Which



	Human Eval ↑	FAD-CLAP-MA ↓	FAD-CLAP-Audio ↓	FAD-PANN ↓	FAD-VGG ↓	FAD-EnCodec ↓
Suno v3.5	1.18	0.24	0.21	0.05	1.60	58.65
Suno v3	0.96	0.20	0.18	0.34	1.27	53.15
Udio	0.58	0.21	0.14	0.49	1.24	18.31
Stable Audio v2	0.29	0.41	0.25	1.69	1.00	34.29
MusicGen Large	0.10	0.24	0.26	1.79	1.53	51.17
Stable Audio v1	-0.10	0.42	0.25	1.51	1.05	28.58
Riffusion	-0.21	0.56	0.38	2.45	3.60	136.61
MusicGen Small	-0.23	0.31	0.33	1.92	1.95	94.17
MusicGen Medium	-0.23	0.27	0.29	1.74	1.72	51.68
Mustango	-0.61	0.65	0.28	1.70	1.77	84.17
AudioLDM 2 Music	-0.86	0.72	0.32	2.10	1.24	59.11
AudioLDM 2 Large	-1.37	0.73	0.29	1.22	2.57	63.38

Figure 3: Scores of the tested metrics for music preference estimation on the generated audio samples and Bradley-Terry parameters of our human evaluation. The cell colors indicate better (green) or worse (white) scores. FAD-PANN scores were multiplied by 1,000.



	Human Eval ↑	LAION-MA ↑	LAION-MS ↑	LAION-MSA ↑	LAION-Audio ↑	LAION ↑	MS 2022 ↑	MS 2023 ↑
Suno v3.5	0.57	0.28	0.30	0.25	0.16	0.11	0.44	0.29
Stable Audio v2	0.42	0.34	0.40	0.32	0.25	0.20	0.50	0.40
Suno v3	0.42	0.27	0.30	0.25	0.16	0.10	0.44	0.31
Udio	0.34	0.26	0.34	0.25	0.17	0.17	0.38	0.28
MTG-Jamendo	0.22	0.27	0.30	0.24	0.14	0.14	0.41	0.29
Stable Audio v1	0.22	0.34	0.39	0.31	0.22	0.24	0.49	0.40
MusicGen Large	-0.06	0.21	0.22	0.21	0.15	0.09	0.46	0.40
Mustango	-0.09	0.22	0.22	0.21	0.10	0.11	0.47	0.44
MusicGen Medium	-0.16	0.20	0.22	0.20	0.15	0.09	0.44	0.40
MusicGen Small	-0.18	0.20	0.21	0.19	0.14	0.06	0.44	0.39
Riffusion	-0.42	0.18	0.22	0.21	0.17	0.08	0.46	0.31
AudioLDM 2 Music	-0.58	0.19	0.17	0.15	0.14	0.12	0.46	0.42
AudioLDM 2 Large	-0.70	0.18	0.16	0.16	0.11	0.11	0.43	0.43

Figure 4: Scores of the tested metrics for text-audio alignment on the generated audio samples and Bradley-Terry parameters of our human evaluation. The cell colors indicate better (green) or worse (white) scores.

music clip is better described by" followed by the combination of tags used to generate the track (in the case of the reference dataset, the original tagging was used). Respondents were limited to a binary selection of either "Music Clip 1" or "Music Clip 2."

For the human evaluation, we randomly select 100 tag combinations and the corresponding generated music for each music generation model, along with the baseline tracks from MTG-Jamendo. As each track for each text description is compared with every other track with that same description, the resulting survey comprises a total of 7,800 music preference and 7,800 text alignment questions. We use the Prolific platform to run the survey [Prolific]. Although previous music evaluation studies have often relied on Amazon Mechanical Turk, [Turk] recent research suggests that Prolific produces higher quality responses [Douglas et al., 2023]. We pre-filter the survey audience to include only fluent English speakers within the age range of 18 to 34 years who reported using music streaming services. Each participant completed a survey consisting of three music preference questions and three text alignment questions. The order of the questions was randomly shuffled for each participant. Additionally, we include an attention check in the form of a text alignment question featuring a high-quality track from MTG-Jamendo paired with static white noise. We present bootstrapped Elo ratings for the music preference and text-audio alignment survey results in fig. 1. Details of our calculation of the Elo ratings can be found in appendix appendix B.

For music preference, commercial models like Suno v3.5, Suno v3, and Udio all outperformed MTG-Jamendo. Notably, Suno v3.5 achieved a significantly higher Elo than all other models. As expected, newer and larger versions of models generally performed better. For instance, Stable Audio v2 exhibited nearly a 5 percentage point increase over Stable Audio v1. Similarly, MusicGen Large outperformed MusicGen Medium and MusicGen Small. Suno v3.5 also obtained the best rating for the text-audio alignment rating with a considerable margin over the other models.

5 Metric Comparison

To compare subjective and objective metrics and thus measure how well the tested metrics align with human perception, we compute the Bradley-Terry parameters [Maystre and Grossglauser, 2015, Bradley and Terry, 1952] to indicate the "strength" of the music generation models for both the pairwise comparisons of music preference and text-audio alignment. We report the results for the objective metrics in fig. 3 and fig. 4 and the correlation of the objective metrics with the Bradley-Terry parameters of the subjective evaluation in fig. 5, where we compute the Pearson correlation coefficient and Spearman's rank correlation coefficient. We use FAD with different embedding models to evaluate how objective metrics can approximate human music preference. We employ VGGish [Hershey et al.,

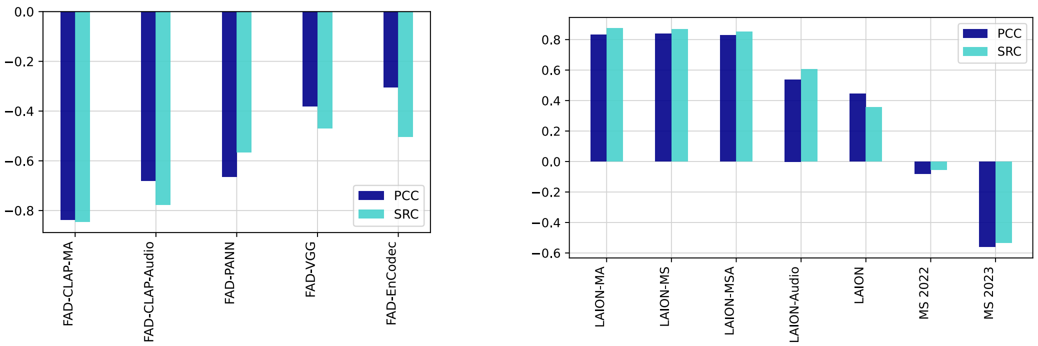


Figure 5: Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SRC) between objective metrics and Bradley-Terry parameters of the human evaluation results. **(Left)** Music preference metric correlations (lower is better). **(Right)** Text-audio alignment metric correlations (higher is better).

2017] (FAD-VGG), and PANN [Kong et al., 2020] (FAD-PANN), which are audio classification models. For CLAP, we use the audioset (FAD-CLAP-Audio) and music-audioset (FAD-CLAP-MA) checkpoints. Additionally, we use the 24 kHz mono version of EnCodec [Défossez et al., 2022] (FAD-EnCodec), an audio compression model. The MS-CLAP [Elizalde et al., 2023, 2024] and LAION-CLAP Wu et al. [2023] models are used to evaluate text-audio alignment. Specifically, for LAION-CLAP, we analyze various checkpoints (LAION, LAION-Audio, LAION-MA, LAION-MS, and LAION-MSA). For the MS-CLAP models, we consider the "2022" (MS 2022) and "2023" (MS 2023) versions. The exact checkpoint names can be found in appendix A. For all CLAP and LAION-CLAP models we compute the mean cosine-similarity between the audio embeddings and the tag-based descriptions for each music generation model.

FAD-CLAP-MA demonstrates the best correlation with human perception of music quality in terms of both linear and rank correlation with human evaluation. This finding aligns with prior FAD results [Gui et al., 2024], computed per song on a smaller set of music generation models. We can also observe that all models tend to rate Riffusion worse than our human preference study suggests and usually rank it at the last place (except for FAD-CLAP-MA). For text-audio alignment, CLAP models trained on music data (LAION-MA, LAION-MS, and LAION-MSA) exhibit the highest correlation with human ratings. Additionally, the strong correlations observed in Pearson's correlation coefficient, as well as Spearman's rank correlation coefficients, indicate that the cosine similarity values from those models demonstrate substantial linear and rank correlation with human judgments. Overall, the rankings for both music quality perception and text-audio alignment suggest that the LAION-MA checkpoint aligns best with human preferences and consistently outperforms others.

6 Conclusions

We present a comprehensive human study on the performance of current generative music models, an emerging field that lacked a comprehensive benchmark until now. We produce a corpus of music by selecting common tag combinations from MTG-Jamendo [Bogdanov et al., 2019] and utilizing a diverse range of both open-source and commercial music generation models. Through a large-scale human survey, we collect detailed feedback on human music preference and text-audio alignment, providing an unbiased ranking of the models. We find that the commercial model Suno [Suno] aligns exceptionally well with human evaluation, outperforming even the reference dataset MTG-Jamendo by considerable margins. Further, it enables us to benchmark and access the alignment of existing metrics with human perception. Among the metrics we tested, CLAP models [Wu et al., 2023] trained on music data most accurately approximate human preferences, both when employed as embedding models for computing FAD scores and for approximating text-audio alignment. We make all associated artifacts publicly available (including human evaluations) to support future research and the development of better metrics.

References

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 preprint arXiv:2301.11325*, 2023.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The mtg-jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States, 2019. URL `http://hdl.handle.net/10230/42015`.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 18 (3):e0279720, 2023.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. Natural language supervision for general-purpose audio representations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 336–340. IEEE, 2024.

Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. *arXiv preprint arXiv:2402.04825*, 2024.

Seth Forsgren and Hayk Martiros. Riffusion-stable diffusion for real-time music generation. *URL https://riffusion. com*, 2022.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting frechet audio distance for generative music evaluation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1331–1335. IEEE, 2024.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, pages 387–392. Citeseer, 2009.

Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. "audioldm 2: Learning holistic audio generation with self-supervised pretraining". *arXiv preprint arXiv:2308.05734*, 2023.

Lucas Maystre and Matthias Grossglauser. Fast and accurate inference of plackett–luce models. *Advances in neural information processing systems*, 28, 2015.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation. *arXiv preprint arXiv:2311.08355*, 2023.

Prolific. Prolific. `https://www.prolific.com/`. Accessed: 2024.

Suno. Suno. `https://suno.com/`. Accessed: May-June, 2024.

Amazon Mechanical Turk. Amazon mechanical turk. `https://www.mturk.com/`. Accessed: 2024.

Udio. Udio. `https://www.udio.com/`. Accessed: May-July, 2024.

Ashvala Vinay and Alexander Lerch. Evaluating generative audio systems and their metrics. *arXiv preprint arXiv:2209.00130*, 2022.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

A CLAP checkpoints

We use the following checkpoints for LAION-CLAP:

- FAD-CLAP-Audio: `630k-audioset-best`
- FAD-CLAP-MA: `music_audioset_epoch_15_esc_90.14`
- LAION: `630k-best`
- LAION-Audio: `630k-audioset-best`
- LAION-MA: `music_audioset_epoch_15_esc_90.14`
- LAION-MS: `music_speech_epoch_15_esc_89.25`
- LAION-MSA: `music_speech_audioset_epoch_15_esc_89.98`

B Elo Ratings Calculation

The Elo ratings were initialized with a base rating of 1,000 and a K-factor of 8. With R_A and R_B as the current Elo ratings of model A and B and R'_A and R'_B the updated ratings, the Elo ratings update formula, for each pairwise comparison of two models, is computed as follows:

$$R'_A = R_A + K \cdot (S_A - \frac{1}{1 + 10^{(R_B - R_A)/400}}) \tag{1}$$

We set $S_A = 1$ if model A wins and $S_A = 0$ if model A loses. The same formula applies to model B, with the variables of A and B switched. We randomly shuffle the pairwise comparisons 10k times and report the mean Elo rating across the bootstrapping procedure for each model.