

CLOUDEAR: INTEGRATING PERCEPTUAL, DIAGNOSTIC, AND CROSS-MODAL EVIDENCE FOR MUSIC EVALUATION

Qiqi He

NetEase Cloud Music

heqiqi02@corp.netease.com

Anqi Huang

NetEase Cloud Music

huanganqi01@corp.netease.com

ABSTRACT

Evaluating generated music requires modeling perceptual quality, technical audio quality, and prompt alignment. Existing methods often overlook signal-level defects, produce compressed quality scores, and miss fine-grained musical attributes in text prompts. We propose *CloudEar*, a pairwise framework for evaluating musicality and prompt alignment. It comprises three components: *i*) a Perceptual Expert for song-quality assessment; *ii*) a Diagnostic Expert for learned and signal-level quality analysis; and *iii*) a Cross-modal Expert for prompt–song alignment based on musical attributes. Experiments show that *CloudEar* outperforms existing song evaluation methods across both pairwise preference accuracy and score-correlation metrics.

1. INTRODUCTION

Evaluating generated music requires assessing both audio quality and musical aesthetics. Signal-level defects such as clipping and abnormal loudness can degrade listening quality, while musical structure, expression, and performance require higher-level perceptual judgment [1]. Existing methods emphasize high-level musical semantics but often overlook technical quality and subtle perceptual differences. Regression models also exhibit central-tendency bias [2], resulting in compressed score differences.

In parallel, prompt alignment poses a separate challenge. General-purpose audio–text models capture broad semantic correspondence but often overlook fine-grained musical attributes such as instrumentation requirements. Consequently, high audio–text similarity does not always indicate whether a generated song follows the specific attributes requested by the prompt.

To address these limitations, we propose *CloudEar*, a pairwise framework for evaluating musicality and prompt alignment. Its Perceptual Expert combines ordinal regression, continuous regression, and pairwise ranking to model human-perceived song quality to mitigate prediction concentration. A Diagnostic Expert complements this

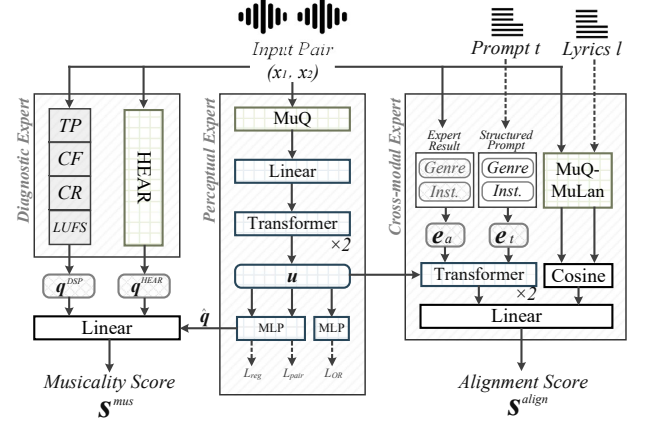


Figure 1. Overview of the proposed *CloudEar* framework.

representation with learned quality estimates and signal-level descriptors. For alignment, a Cross-modal Expert compares genre and instrumentation attributes extracted independently from the audio and prompt. Experiments on rating and pairwise-preference benchmarks show that *CloudEar* consistently improves both musicality and alignment evaluation over existing methods.

2. PROPOSED FRAMEWORK

As illustrated in Fig. 1, given a prompt t and two generated songs (x_1, x_2) , *CloudEar* predicts their pairwise preferences in musicality and prompt alignment.

2.1 Perceptual Expert

Following SongBench [3], we randomly sample a 30-second clip from each song x_i and extract MuQ features [4], which are processed by a Transformer adapter to obtain the perceptual representation u_i .

To alleviate the tendency of regression models to concentrate predictions around the mean [2], we jointly optimize ordinal regression, continuous regression, and pairwise ranking objectives. The ordinal loss L_{OR} [5] supervises ordered quality levels, while an MLP regression head predicts the quality score $\hat{q}_i$ using $L_{reg} = (\hat{q}_i - y_i)^2$, where y_i is the ground-truth score. For a training pair with $y_1 > y_2$, the pairwise loss encourages a higher score for the preferred song:

$$L_{pair} = -\log \sigma(\hat{q}_1 - \hat{q}_2), \quad (1)$$



where σ denotes the sigmoid function. The overall training objective is

$$L_{PE} = 0.4L_{OR} + 0.4L_{reg} + 0.2L_{pair}. \quad (2)$$

2.2 Diagnostic Expert

The Diagnostic Expert complements high-level perceptual modeling with signal-level features and learned quality estimates. We extract true-peak level (TP), clipping ratio (CR), crest factor (CF), and integrated loudness (LUFS) to form the DSP feature vector $\mathbf{q}_i^{\text{DSP}}$, with loudness and true-peak measurements following standard practices [6, 7]. We also use HEAR [8], which combines MuQ and MusicFM representations, to obtain the quality score q_i^{HEAR} . The DSP feature vector, HEAR score, and perceptual score $\hat{q}_i$ are concatenated and fed into a linear layer to produce the musicality score s_i^{mus} .

The musicality fusion layer is trained using binary cross-entropy (BCE) between the pairwise preference label and the predicted probability $\sigma(s_1^{\text{mus}} - s_2^{\text{mus}})$. At inference time, *CloudEar* selects the song with the higher musicality score.

2.3 Cross-modal Expert

The Cross-modal Expert assesses genre and instrument alignment between the prompt and song. On the audio side, Genre Discogs400 [9] predicts 400 music styles, and MTG-Jamendo Instrument [10] predicts instrument tags. Their outputs are concatenated into $e_{a,i}$. On the prompt side, we use Qwen3-30B [11] to extract structured genre and instrument information from t . The extracted attributes are mapped to label vectors using the same vocabularies as the audio classifiers. These vectors are concatenated to form the prompt representation e_t .

A two-layer Transformer jointly models u_i , $e_{a,i}$, and e_t to obtain h_i . We also compute the cosine distance between the input audio x_i and lyrics l using MuQ-MuLan [4]:

$$d_{i,l} = 1 - \cos(f_a(x_i), f_t(l)), \quad (3)$$

where f_a and f_t denote its audio and text encoders. A linear layer maps $[h_i; d_{i,l}]$ to the alignment score s_i^{align} . The alignment branch is trained with BCE on $\sigma(s_1^{\text{align}} - s_2^{\text{align}})$. At inference time, the song with the higher score is selected.

3. EXPERIMENTS

3.1 Datasets & Metrics

We train the Perceptual Expert on AIME [12], MusicArena [13], and approximately 7,000 hours of in-house human-rated music, mainly English songs from five generation models. For the in-house dataset, we retain only songs rated by more than 200 listeners.

For pairwise training, we additionally use MusicEval [14] and the training split of CMI-Pref [15]. Their musicality annotations train the Perceptual Expert and musicality fusion layer, while

Table 1. Comparison and ablation results. Mus./Ali.: musicality/alignment. MuQ uses MuQ-Eval [17] for Mus. and MuQ-MuLan [4] audio-text similarity for Ali.

Method	CMI-Test		MA.	PAM-Music	
	Mus. P-ACC	Ali. P-ACC	P-ACC	SRCC	LRCC
CMI-RM [15]	.7833	.7389	.7194	.6626	.6379
HEAR [8]	.7882	.6970	.5769	.6180	.6230
SongBench [3]	.7340	.6823	.6746	.5437	.5123
MuQ [4, 17]	.5936	.6429	.6478	.3168	.3608
<i>CloudEar</i>	.8179	.7635	.7657	.7257	.7229
<i>PE + HEAR</i>	.7805	–	.7325	.6872	.7008
<i>PE + DSP</i>	.7620	–	.6845	.6351	.6626
<i>PE only</i>	.7592	.6743	.6736	.6056	.6515

their alignment annotations train the Cross-modal Expert. Musicality pairs use within-dataset rating gaps above 0.2 after $[0, 1]$ normalization while alignment pairs share a prompt.

Following prior work [15], evaluation uses PAM-Music [16], CMI-Test, and the MusicArena test set (MA.): 1,340 song pairs generated in Jul.–Dec. 2025, with no training-data overlap. We report SRCC and LRCC for scalar ratings and P-ACC for A/B preferences.

3.2 Experimental Results

Ablation Study. The full *CloudEar* outperforms all reduced configurations on musicality, confirming their complementary contributions. Both HEAR and DSP improve over *PE only*, demonstrating complementary information beyond the perceptual representation. The alignment gain over *PE only*, trained directly on alignment labels, supports the benefit of the Cross-modal Expert. The Perceptual Expert alone already outperforms SongBench on CMI-Test musicality and remains competitive on MusicArena-Test.

Comparison with Prior Work. For baseline evaluation, SongBench and HEAR use musicality scores for both tasks. *CloudEar* outperforms all compared baselines on the reported benchmarks. Its higher musicality and alignment P-ACC indicate better agreement with human judgments of quality and prompt alignment, respectively. Higher SRCC and LRCC on PAM-Music further demonstrate stronger rank and linear correlations with human quality ratings. Notably, SongBench and HEAR achieve above-chance alignment accuracy without prompts, suggesting a correlation between musicality and alignment preferences.

4. CONCLUSION

We presented *CloudEar*, a pairwise framework for evaluating the musicality and prompt alignment of generated songs. *CloudEar* integrates perceptual, diagnostic, and cross-modal experts to model song quality and fine-grained prompt alignment. Experiments on three human-annotated benchmarks demonstrate consistent improvements over existing methods. Future work will explain which musical factors lead to each preference.

5. AI USAGE STATEMENT

Qwen3-30B is used to extract structured genre and instrument attributes from input prompts, as described in Section 2.3. The authors take full responsibility for the content of this paper.

6. REFERENCES

- [1] J. A. Sloboda, *The Musical Mind: The Cognitive Psychology of Music*. Oxford: Clarendon Press, 1985.
- [2] H. Lee and S. Chen, "Systematic bias of machine learning regression models and correction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 6, pp. 4974–4983, 2025.
- [3] D. Wu, S. Lei, W. Tan, G. Li, Y. Wang, H. Zhang, L. Zuo, and Z. Wu, "Songbench: A fine-grained multi-aspect benchmark for song quality assessment," 2026. [Online]. Available: <https://arxiv.org/abs/2604.25937>
- [4] H. Zhu, Y. Zhou, H. Chen, J. Yu, Z. Ma, R. Gu, Y. Luo, W. Tan, and X. Chen, "Muq: Self-supervised music representation learning with mel residual vector quantization," *arXiv preprint arXiv:2501.01108*, 2025. [Online]. Available: <https://arxiv.org/abs/2501.01108>
- [5] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.
- [6] International Telecommunication Union, "Algorithms to measure audio programme loudness and true-peak audio level," International Telecommunication Union, Tech. Rep. Recommendation ITU-R BS.1770-5, 2023. [Online]. Available: <https://www.itu.int/rec/R-REC-BS.1770>
- [7] European Broadcasting Union, "Loudness normalisation and permitted maximum level of audio signals," European Broadcasting Union, Tech. Rep. EBU Recommendation R 128, 2020. [Online]. Available: <https://tech.ebu.ch/publications/r128>
- [8] S. Liu, Y. Jin, R. Lin, S. Chen *et al.*, "HEAR: Hierarchically enhanced aesthetic representations for multi-dimensional music evaluation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026.
- [9] P. Alonso-Jiménez, X. Serra, and D. Bogdanov, "Music representation learning based on editorial metadata from discogs," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2022.
- [10] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-Jamendo dataset for automatic music tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning*, Long Beach, CA, USA, 2019. [Online]. Available: <http://hdl.handle.net/10230/42015>
- [11] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.
- [12] F. Grötschla, A. Solak, L. A. Lanzendörfer, and R. Wattenhofer, "Benchmarking music generation models and metrics via human preference studies," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.
- [13] Y. Kim, W. Chi, A. N. Angelopoulos, W.-L. Chiang, K. Saito, S. Watanabe, Y. Mitsufuji, and C. Donahue, "Music arena: Live evaluation for text-to-music," 2025. [Online]. Available: <https://arxiv.org/abs/2507.20900>
- [14] C. Liu, H. Wang, J. Zhao, S. Zhao, H. Bu, X. Xu, J. Zhou, H. Sun, and Y. Qin, "Musiceval: A generative music corpus with expert ratings for automatic text-to-music evaluation," 2025. [Online]. Available: <https://arxiv.org/abs/2501.10811>
- [15] Y. Ma, H. Xia, H. Gao, W. Chen, Y. Ye, Y. Yang, S. Chang, M. Ding, Y. Li, R. Yuan, S. Dixon, and E. Benetos, "CMI-RewardBench: Evaluating music reward models with compositional multimodal instruction," 2026. [Online]. Available: <https://arxiv.org/abs/2603.00610>
- [16] S. Deshmukh, B. Elizalde, R. Wang, B. Raj, and R. Singh, "PAM: Prompting audio-language models for audio quality assessment," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2024.
- [17] D. Zhu and Z. Li, "MuQ-Eval: An open-source per-sample quality metric for AI music generation evaluation," 2026. [Online]. Available: <https://arxiv.org/abs/2603.22677>