WHEN NOISE LOWERS THE LOSS: RETHINKING LIKELIHOOD-BASED EVALUATION IN MUSIC LLMS

Xiaosha Li[1], *Chun Liu*[2], *Ziyu Wang*[34]

[1]Georgia Institute of Technology [2]ByteDance Inc.
[3]Courant Institute of Mathematical Sciences, New York University
[4] Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE
xiaosha@gatech.edu, chun.liu@bytedance.com, ziyu.wang@nyu.edu

ABSTRACT

The rise of large language models for music (Music LLMs) demands robust methods of evaluating output quality, especially in distinguishing high-quality compositions from "garbage music". Curiously, we observe that the standard cross-entropy loss—a core training metric—often decrease when models encounter systematically corrupted music, undermining its validity as a standalone quality indicator. To investigate this paradox, we introduce *noise injection experiment*, where controlled noise signal of varying lengths are injected into musical contexts. We hypothesize that a model's loss reacting positively to these perturbations, specifically a sharp increase ("Peak" area) for short injection, can serve as a proxy for its ability to discern musical integrity. Experiments with MusicGen models in the audio waveform domain confirm that Music LLMs respond more strongly to local, texture-level disruptions than to global semantic corruption. Beyond exposing this bias, our results highlight a new principle: the shape of the loss curve—rather than its absolute value—encodes critical information about the quality of the generated content (i.e., model behavior). We envision this profile-based evaluation as a label-free, model-intrinsic framework for assessing musical quality—opening the door to more principled training objectives and sharper benchmarks.[1]

Index Terms— Loss, noise, music LLMs, LLM evaluation, exposure bias

1. INTRODUCTION

In natural language processing, a common evaluation method is to feed text into a large language model (LLM) and compute its likelihood (or, equivalently, its loss) [1]. The intuition is straightforward: A sequence assigned with a higher likelihood is considered more consistent with the model's learned distribution, and therefore "better". With the rise of music LLMs, it is natural to extend this idea to music evaluation: input a musical sequence into the model and use likelihood as a proxy for quality.

However, the reliability of this approach in the music domain is far from established. While artifacts and biases of likelihood-based evaluation have been documented for text-based LLMs [2, 3], it remains unclear how these issues manifest in music or whether they align with human judgments of quality. To investigate this, we design a *noise injection experiment* in which perturbations (e.g. white noise) are added to musical sequences and the resulting changes in likelihood are measured (Figure 1). Intuitively, one would expect that corrupting music with noise decreases its likelihood, thereby increasing the loss. Surprisingly, we observe the opposite: adding noise frequently *lowers* the loss.

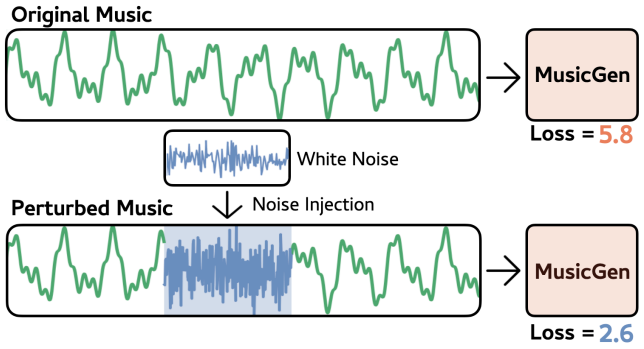


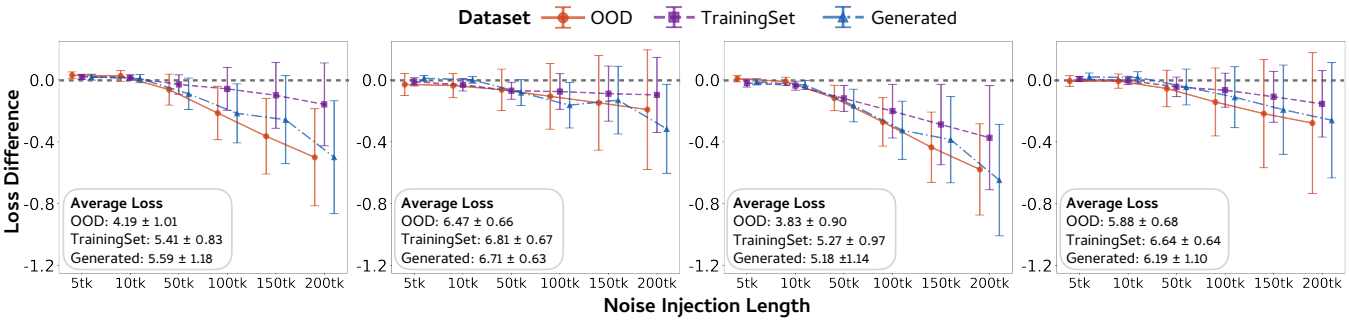
Fig. 1. Context amnesia effect revealed by noise injection experiment.

Our analysis reveals that the unexpected reduction in loss under noise arises from changes in per-token likelihood. At the onset of a noise segment, the model reacts with a sharp spike in loss, signaling recognition of inconsistency with the preceding context. Yet almost immediately afterward, the loss drops and remains low for the duration of the noise, regardless of the original context. This happens because music LLMs already assign relatively low loss to certain forms of noise, whose regularity makes it easier to predict than real music. Once the perturbation ends, the loss realigns with the original context, but with significantly higher variance. In other words, the model briefly "resists" the disturbance, then readily "forgets" the prior musical material and adapts to the noise as if it were the prevailing context. This behavior emerges consistently across noise types, musical styles, and transformer-based model variants. We refer to this phenomenon as the *Context Amnesia Effect* (Figure 1).

This finding highlights a fundamental limitation of likelihood-based evaluation in music. When measured by loss, LLMs can reliably detect only very short-term inconsistencies (e.g., onset noise) but fail to register longer-term structural degradations (e.g., phrase reorderings). In fact, the loss response to perturbations is highly inconsistent: it may rise, fall, or remain unchanged, making it an unreliable indicator of musical quality. This unpredictability is a concrete manifestation of exposure bias in music LLMs.

At the same time, our results point to a more promising direction. While absolute loss values are uninformative, the local dynamics of the loss curve carry meaningful signals. In particular,

[1]https://noiseloss.github.io



(a) MusicGen Small (b) MusicGen Medium (c) MusicGen Melody (d) MusicGen Large

Fig. 2. Comparison of model performance under white noise (Std across song differences)

we consistently observe a sharp peak at the onset of perturbation, followed by assimilation and recovery phases, which appear more reliable than raw likelihood in capturing model behavior. This suggests that profile-based evaluation focusing on the shape of token-wise loss curves rather than absolute values may provide a stronger foundation for future methods of automatic music evaluation.

2. RELATED WORK

LLM-based evaluation leverages large models to assess others through three major paradigms. Automatic methods include likelihood-based metrics, which measure probability fit but often neglect content quality [1, 4], and prompt-based multiple-choice tests that expose paradoxes in evaluative ability [5]. In contrast, non-automatic approaches such as human preference arenas rank models via large-scale pairwise voting [6]. Recent extensions like G-Eval and GPTScore, employ LLMs directly as judges—either through structured reasoning with probability-weighted scoring [7] or flexible instruction-based evaluation [8]—and report stronger correlations with human ratings in language tasks. However, applying these methods to music remains challenging: musical semantics are ambiguous; salient moments and long-range structures are difficult to capture; and high-probability continuations often fail to reflect high-quality music.

Automatic evaluation offers scalability and reduced cost, but its reliability remains uncertain. Although likelihood is an effective training objective, using it for decoding yields bland and repetitive text, diverging from the creativity and diversity valued in human language [2]. LLM-based evaluators further exhibit likelihood bias, overrating high-likelihood but superficial outputs [3], and suffer from inconsistency, with familiarity and anchoring effects and sensitivity to prompts that do not affect human judgment [9]. More broadly, such artifacts echo shortcut learning, where models exploit spurious correlations rather than demonstrating genuine ability, as observed across vision and language domains [10, 11].

In music, evaluation continues to rely primarily on human ratings, with objective metrics emerging only recently [12]. Fréchet Audio Distance (FAD) and MAD probe fidelity, musicality, and diversity [13, 14], while platforms like Music Arena scale evaluation via listener preferences [15]. CMI-Bench reframes music understanding into instruction-following tasks [16]. Audiobox-Aesthetics represents an early attempt to integrate human ratings into automated evaluation through a weighted model, though its dimensions remain limited [17]. Unlike in language domain, no established methodology exists for applying LLM-as-judge frameworks to music, motivating our investigation.

3. NOISE INJECTION EXPERIMENT

In this section, we introduce the noise injection experiment, which demonstrates the counterintuitive effect that model prediction loss decreases (or equivalently, likelihood increases) when perturbations are applied to the input audio.. The experimental setup is described in Section 3.1, and the results are presented in Section 3.2.

3.1. Experiment Setting

Given an audio signal $x_{1:T}$, where T is the number of tokens, a generative model computes the loss $\ell(x_{1:T})$, typically defined as its negative log-likelihood:

$$\ell(x_{1:T}) = -\sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}), \tag{1}$$

We define the perturbed signal $x'_{1:T}$ as

$$x'_i = \begin{cases} x_i, & i \notin \mathcal{I}, \\ \epsilon_i, & i \in \mathcal{I}, \end{cases} \tag{2}$$

where $\mathcal{I}$ is the perturbed time steps and ϵ_i denotes the given noise. Similarly, the perturbed sequence $x'_{1:T}$ has the loss

$$\ell(x'_{1:T}) = -\sum_{t=1}^{T} \log p_\theta(x'_t \mid x'_{<t}). \tag{3}$$

We also define the *token-wise loss difference* at time step t as

$$\Delta\ell_t = -\log p_\theta(x'_t|x'_{<t}) + \log p_\theta(x_t|x_{<t}), \tag{4}$$

which measures the change of sequence likelihood under perturbation from that of the original sequence at token t.

We first use white noise with controlled loudness (-30 to -12 dB) as injected noise. For an audio consisting of 750 tokens (15 seconds), the noise is injected at 250-th frame (5 seconds), which provides sufficient history for the model to establish context. Perturbation lengths are set to $5, 10, 50, 100, 150, 200$ tokens, which corresponds to $0.1, 0.2, 1.0, 2.0, 3.0, 4.0$ seconds (i.e., a token ≈ 20 ms). The chosen lengths cover multiple levels of musical perturbation, disrupting semantics at approximately the frame, note, beat, and measure levels.

3.2. Experiment Results

In our experiments, we evaluate noise injection on three types of music data: (1) **TrainingSet**: a subset of the ShutterStock [2] training corpus used for MusicGen, consisting of 20 songs; (2) **Generated**: 140 samples produced by MusicGen-Small under broad generation settings (top-k = 10, 50, 100, 150, 200, 250, 500, etc); and (3) **Out of Distribution (OOD)**: 78 classical pieces from the ASAP dataset [3], spanning a wide range of composers and styles.

We evaluate autoregressive LLMs in waveform, including MusicGen (Small(300M)/ Medium(1.5B)/ Large(3.3B)/ Melody(1.5B)) models [18].

We compute the loss difference $\Delta\ell = \ell(x'_{1:T}) - \ell(x_{1:T})$, and the results are presented in Fig. 2. Across all models, datasets, and noise lengths, we observe a consistent pattern: when the injected noise is short, the loss difference remains close to zero; as the noise length increases, the loss difference becomes negative, indicating that longer perturbations systematically decrease the loss. To verify the robustness of this trend, we apply both Pearson and Spearman correlation tests between perturbation length and average loss difference. In large-sample settings ($\approx$ 78 points), both Pearson and Spearman correlations exceed 0.85 with $p < 0.001$, indicating a highly significant negative trend. Even in small-sample settings (≈ 6 groups), correlations remain strong ($r < -0.91$) and statistically significant ($p < 0.05$), corroborated by linear regression tests showing significantly negative slopes.

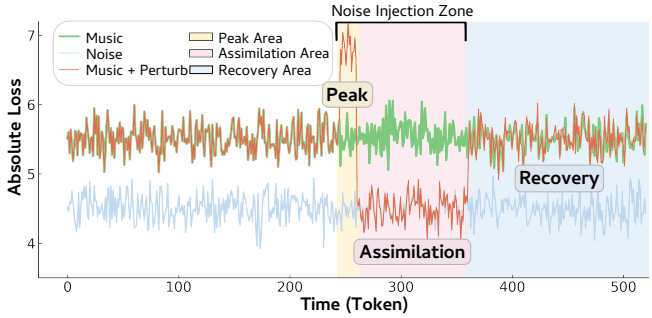


Fig. 3. Loss curve of noise injection experiment.

4. ANALYSIS OF LOSS DYNAMICS UNDER NOISE INJECTION

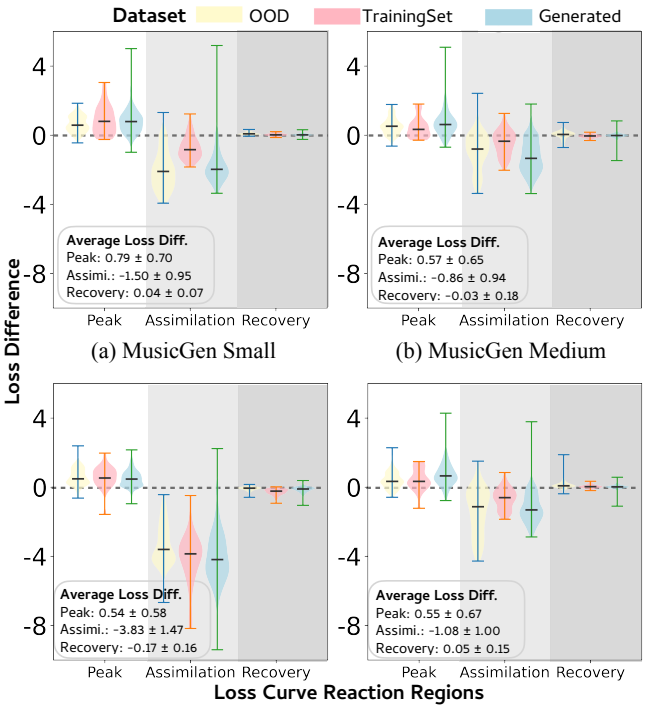
We now analyze in detail how loss behaves in the *noise injection* experiment. Our goals are twofold: first, to explain the counterintuitive trends observed in Section 1 and second, to quantify how perturbations reshape token-wise loss dynamics. We first use per-token visualization to highlight a consistent three-stage effect (Section 4.1), then validate this effect with an automatic region-detection experiment (Section 4.2).

4.1. Token-wise Loss Dynamics

To investigate the counterintuitive behavior, we analyze the *noise injection* setting in Eq. (6) at the per-token level. We compute the loss difference $\Delta\ell_t$ and visualize token-wise loss curves for original music, pure noise, and music+perturb (Fig. 3), which reveals how perturbation reshapes the loss dynamics.

We term this phenomenon the *Context Amnesia Effect*, characterized by three consistent regions across different perturbation lengths:

- **Peak area:** In the first ~100 ms (≈5 tokens), loss spikes due to local inconsistency with the preceding context.
- **Assimilation area:** Within the perturbation window, loss rapidly decreases and stabilizes at a low value, largely insensitive to context.
- **Recovery area:** After the perturbation ends, predictions become unstable and loss oscillates around the baseline, reflecting exposure bias [1].



(a) MusicGen Small (b) MusicGen Medium

(c) MusicGen Melody (d) MusicGen Large

Fig. 4. Three-stage behavior after noise injection.

4.2. Validation through Automated Region Detection

To confirm that these three regions are systematic, we conduct a region-detection experiment. We apply a moving average to $\Delta\ell_t$ to suppress local fluctuations and automatically identify boundaries where the loss crosses zero for several consecutive tokens. This procedure robustly detects the onset peak, the assimilation plateau, and the unstable recovery phase (Fig. 4).

These results confirm the qualitative analysis in Section 4.1: models reliably detect short, local perturbations but fail to register long-range structural disruptions. This *Context Amnesia Effect* underlies the unreliability of loss-based evaluation for music.

5. DISCUSSION

In the previous sections, we demonstrated that adding white noise counterintuitively lowers the loss and introduced the *Context Amne-*

[2] https://www.shutterstock.com

[3] https://github.com/fosfrancesco/asap-dataset

sia Effect to explain this behavior. While these results clarify why model loss fails to capture long-range structural disruptions, it remains unclear whether this limitation is unique to noise injection or reflects a more general weakness of loss as an evaluation metric. In this section, we extend our analysis to more realistic settings specifically, those where model loss would naturally be used to assess musical quality. We first examine order shuffling as an alternative perturbation in Section 5.1, showing that the same patterns of short-range sensitivity and long-range insensitivity persist. We then relate this phenomenon to the broader concept of exposure bias [1] in Section 5.2, highlighting its implications for evaluating generative music models.

5.1. Alternative Perturbation: Order Shuffling

To further examine our findings, we extend the analysis to another type of perturbation: order shuffling. Unlike noise injection, shuffling preserves the same amount of information while disrupting musical form across multiple levels. In this experiment, we shuffled segments of different lengths (1, 2, 5, 10, 35, 50, 70, 100, 150, and 200 tokens) to cover more broader scenarios. As shown in Fig. 5, the results mirror those of noise injection: immediate, short-span shuffles produce a sharp spike in loss, but as the shuffle length increases, the loss curve remains nearly unchanged. This suggests the model is insensitive to disruptions in musical order. Consequently, relying solely on global loss to evaluate a music model's performance is unreliable, particularly when the perturbation targets long-range structural coherence, where loss provides almost no meaningful signal.

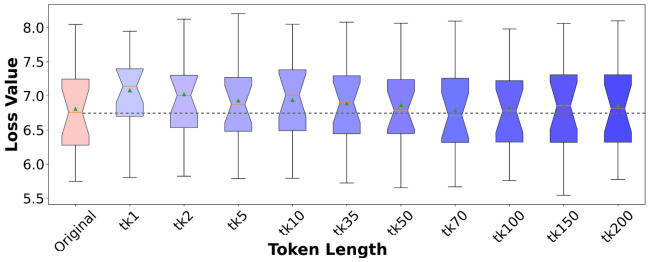


Fig. 5. Example of shuffle order perturbation with experimental results.

5.2. Relation to Exposure Bias

The Context Amnesia Effect can be related to the more well-recognized phenomenon of *exposure bias* [1], where, at inference-time generation, auto-regressive models struggle to recover once mistakes are introduced into a sequence. In our case, injected noise creates tokens entirely unfamiliar to the model from its training data. From the perspective of exposure bias, one would expect the model's predictions to degrade following the introduction of such mistakes. This expectation is consistent with what we observe: a brief spike in loss at noise onset, followed by adaptation to the disturbance and increasingly random subsequent predictions. It appears that after such mistakes, the model effectively shortens its context window, relying only on the corrupted input. This demonstrates that exposure bias is not only impairs generation but also undermines the reliability of likelihood-based evaluation. In music, where surprise, tension, and novelty are intrinsic to the art form, even an original but unfamiliar passage risks being misperceived as "errors", leading the model to undervalue it in the same manner as noise.

Our findings suggest current LLMs cannot reliably use absolute loss to distinguish between works of differing quality (e.g., between canonical classical compositions and generic training-set samples). We encourage future research to further explore the depth of this limitation in music evaluation.

6. CONCLUSION

In this work, we investigated the reliability of loss-based evaluation for music LLMs, anchored by a counterintuitive phenomenon observed in our *noise injection* experiment. We identified a key pattern termed the *Context Amnesia Effect*: when perturbations are introduced, the loss curve consistently exhibits three characteristic regions—**Peak**, **Assimilation**, and **Recovery**. By visualizing loss curves and token-wise loss differences, we demonstrated how models detect only instantaneous inconsistencies while failing to respond to longer-term structural changes. Extending our analysis to additional perturbations, such as order shuffling, further confirmed that absolute loss is unreliable for evaluating musical quality, especially at the compositional level.

Our findings underscore a fundamental limitation of likelihood-based evaluation for music: absolute loss values cannot function as a stable indicator of musical quality. Instead, the shape and local dynamics of the loss curve—particularly the onset peak—offer clearer, more consistent signals. We frame this profile-based perspective as an initial step toward developing more reliable automatic evaluation frameworks, ones that better align with music's unique structural characteristics.

7. ACKNOWLEDGMENT

The authors would like to thank Yikang Shen for his valuable guidance and Wenye Ma for her helpful assistance.

8. REFERENCES

[1] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba, "Sequence level training with recurrent neural networks," in *International Conference on Learning Representations (ICLR)*, 2016.

[2] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi, "The curious case of neural text degeneration," in *International Conference on Learning Representations (ICLR)*, 2020.

[3] Masanari Ohi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, and Naoaki Okazaki, "Likelihood-based mitigation of evaluation bias in large language models," *arXiv preprint arXiv:2402.15987*, 2024.

[4] Cong Xu, Zhangchi Zhu, Jun Wang, Jianyong Wang, and Wei Zhang, "Understanding the role of cross-entropy loss in fairly evaluating large language model-based recommendation," in *Proceedings of the ACM Conference on Recommender Systems (RecSys)*, 2024.

[5] Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Raghavi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi, "The generative ai paradox: What it can create, it may not understand," *arXiv preprint arXiv:2311.00059*, 2023.

[6] Wei-Lin Chiang, Lianmin Zheng, Siyuan Zhuang, Eric Wallace, Tianjun Li, Yingbo Sheng, Rose Wu, et al., "Chatbot arena: An open platform for large language model evaluation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[7] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," *arXiv preprint arXiv:2303.16634*, 2023.

[8] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu, "Gptscore: Evaluate as you desire," *arXiv preprint arXiv:2302.04166*, 2023.

[9] Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara, "Large language models are inconsistent and biased evaluators," *arXiv preprint arXiv:2405.01724*, 2025.

[10] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[11] Aarohi Srivastava et al., "Do llms overcome shortcut learning?," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[12] Alexander Lerch, Claire Arthur, Nick Bryan-Kinns, Corey Ford, Qianyi Sun, and Ashvala Vinay, "Survey on the evaluation of generative models in music," *arXiv preprint arXiv:2506.05104*, 2025.

[13] Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou, "Adapting fréchet audio distance for generative music evaluation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, 2024, IEEE.

[14] Yichen Huang and Chris Donahue, "Aligning text-to-music evaluation with human preferences," *arXiv preprint arXiv:2503.16669*, 2024.

[15] Yonghyun Kim, Wayne Chi, Anastasios N. Angelopoulos, Wei-Lin Chiang, Koichi Saito, Shinji Watanabe, Yuki Mitsufuji, and Chris Donahue, "Music arena: Live evaluation for text-to-music," *arXiv preprint arXiv:2507.20900*, 2025.

[16] Yuxuan Wang, Ming Sun, Hao Chen, Rui Zhang, Kat Agres, and Yi-Hsuan Yang, "Cmi-bench: A comprehensive benchmark for evaluating music instruction following," *arXiv preprint arXiv:2506.12285*, 2025.

[17] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu, "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," *arXiv preprint arXiv:2502.05139*, 2025.

[18] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, "Simple and controllable music generation," in *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.