ENHANCING TEXT-TO-MUSIC GENERATION THROUGH RETRIEVAL-AUGMENTED PROMPT REWRITE 2025

Anonymous Authors
Anonymous Affiliations
anonymous@ismir.net

ABSTRACT

This paper evaluates the extent to which expertise in prompt construction influences the quality of the music generation output. We propose a **Retrieval-Augmented Prompt Rewrite** system (RAG) [1] that transforms novice prompts into expert descriptions using CLAP. Our method helps preserve user intent and bypass the need for extensive domain training of the user. Given novice-level prompts, participants selected relevant terminologies from top-k most textually or audibly similar MusicCaps captions, which were fed into GPT 3.5 to create succinct, expert-level rewrites. These rewrites were then used to generate music with Stable Audio 2.0. To mitigate anchoring bias toward expert prompts, we implemented a counterbalanced design and conducted a subjective study to evaluate the effectiveness of RAG. We generated rewrites using a traditional LoRA fine-tuning method as our baseline. Participants evaluated the *expertness*, *musicality*, *production quality* and *preference* of music generated from novice and expert prompts. Both RAG and LoRA rewrites significantly improve music generation across all NLP and subjective metrics, with RAG outperforming LoRA overall. Finally, the subjective results largely align with Meta's Audiobox Aesthetics metrics.

1. INTRODUCTION

Text-to-music platforms such as Stable Audio [1], Suno [2], and Riffusion [3] enable users to express creative intent through text prompts. However, models trained on prompts with domain-specific semantics [2] often encounter underspecified real-world queries [3] and out-of-distribution prompts leads to subpar outputs at inference time.

To address this description gap, we propose a Retrieval-Augmented Prompt Rewrite system (RAG) that helps novices craft precise, expressive prompts without requiring musical training. Our approach uses CLAP-based retrieval [4] to preserve and enrich user intent. Pre-computed CLAP embeddings from MusicCaps [2] enable retrieval of audio and captions most similar to the user's novice query. Users then select keywords to guide GPT-3.5 [5] in generating an expert-level rewrite. For example, a novice prompt like "*Calming classical music similar to Bach with harp*" becomes "*Heavenly, melancholic ballads with harp arpeggios, similar to calming classical Bach*" (Figure 1).

[1] GitHub link redacted for review
[2] https://suno.com
[3] https://github.com/riffusion/riffusion-hobby

© F. Author, S. Author, and T. Author. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** F. Author, S. Author, and T. Author, "Enhancing Text-to-Music Generation through Retrieval-Augmented Prompt Rewrite 2025", submitted to *ISMIR*, 2025.

2. RELATED WORKS

Challenges in Text-to-Music Prompt Construction. Underspecified prompts often yield generic outputs. Zang and Zhang [6] identify this "one-to-many mapping" problem between one vague prompt and many valid interpretations and propose the use of LLMs for aligning model outputs with user intent. Other efforts include rank-based alignment [7] and intent taxonomies for retrieval scenarios [3]. These approaches prioritize cross-modal similarity scores or retrieval over expressive generation.

Retrieval-Augmented Generation.

RAG [8] combines a retrieval module with a sequence-to-sequence generator for *knowledge-intensive* tasks. The original framework uses a pre-trained retriever—comprising a query encoder and a dense document index—and a pretrained generator, which are fine-tuned jointly to adapt to the specific tasks.

RECAP [9] applies RAG to audio captioning by incorporating retrieved captions as contextual input. We extend this to the reverse process of text-to-music generation, treating novice prompts as out-of-distribution inputs and enriching them with retrieved textual descriptions from a CLAP-based index. We rely on pre-trained retriever and generator components—CLAP and GPT 3.5, respectively.

Re-AudioLDM [10] addresses diffusion-based models' poor performance on rare events in audio generation. The authors propose a retrieval-augmented framework that uses CLAP to retrieve top-k relevant text–audio pairs and incorporates those features (via AudioMAE and T5 encoders) into the latent diffusion model through cross-attention. Our system prioritizes user interaction and improves music generation quality without requiring model fine-tuning.

Contrastive Language-Audio Pretraining. CLAP [4] aligns audio and text in a shared embedding space via contrastive learning. We use a CLAP checkpoint (music_audioset_epoch_15_esc_90.14.pt) trained on music + Audioset + LAION-Audio-630k, because it is one of the larger models best suited for music related tasks.

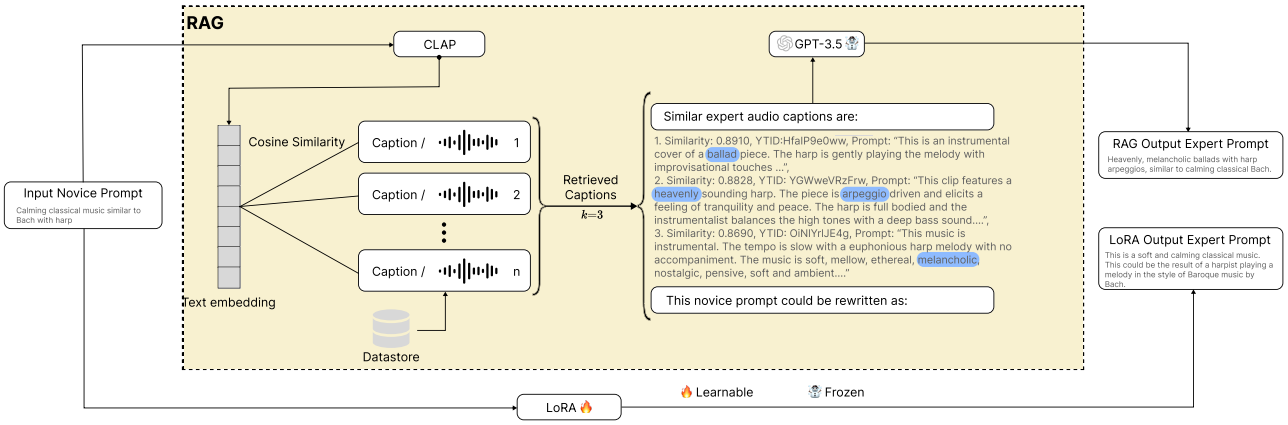

Figure 1. Overview of two novice-to-expert prompt rewrite methods: (1) **RAG**, a retrieval-augmented generation method that uses CLAP-based similarity to retrieve the top-$k = 3$ most relevant audio captions; participants then select keywords (highlighted in blue) to guide GPT-3.5 in generating a custom expert-level prompt; and (2) **LoRA**, a fine-tuned model.

3. METHOD

3.1 Data

We used the MusicCaps dataset introduced by Agostinelli et al. [2] as the RAG datastore and as the basis for LoRA fine-tuning. Out of the 5,521 music samples—including 1,000 genre-balanced examples spanning 24 genres—we were able to download 5,353 for computing text and audio embeddings. Each entry has an aspect list detailing musical features (e.g., "pop, mellow piano, high-pitched vocal") and a musician-written caption describing the 10 second YouTube music track. To tackle the scarcity of novice-style music descriptions, we leveraged GPT-3.5 for MusicCaps prompt simplification by providing 24 novice-style examples, one per genre. With in-context learning, we generated a new set of novice captions, resulting in 5,521 novice-expert caption pairs for LoRA fine-tuning.

3.2 Baseline: LoRA Model

Low-rank Adaptation (LoRA) [11] is a parameter-efficient fine-tuning method for adapting large language models to downstream tasks with reduced computational cost. We selected LLaMA-3 for LoRA as it supports parameter-efficient fine-tuning and is open-sourced. Our preliminary results showed that LoRA outperformed in-context baselines (GPT-3.5 and LLaMA-3) on accuracy metrics such as BLEU and METEOR, and LoRA achieved an 89.49% win rate in LLM-as-a-judge evaluations compared to full fine-tuning. Consequently, we used LoRA as our fine-tuned baseline despite the backbone difference from RAG.

We fine-tuned LLaMA-3.1-8B-Instruct [12] on the novice–expert paired dataset using a batch size of 4, 1500 training steps, and LoRA applied to the top 8 transformer layers. The model learned to transform novice prompts into expert-level descriptions, preserving key musical attributes (instruments, genre, mood).

3.3 RAG Procedure and Counterbalanced Design

Participants enriched novice prompts by selecting keywords from retrieved captions on a StreamLit user interface and each novice prompt is passed through GPT-3.5 alongside the keywords to generate the RAG expert rewrite; experts prompts were used to generate audio via Stable Audio 2.0 (See Appendix for full RAG procedure).

To reflect the real-world text-to-music iterative workflows, we allowed participants to listen to the novice music generation and then rewrite prompts based on the initial output. However, rating their own rewrites can introduce anchoring bias. To mitigate this, we use a counterbalanced design: 24 participants were split into two groups, group 1 rewrites the first three prompts and rates group 2's rewrites of the remaining three, and vice versa.

3.4 Evaluation Methods

3.4.1 NLP Metrics

We used four NLP metrics to evaluate the performance of our models in three different areas: BLEU [13] for semantic fidelity, Type-Token Ratio [14] (TTR) and Measure of Textual Lexical Diversity [15] (MTLD) for lexical diversity, and Flesch Reading Ease [16] (FRE) for complexity.

3.4.2 Meta Audiobox Aesthetics Metrics

To complement our survey, we applied Meta Audiobox Metrics [17]—a model-based evaluation method that rates audio along four axes: *Content Usefulness* (CU: potential for reuse in content creation), *Production Complexity* (PC: complexity of audio scene, measured by number of audio components), *Production Quality* (PQ: technical aspects such as clarity and fidelity), and *Content Enjoyment* (CE: emotional impact, artistic expression, and subjective impact). These approximately map onto our survey dimensions of *expertness*, *musicality*, *production quality*, and *preference*, respectively.

4. RESULTS

4.1 NLP Results

In our study, the RAG rewrites consistently achieved higher BLEU scores compared to the LoRA rewrites, suggesting they may better preserve the original intent in the novice prompt (See Table 1). RAG rewrites show a clear advantage over LoRA in both MTLD and TTR scores, indicating that RAG produces more lexically diverse outputs. Finally, RAG also significantly surpasses LoRA in FRE, which indicates that the RAG rewrites are significantly more complex than LoRA rewrites.

Model	B1	B2	B3	B4	TTR	MTLD	FRE
LoRA	0.19	0.11	0.06	0.03	0.42	34.29	76.93
RAG	**0.28**	**0.17**	**0.12**	**0.08**	**0.58**	**86.20**	**32.33**

Table 1. Results of NLP metrics to evaluate rewrite's adherence to novice prompt (BLEU 1–4), lexical diversity (TTR, MTLD), and complexity (FRE). Bold values indicate the best performance.

4.2 Survey Results

4.2.1 *Paired t-tests.*

We first conducted three separate paired-sample t-tests to compare subjective ratings across four evaluation dimensions corresponding to questions 2 to 5 in the survey—*expertness*, *musicality*, *production quality*, and *preference*—for music generated from novice prompts, RAG rewrites, and LoRA rewrites. Both RAG and LoRA outperform novice prompts across all four dimensions ($p < 0.01$), which is significant even after Bonferroni correction for multiple comparison at an alpha-level of 0.05, but no statistically significant differences were found between RAG and LoRA. To explore this further, we turn to regression models to investigate performance difference between LoRA and RAG.

4.2.2 *OLS Regression: Score ∼ Version.*

We first fit a simple OLS model with the rewrite method as a categorical predictor, where the levels correspond to the three prompt versions. As shown in Table 2, the model reveals a consistent ranking: **RAG > LoRA > Novice** across all four metrics. Both RAG and LoRA show significant improvements over novice prompts ($p < 0.001$), with RAG yielding higher intercepts than LoRA in all four metrics. For instance, *production quality* increases by 0.76 with LoRA and 0.99 with RAG, on a 3-point scale. Similarly, Figure 2 shows RAG exhibits higher proportions of top ratings (score = 3, shown in green) compared to LoRA and Novice, with mean scores consistently increasing across versions. For example, in the *musicality* dimension, the average score rises from 1.56 (Novice) to 2.19 (LoRA) and further to 2.25 (RAG). While pairwise t-tests only test for statistical significance, OLS results shows the effect size of RAG is consistently larger than that of LoRA.

Score	Intercept	LoRA	RAG	Adjusted R^2
Expertness	1.64	0.50†	**0.58**†	0.09
Musicality	1.56	0.64†	**0.69**†	0.14
Production	1.42	0.76†	**0.99**†	0.26
Preference	1.58	0.54†	**0.71**†	0.13
CU	2.25	**0.29**†	0.27†	–
PC	2.02	-0.09	0.05	–
PQ	2.36	0.18†	**0.20**†	–
CE	2.23	0.19*	**0.21**†	–

Table 2. Rewrite Version Effects for Survey (OLS) and Audiobox (Mixed Effect Model where Version is considered as fixed effect and PromptID as random intercept). $\dagger p < 0.001$, $*p < 0.1$

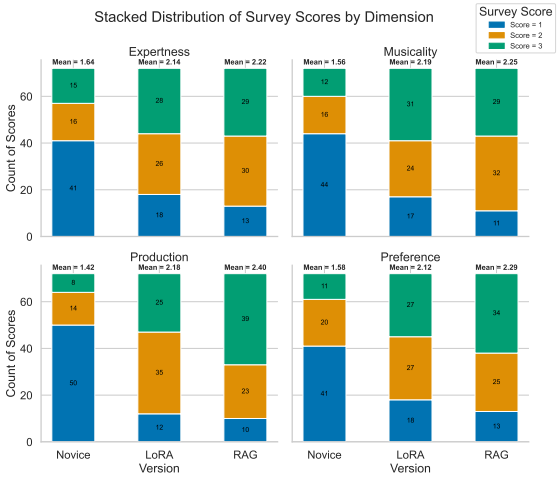

Figure 2. Stacked bar plots showing the distribution of survey scores (1–3) across rewrite versions for each evaluation dimension. Mean scores of each version are annotated above each bar.

4.2.3 *Prompt-Specific Effects: Score ∼ Version × PromptID.*

To assess whether rewrite effectiveness is influenced by prompt content, we added interaction terms with `PromptID`. Table 3 results show that LoRA exhibits positive interactions (p-value < 0.05) with Prompts 2, 4, 5, and 6 across various metrics, suggesting LoRA's improvement is prompt-dependent (See Appendix Section 8.6). In contrast, RAG has minimal interactions with `PromptID`, indicating that its effectiveness is robust across prompts.

4.2.4 *Participant Variance: Score ∼ Version × PromptID + (1|ParticipantID).*

A mixed-effects model treating `ParticipantID` as a random intercept resulted in a group variance ≈ 0. This suggests that participant-specific variation is minimal and does not significantly influence the scores, confirming the robustness of observed version effects across listeners.

Overall, both RAG and LoRA significantly improve generation quality. RAG achieves the strongest performance and robustness across prompts and users.

Score	Intercept	Prompt	LoRA Interaction	RAG Interaction	Adjusted R^2
Expertness	2.00	P5 (−0.67), P6 (−0.83)	P2 (+0.92), P4 (+1.33), P5 (+1.42), P6 (+1.67)	None	0.15
Musicality	1.83	P6 (−0.67)	P4 (+1.08), P5 (+0.92), P6 (+1.25)	None	0.18
Production	1.50	None	P6 (+0.92)	P2 (−0.75)	0.28
Preference	1.75	None	P5 (+1.08), P6 (+1.17)	None	0.16
CU	2.38	P2 (+0.22), P5 (−0.98), P6 (−0.18)	P5 (+0.85)	P2 (−0.18), P5 (+0.95)	0.81
PC	2.28	P2 (−0.92)	None	None	0.46
PQ	2.45	P5 (−0.82)	P5 (+0.77)	P5 (+0.75)	0.73
CE	2.41	P5 (−1.05)	P5 (+0.92)	P5 (+0.94)	0.61

Table 3. OLS Prompt and Version Interaction Effects ($p < .05$) on Survey (top) and Audiobox (bottom) Scores

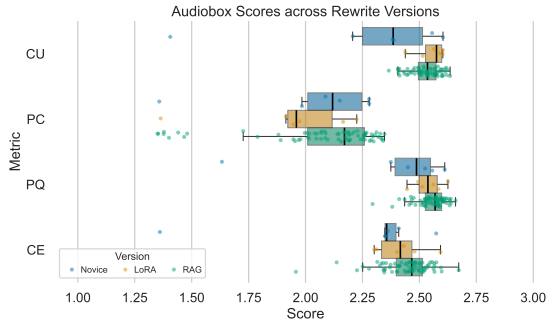

Figure 3. Audiobox evaluation scores (CU, PC, PQ, CE) for music generated using Novice, LoRA, and RAG prompts. Each point represents the score of an individual audio clip, while boxplots summarize the score distribution for each rewrite method in that dimension.

4.3 Audiobox Metric Results

4.3.1 Mixed-Effects Model:
Score $\sim$ *Version* $+ (1|$*PromptID*$)$.

The model was able to isolate the effect of Version from prompt-level variability by including PromptID as a random intercept. It converged with non-zero prompt variance, confirming that when prompt dependence was accounted for, we still finding a significant fixed effect for Version. As shown in Table 2 and Figure 3, both RAG and LoRA outperform Novice generations for CU, PQ, and CE. The definitions of PQ and CE align best with the survey dimensions, and the Audiobox results on these two dimensions are also the most consistent with the the survey findings, with average scores of RAG > LoRA > Novice. For instance, PC, as computed by Audiobox, captures the number of audio components, whereas our survey's definition of *musicality* includes broader artistic attributes such as instrument usage, genre appropriateness, and emotional impact. Therefore, the lack of significant improvement in PC is expected, given that our rewriting methods do not inherently favor increased instrumentation.

4.3.2 Prompt-Specific Effects:
Score $\sim$ *Version* $\times$ *PromptID*.

To further examine version effects at the prompt level, we ran an OLS model with interactions. As shown in Table 3, PC remains largely unaffected, while other metrics show significant interactions with certain prompts—especially Prompt 5, which had the lowest base score but saw significant boosts from both LoRA and RAG.

In conclusion, Audiobox results in the PQ and CE dimensions corroborate our survey findings, further validating that RAG rewrites produces the most consistent and high-quality musical output.

5. DISCUSSION

Our proposed rewrite methods successfully address the "one-to-many mapping" challenge posed by underspecified prompts by adding more expert-level musical attributes that reduces the scope of potential generations, as evidenced by the reduced score variance in the RAG and LoRA groups compared to the Novice group (See Figure A1 in Appendix Section 8.6).

We evaluated six prompts covering *R&B*, *classical*, *pop*, *soul*, *indie*, and *jazz*, ensuring as much stylistic coverage and semantic diversity as possible. To ensure the validity of our comparisons despite the limited number of prompts, we explicitly modeled prompt-level variability: in Section 4.3.1, the mixed effect model treated `PromptID` as a random intercept. This approach separates the variance explained by prompt-specific effects, ensuring that the fixed effect of `Version` reflects differences between rewrite methods after accounting for prompt variability. Similarly, analyses in Sections 4.2.3 and 4.3.2 contained 12 observations per `PromptID` in all three levels, enabling us to separate the method effectiveness from prompt variability.

Analysis on the musical ability of survey participants (Question 1), use of more novice prompts, as well as augmentation of the datastore with more expert caption-music pairs may further improve the proposed rewrite methods.

6. CONCLUSION

Our findings show that while LoRA rewrites improve music generation, RAG consistently outperforms LoRA, demonstrating superior robustness and greater preservation of user intent. By allowing the selection of relevant terminologies, RAG more effectively bridges the gap between novice and expert creators without demanding extensive domain knowledge. The resulting prompts generates music that consistently score higher across subjective and objective evaluations. These insights demonstrate the potential of retrieval-augmented methods to enhance creative workflows in real-world applications, particularly in industry contexts where high-quality music generation with minimal barriers to entry for users is of high priority.

7. REFERENCES

[1] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Long-form music generation with latent diffusion," 2024. [Online]. Available: https://arxiv.org/abs/2404.10301

[2] A. A. et al., "Musiclm: Generating music from text," 2023. [Online]. Available: https://arxiv.org/abs/2301.11325

[3] S. Doh, K. Choi, D. Kwon, T. Kim, and J. Nam, "Music discovery dialogue generation using human intent analysis and large language models," 2024. [Online]. Available: https://arxiv.org/abs/2411.07439

[4] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "Clap: Learning audio concepts from natural language supervision," 2022. [Online]. Available: https://arxiv.org/abs/2206.04769

[5] T. B. B. et al., "Language models are few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[6] Y. Zang and Y. Zhang, "The interpretation gap in text-to-music generation models," 2024. [Online]. Available: https://arxiv.org/abs/2407.10328

[7] E. Chang, S. Srinivasan, M. Luthra, P.-J. Lin, V. Nagaraja, F. Iandola, Z. Liu, Z. Ni, C. Zhao, Y. Shi, and V. Chandra, "On the open prompt challenge in conditional audio generation," 2023. [Online]. Available: https://arxiv.org/abs/2311.00897

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021. [Online]. Available: https://arxiv.org/abs/2005.11401

[9] S. Ghosh, S. Kumar, C. K. R. Evuru, R. Duraiswami, and D. Manocha, "Recap: Retrieval-augmented audio captioning," 2024. [Online]. Available: https://arxiv.org/abs/2309.09836

[10] Y. Yuan, H. Liu, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Retrieval-augmented text-to-audio generation," 2024. [Online]. Available: https://arxiv.org/abs/2309.08051

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[12] A. Grattafiori and et al., "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, 2002, p. 311–318.

[14] E. Castello, *Text Complexity and Reading Comprehension Tests*. Peter Lang, 2008.

[15] P. M. McCarthy, "An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (mtld)." 2005.

[16] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom, "Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for navy enlisted personnel," in *Research Branch Report*, C. of Naval Technical Training, Ed. Memphis, USA: Naval Air Station Memphis, 1975, pp. 8–75.

[17] A. T. et al., "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," 2025. [Online]. Available: https://arxiv.org/abs/2502.05139

8. APPENDIX

8.1 RAG Procedure

1. Novice Prompt: Participants are shown a novice-level text prompt and listened to its corresponding generated audio. They identify areas for improvement (e.g., better instrumentation, unclear style).
2. Prompt Refinement: Using the StreamLit interface (see GitHub), participants modify the original prompt into an "expert-level" description. This involves selecting keywords from retrieved textual or audio examples to add details about instrumentation, mood, genre, or other musical attributes they deemed important for generating a more expert-level musical output.
3. Music Generation: The refined prompt is then processed by Stable Audio 2.0, producing a 30-second music output. Repeat Steps 1–3 for three prompts.
4. Evaluation: Each participant ranks three versions of music (Novice, LoRA, RAG) generated by each of the three prompts rewritten by the other participant. Survey questions are listed below.

8.2 Survey Questions

1. Q1 (Musical Ability): "How familiar are you with the current genre under evaluation?"
2. Q2 (Expertness): "Which version of the generated music sounds most like it was composed by an expert musician?"
3. Q3 (Musicality): "Which version is the most musical, considering instrument usage, genre alignment, and emotional conveyance?"
4. Q4 (Production Quality): "Which version sounds the most professional in terms of clarity, balance, mixing, and overall naturalness?"
5. Q5 (Preference): "Which version do you prefer overall?"
6. Q6 (Text-to-Music Consistency): "Did you notice any inconsistencies in how well the generated music adhered to the text prompt? If so, which version had the most issues?"

For questions 2 to 5, we converted the user rankings for the three music versions (Novice, LoRA, RAG) into a numeric scale, assigning a score of 1 to the version originally ranked last, 2 to the version ranked second, and 3 to the version ranked highest.

8.3 Audiobox Data Imbalance & Modeling

Unlike the survey, where multiple participants rated the same pieces of audio for the Novice and LoRA groups, the Audiobox metrics are computed directly from the audio itself, yielding only one set of scores (4 dimensions) per clip, and a total of 6 sets of scores (6 PromptIDs) for each of the two groups. In contrast, RAG was still evaluated on 72 pieces of audio, since each of the 12 pairs of participants generated one distinct RAG rewrite for each PromptID. This data imbalance precludes paired t-tests, so we used linear models for analysis.

8.4 Diffusion Randomness

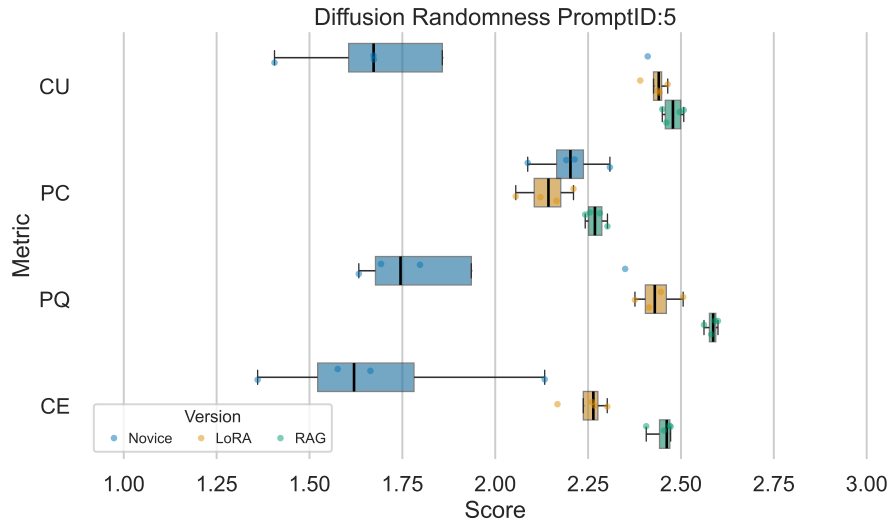

Figure A1. Audiobox scores for music generated from the same Novice, LoRA, or RAG prompt (PromptID = 5).

We did not explicitly model the variability inherent in diffusion process in the experiment (e.g., generating multiple musical outputs per prompt), but rather assumed minimal variation across outputs from the same text input. However, if we generate multiple audio for the same prompt in each group (here we take PromptID = 5 as an example), we can find the resulting rewrite groups' Audiobox scores has higher mean and lower variance than that of the Novice group in the CU, PQ, and CE dimensions, as shown in Figure A1, which aligns with our Audiobox Analysis results in Section 4.3. This indicates that the effectiveness of rewrite methods is robust to random fluctuations in diffusion-based generation. Higher average Audiobox scores show that rewrites better leverage the capabilities of the text-to-music model, and the lower variance in rewrite groups suggests more consistent outputs and improved handling of underspecified prompts.

Further comparison between the LoRA group with the RAG group reveals that RAG method better capture user intent. While LoRA-based rewrites reduced ambiguity by mimicking expert-style prompts from MusicCaps, rigid fine-tuning limit user control. In contrast, RAG embraces the one-to-many nature of the task: it retrieves multiple relevant candidate prompts and enables refinement through personalized keyword selection. This flexibility is also reflected in NLP metrics, where RAG achieves higher lexical diversity, greater textual complexity, and consistently higher BLEU scores than LoRA—indicating more specific, expert-level rewrites that better capture user intent.

8.5 Text-to-Music Consistency

To assess text-to-music consistency, as discussed in Q6, we computed the CLAP score for each audio and prompt pair. The 72 RAG prompt-audio pairs achieved the highest mean CLAP score (0.4987, sd=0.03), followed by 6 LoRA prompt-audio pairs (0.4621) and 6 Novice prompt-audio pairs (0.4266). However, this result contrasts with our survey Q6 responses, where LoRA received the highest inconsistency vote. This discrepancy could be caused by Stable Audio model's difficulty in generating human vocals when prompted, which many participants identified as the source of inconsistency.

8.6 Prompt-specific Variation

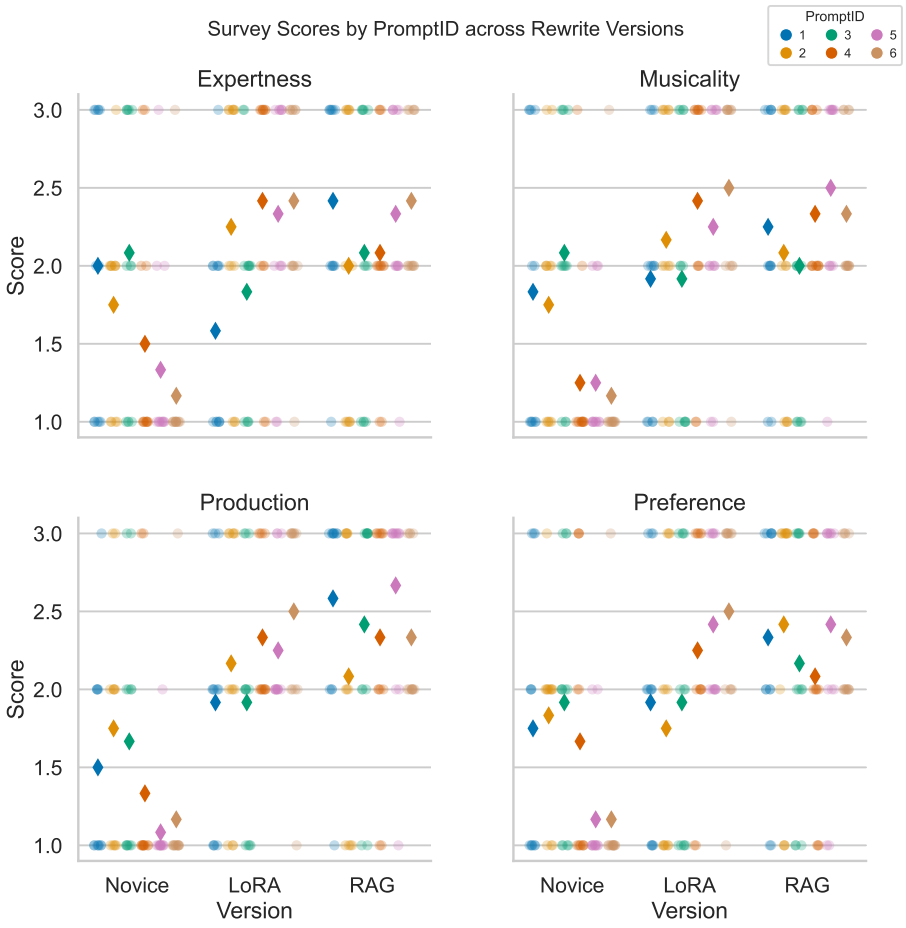

Figure A2. Survey scores (Questions 2-5) for four evaluation metrics for music generated using Novice Baseline, LoRA, RAG prompts across PromptIDs. Each circle represents a participant rating. Diamonds indicate the mean score for each rewrite method within each PromptID.

This figure illustrates participant ratings for four evaluation metrics across three prompt versions when blocked by PromptID. Each circle represents an individual participant rating for a specific prompt (color-coded by PromptID, jittering used to avoid overlap between participant ratings and reveal the underlying density), while diamonds indicate the mean score for each version within each prompt.

Overall, Novice prompts consistently receive the lowest scores across all metrics, while both rewrite methods show substantial improvement. Among the two, RAG generally achieves the higher mean ratings with less prompt-level variation. The tighter cluster of diamonds often near the top of the scale represents greater improvement and higher consistency. In contrast, LoRA improvements appear more prompt-dependent and is clustered more sparsely, as certain prompts (e.g., Prompt 4, 5 and 6, shown in red, brown and pink) show larger gains while others (e.g., Prompt 1, 2 and 3, shown in blue, yellow and green) exhibit smaller differences. This complements the results of Table 3, where LoRA's effect interacts more with PromptID. These patterns suggest that RAG method's improvement to music generation is more generalizable when individuals could tailor the rewrites.