WHEN NOISE LOWERS THE LOSS: RETHINKING LIKELIHOOD-BASED EVALUATION IN MUSIC LARGE LANGUAGE MODELS

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)
xiaosha@gatech.edu, chun.liu@bytedance.com, ziyu.wang@nyu.edu

ABSTRACT

The rise of music large language models (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, 4, 5], 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
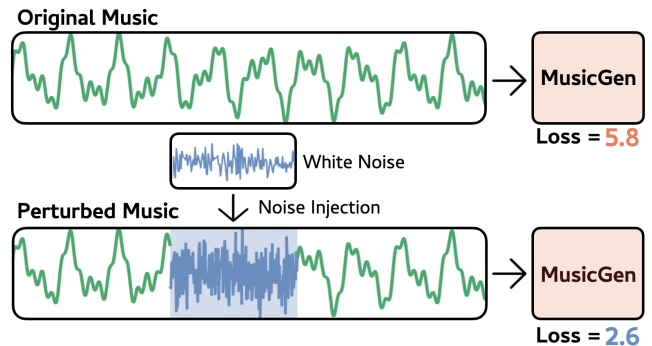

Fig. 1. The figure shows that injecting noise into the input audio can unexpectedly reduce the loss; we refer to this phenomenon as the Context Amnesia Effect.

that corrupting music with noise decreases its likelihood, thereby increasing the loss. Surprisingly, we observe the opposite: adding noise frequently *lowers* the loss.

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*.

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*[3] in music LLMs.

At the same time, our results point to a more promising direction. While absolute loss values are uninformative, the local dy-
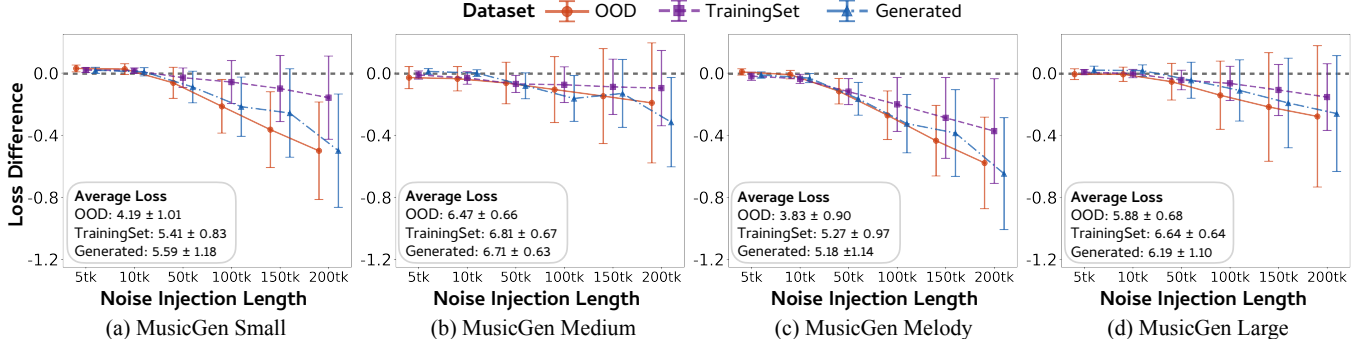

Fig. 2. Comparison of model performance under white-noise injection on the dataset, with standard deviation measured across songs.

namics of the loss curve carry meaningful signals. In particular, 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 other models' outputs through several major paradigms. Automatic methods include likelihood-based metrics, which measure probability fit but often neglect content quality [5], and prompt-based multiple-choice tests that expose paradoxes in evaluative ability [6]. In contrast, non-automatic approaches such as human preference arenas rank models via large-scale pairwise voting [7]. Recent extensions like G-Eval and GPTScore, employ LLMs directly as judges—either through structured reasoning with probability-weighted scoring [8] or flexible instruction-based evaluation [9]—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 structure are hard to capture, and high-probability continuations often fail to reflect quality.

Automatic evaluation offers scalability and reduced cost, but its reliability remains uncertain. Although likelihood is an effective training objective, it suffers from exposure bias and objective mismatch at inference time [3, 10]. Using likelihood for decoding further yields bland and repetitive text, diverging from the creativity and diversity valued in human language [4]. LLM-based evaluators also exhibit likelihood bias, overrating high-likelihood but superficial outputs [5], and suffer from inconsistency, with familiarity and anchoring effects and sensitivity to prompts that do not affect human judgment [11]. More broadly, such artifacts echo shortcut learning, where models exploit spurious correlations rather than genuine ability, as observed across vision and language domains [12, 13].

In music, evaluation continues to rely primarily on human ratings [14], with objective metrics emerging only recently [15]. Fréchet Audio Distance (FAD) and MAD probe fidelity, musicality, and diversity [16, 17], while platforms like Music Arena scale evaluation via listener preferences [18]. CMI-Bench reframes music understanding into instruction-following tasks [19]. Audiobox-Aesthetics represents an early attempt to integrate human ratings into automated evaluation through a weighted model, though its dimensions remain limited [20]. Unlike in the language domain, no

established methodology exists for using large language models as automatic judges of 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 tokenized as $x_{1:T}$, where T is the sequence length (e.g., via EnCodec [21] in MusicGen), a generative model computes the loss $\ell(x_{1:T})$ as the negative log-likelihood:

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

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

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

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

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

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

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

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, where the noise loudness is matched to the original audio to account for variations in overall amplitude across datasets, as we empirically observe that the resulting loss magnitude is sensitive to the noise loudness level. For an audio consisting of 750 tokens (15 seconds), the noise is injected at the 250-th token (5 seconds), which provides sufficient history for the model to establish context. Here, tokens correspond to RVQ-based audio tokens at a rate of 50 Hz, meaning that each token represents approximately 20 ms of audio (with

all signals resampled to 32 kHz). Perturbation lengths are set to 5, 10, 50, 100, 150, 200 tokens, corresponding to 0.1, 0.2, 1.0, 2.0, 3.0, 4.0 seconds. 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 [22].

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 the large-sample setting, where each data point represents a specific (model, dataset, noise-length) combination ($\approx$ 78 points total), both Pearson and Spearman correlation coefficients are strongly negative ($r < -0.85$, $p < 0.001$), confirming a highly significant negative trend. Even in the small-sample setting, with results aggregated by noise length ($\approx$ 6 groups), correlations remain robust ($r < -0.91$) and statistically significant ($p < 0.05$), a finding further corroborated by the significantly negative slopes observed in linear regression tests.

To further examine the generalizability of this effect, we repeat the analysis on the YuE (1B) [23] autoregressive model using three datasets: TrainingSet, OOD, and Generated data produced by YuE itself. For both TrainingSet and OOD data, Pearson correlation tests reveal a consistent and significant negative relationship between noise length and loss difference ($r < -0.90$, $p < 0.01$), corroborated by significantly negative regression slopes. For the Generated data, although no statistically significant correlation between noise length and loss difference is observed ($p > 0.05$), all noise-injected samples consistently exhibit lower loss values compared to their corresponding clean generated counterparts. This generalized finding indicates that the absolute loss value alone is insufficient to determine whether a musical sequence has been perturbed by noise, particularly in model-generated music. Visualizations of these experiments are available on our demo page.

4. ANALYSIS OF LOSS DYNAMICS

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 3 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).
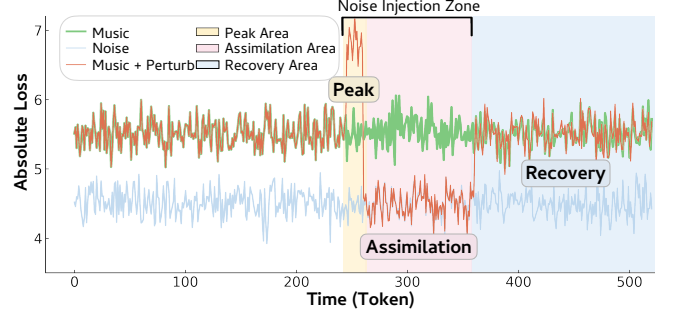
[2]https://www.shutterstock.com
[3]https://github.com/fosfrancesco/asap-dataset

Fig. 3. Loss curve of noise injection experiment.

4.1. Token-Wise Loss Dynamics

To investigate the counterintuitive behavior, we analyze the noise injection setting in Eq. (4) 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 across all models and datasets in our experiments (Fig. 3), revealing how perturbation reshapes the loss dynamics.
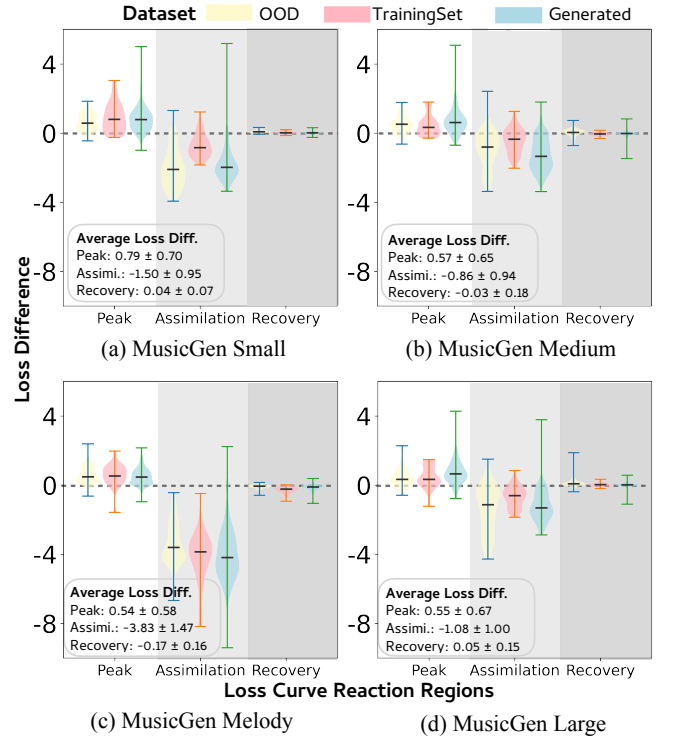

Fig. 4. Three-stage behavior after noise injection.

Across different perturbation lengths, we observe three characteristic regions in the token-wise loss trajectory:

- **Peak area:** In the first $\sim$100 ms ($\approx$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.

This three-stage behavior suggests that the model transiently loses effective access to prior musical context, as if earlier information were forgotten, which we term the Context Amnesia Effect.

4.2. Validation Through Automated Region Detection

To confirm that these three regions are consistent, 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 five consecutive tokens. This procedure robustly detects the onset peak, the assimilation plateau, and the unstable recovery phase (Fig. 4).

These results validate the hypothesis suggested by the qualitative analysis in Section 4.1: models respond strongly to short, local perturbations but fail to reflect 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 Amnesia 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 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 perturbation: order shuffling. Unlike noise injection, shuffling preserves the same amount of information while disrupting musical order in a structured manner. We shuffle segments of different lengths (1, 2, 5, 10, 35, 50, 70, 100, 150, and 200 tokens) to cover a broad range of perturbation scales, from local rearrangements to long-range disruptions of musical structure.

As shown in Fig. 5, the results closely mirror those of noise injection. Short shuffles produce a sharp spike in loss, as the perturbation is dominated by the peak region. As the shuffle length increases, the assimilation region becomes longer, during which the model adapts to the perturbed order and effectively loses access to the original context, causing the loss to remain close to the baseline. Although long-span shuffling introduces substantial and structured degradation of musical form, the global loss remains largely unchanged, indicating the unreliability of global loss for evaluating model performance under long-range structural disruptions.

5.2. Relation to Exposure Bias

The Context Amnesia Effect is closely related to the well-known phenomenon of exposure bias, where auto-regressive models strug-
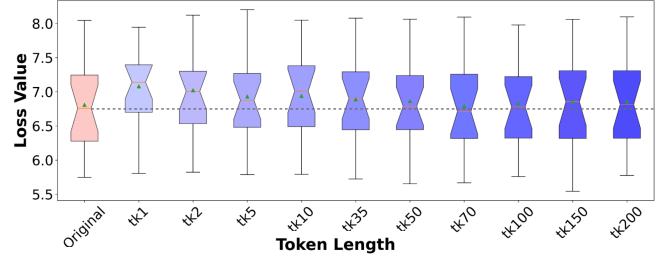

Fig. 5. Example of shuffle order perturbation with experimental results.

gle to recover after errors during inference-time generation. In our setting, injected noise produces tokens that are out-of-distribution relative to the training data, leading to a brief loss spike at noise onset followed by rapid adaptation and increasingly unstable predictions. After such errors, the model effectively shortens its usable context and relies on corrupted input, indicating that exposure bias not only degrades generation but also undermines the reliability of likelihood-based evaluation. In music, where surprise, tension, and novelty are essential, unfamiliar yet meaningful passages may therefore be misinterpreted as perturbations and undervalued 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

8. REFERENCES

[1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin, "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, 2003.

[2] Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Chi Kit Cheung, "Why Exposure Bias Matters: An Imitation Learning Perspective of Error Accumulation in Language Generation," *Findings of the Association for Computational Linguistics (ACL)*, 2022.

[3] 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.

[4] 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.

[5] Masanari Ohi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, and Naoaki Okazaki, "Likelihood-based Mitigation of Evaluation Bias in Large Language Models," in *Findings of the Association for Computational Linguistics (ACL)*, 2024, pp. 3237–3245.

[6] 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," in *International Conference on Learning Representations (ICLR)*, 2024.

[7] Wei-Lin Chiang, Lianmin Zheng, Siyuan Zhuang, Eric Wallace, Tianjun Li, Yingbo Sheng, Rose Wu, et al., "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[8] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu, "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 2511–2522.

[9] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu, "GPTScore: Evaluate as You Desire," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2024, pp. 6556–6576.

[10] 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.

[11] Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara, "Large Language Models are Inconsistent and Biased Evaluators," *arXiv preprint arXiv:2405.01724*, 2024.

[12] 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.

[13] Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu, "Do LLMs Overcome Shortcut Learning? An Evaluation of Shortcut Challenges in Large Language Models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024, pp. 12188–12200.

[14] Florian Grötschla, Ahmet Solak, Luca A. Lanzendörfer, and Roger Wattenhofer, "Benchmarking Music Generation Models and Metrics via Human Preference Studies," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025, IEEE.

[15] Alexander Lerch, Claire Arthur, Nick Bryan-Kinns, Corey Ford, Qianyi Sun, and Ashvala Vinay, "Survey on the Evaluation of Generative Models in Music," *ACM Computing Surveys*, vol. 58, no. 4, pp. 99, 2025.

[16] Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou, "Adapting Fréchet Audio Distance for Generative Music Evaluation," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, pp. 1331–1335, IEEE.

[17] Yichen Huang and Chris Donahue, "Aligning Text-to-Music Evaluation with Human Preferences," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2025.

[18] 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," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[19] Yuxuan Wang, Ming Sun, Hao Chen, Rui Zhang, Kat Agres, and Yi-Hsuan Yang, "CMI-Bench: A Comprehensive Benchmark for Evaluating Music Instruction Following," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2025.

[20] 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.

[21] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research (TMLR)*, 2023, Accepted September 2023.

[22] 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.

[23] Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, et al., "Yue: Scaling Open Foundation Models for Long-Form Music Generation," *arXiv preprint arXiv:2503.08638*, 2025.