Enhancing Text-to-Music Generation through Retrieval-Augmented Prompt Rewrite

Meiying Ding[1] **Sunny Yang**[2] **Chenkai Hu**[2] **Juhua Huang**[2]

Brian McFee[1,2]
[1] Steinhardt, New York University, United States
[2] Center for Data Science, New York University, United States
{miyading, sy2577, ckh326, jh9029, brian.mcfee}@nyu.edu

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 expert-level rewrites. These rewrites were then used to generate music with Stable Audio 2.0. We conducted a subjective study to evaluate the effectiveness of RAG against a LoRA fine-tuning 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. The subjective results largely align with Meta's Audiobox Aesthetics metrics.

1 Introduction

Text-to-music generation platforms, such as Suno and Riffusion, provide users with creative tools to generate music from text prompts. However, models trained on prompts with domain-specific semantics [6] often encounter underspecified real-world queries [4], leading to subpar outputs at inference time. Zang and Zhang [12] identify this "one-to-many mapping" problem and propose using LLMs to align outputs with user intent. Other methods include rank-based alignment [3] and intent-driven retrieval taxonomies [4], which emphasize cross-modal similarity expressive generation.

Retrieval-Augmented Generation (RAG) [9] combines retrieval with seq2seq models for *knowledge-intensive* tasks. RECAP [7] applies this to audio captioning by retrieving textual descriptions as context. We reverse this for text-to-music: novice prompts are enriched using CLAP [5] and GPT-3.5 [2]. Unlike Re-AudioLDM [11], which fuses retrieved features into a latent diffusion model, our method emphasizes interactive prompt rewriting without model fine-tuning. MusicCaps [1] serves as the RAG datastore and for LoRA fine-tuning.

[1] https://github.com/miyading/ismir_text_rewrite

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

2 Method

Baseline: LoRA Model. We fine-tuned LLaMA-3.1-8B-Instruct [8] on a novice–expert paired dataset. Preliminary results showed that LoRA outperformed in-context baselines on accuracy metrics and achieved a 90% win rate in LLM-as-a-judge evaluations against full fine-tuning.

Proposed RAG Procedure. Participants performed the rewrite in a StreamLit interface. First, each novice prompt was embedded using the CLAP model [5]. We retrieved top-k captions and music clips by comparing the novice prompt's CLAP text embedding to both the precomputed text and audio embeddings in the MusicCaps datastore. Retrieved examples included descriptive captions and aspect lists, from which users selected relevant musical terms. These terms were passed to GPT-3.5 [2] along with the original novice prompt to generate an expert-level rewrite. The final rewritten prompt was then given to Stable Audio 2.0 to generate a 30-second music clip (See Figure A1 in Appendix).

3 Results

3.1 Subjective and Objective Results Table

Evaluation Models	**Key Findings and Effect Sizes**
Survey Results	
Model 1: Paired t-tests	RAG and LoRA outperform Novice prompts across all metrics ($p < 0.01$).
Model 2: OLS	RAG $>$ LoRA $>$ Novice across all four metrics ($p < 0.001$). LoRA effect sizes: Expertness $+0.50$, Musicality $+0.64$, Production $+0.76$, Preference $+0.54$ RAG effect sizes: Expertness $+0.58$, Musicality $+0.69$, Production $+0.99$, Preference $+0.71$
Model 3: OLS, PromptID interaction	LoRA shows prompt-specific gains: P2 ($+0.92$), P4 ($+1.33$), P5 ($+1.42$), P6 ($+1.67$). RAG shows minimal interaction effects; only P2 (-0.75) significant.
Model 4: Mixed-Effects, ParticipantID random intercept	Participant-level variance ≈ 0, indicating consistent effects across listeners.
Audiobox Results	
Model 5: Mixed-Effects, PromptID random intercept	RAG and LoRA outperform Novice for CU, PQ, and CE. PC shows no significant gains. LoRA effect sizes: CU $+0.29^{\dagger}$, PQ $+0.18^{\dagger}$, CE $+0.19^{*}$, PC -0.09 RAG effect sizes: CU $+0.27^{\dagger}$, PQ $+0.20^{\dagger}$, CE $+0.21^{\dagger}$, PC $+0.05$
Model 6: OLS, PromptID interaction	Prompt 5 shows the strongest version-based improvements across CU, PQ, and CE.

Table 1: Summary of Survey and Audiobox Models: model and key findings with effect sizes for Novice, RAG, LoRA prompt versions, and effect sizes for PromptIDs. Significance levels: $\dagger p < 0.001$, $*p < 0.1$

Survey Results. Across all analyses, both RAG and LoRA significantly outperform novice prompts in all four metrics (See Table 1). While paired t-tests (with Bonferroni correction) exhibit no significant difference between RAG and LoRA, OLS shows that RAG achieves larger effect sizes—for example, a $+0.99$ boost in *production quality* vs. $+0.76$ for LoRA. Model 3 reveals prompt-specific variation for LoRA, whereas RAG remains robust across different prompt context. Model 4 finds negligible participant-level variance, indicating the consistency of the version effects across listeners.

Audiobox Results. We used the Meta Audiobox Aesthetics model [10] to evaluate generation quality across four perceptual dimensions: Content Usefulness (CU), Production Complexity (PC), Production Quality (PQ), and Content Enjoyment (CE). Model 5 shows that both RAG and LoRA improve CU, PQ, and CE over novice prompts, with no gain in PC—consistent with the fact that our rewrite method do not inherently favor more audio components. OLS with prompt interaction further highlights the capacity of both methods to remediate low-quality novice prompts. PQ and CE closely align with *production quality* and *preference*; the result of CU align with the *expertness* dimension, suggesting expert-like tracks may also be more reusable for downstream production.

4 Conclusion

Our findings show that while both RAG and LoRA improve music generation from novice prompts, RAG consistently outperforms LoRA across subjective and objective evaluations. In addition to higher mean scores, RAG prompts exhibit lower score variance, indicating more focused outputs. This shows that enriching underspecified prompts with expert-level attributes narrows the generative space and mitigates the one-to-many mapping. RAG also shows robust performance across a stylistically diverse prompt set, and effect sizes are significant after accounting for prompt-level variability. These results highlight the potential of RAG methods to enhance creative workflows, particularly in industry settings where high-quality generation with minimal barriers to entry for users is of high priority.

References

[1] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank. MusicLM: Generating music from text, 2023. arXiv:2301.11325.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, and D. Amodei. Language models are few-shot learners, 2020. arXiv:2005.14165.

[3] 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. arXiv:2311.00897.

[4] S. Doh, K. Choi, D. Kwon, T. Kim, and J. Nam. Music discovery dialogue generation using human intent analysis and large language models, 2024. arXiv:2411.07439.

[5] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang. CLAP: Learning audio concepts from natural language supervision, 2022. arXiv:2206.04769.

[6] Z. Evans, J. D. Parker, C. J. Carr, Z. Zukowski, J. Taylor, and J. Pons. Long-form music generation with latent diffusion, 2024. arXiv:2404.10301.

[7] S. Ghosh, S. Kumar, C. K. R. Evuru, R. Duraiswami, and D. Manocha. RECAP: Retrieval-augmented audio captioning, 2024. arXiv:2309.09836.

[8] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, and J. J. et al. The Llama 3 herd of models, 2024. arXiv:2407.21783.

[9] 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. arXiv:2005.11401.

[10] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, C. Wood, A. Lee, and W.-N. Hsu. Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound, 2025. arXiv:2502.05139.

[11] Y. Yuan, H. Liu, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang. Retrieval-augmented text-to-audio generation, 2024. arXiv:2309.08051.

[12] Y. Zang and Y. Zhang. The interpretation gap in text-to-music generation models, 2024. arXiv:2407.10328.

5 Appendix

5.1 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.

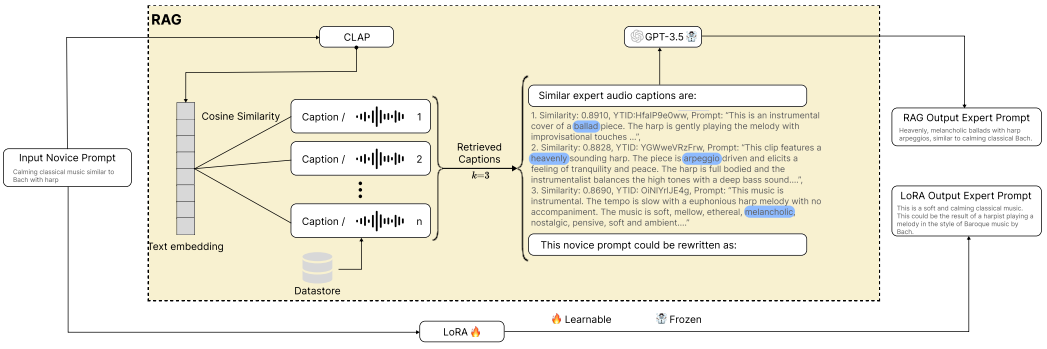

Figure A1: Overview of two novice-to-expert prompt rewrite methods: (1) **RAG**, a retrieval-augmented generation method that uses CLAP-based similarity to retrieve the top-$k = 3$ most relevant audio captions; participants then select keywords (highlighted in blue) to guide GPT-3.5 in generating a custom expert-level prompt; and (2) **LoRA**, a fine-tuned model.

5.2 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.

5.3 Counterbalanced Experiment Design

Allowing users to listen to the music corresponding to the novice prompt is essential since it mimics the real-life iterative workflow for users of a text-to-music generation platform, where users would frequently generate an initial piece, reflect on its shortcomings, and then refine the text prompt to improve alignment with their creative goals. However, a key challenge in this setup is the risk of anchoring bias, which could arise if participants listen to their own novice-generated music before creating or evaluating the expert versions.

To manage this risk, we adopt a counterbalanced design featuring two main groups of participants: the first group (participants 1-12) rewrites novice prompts into expert-level prompts for the first three of six total prompts and then evaluates the music generated from the last three prompts (both novice and expert versions), and vice versa. By separating the rewriting phase from the evaluation phase, we can more accurately measure the efficacy of the refined prompts and reduce bias caused by direct involvement in prompt refinement. In this study, 24 participants are recruited and randomly assigned to either Group 1 or Group 2, resulting in 72 data points. Participants were paired during each time slot so that every participant had corresponding music outputs to evaluate.

5.4 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.

5.5 Diffusion Randomness

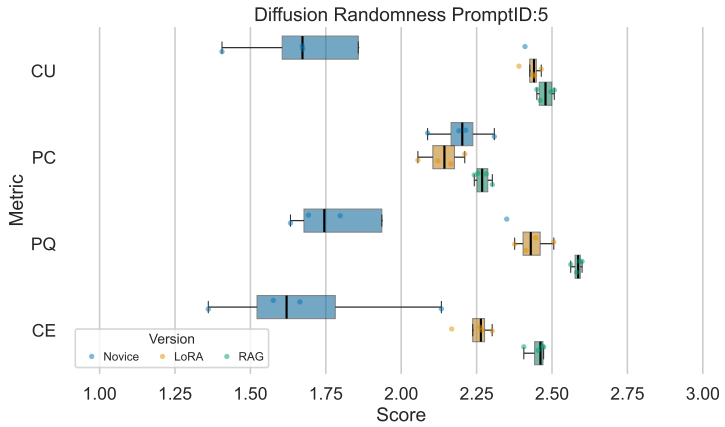

Figure A2: Audiobox scores for music generated from the same Novice, LoRA, or RAG prompt (PromptID = 5).

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 A2, 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.

5.6 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.

5.7 Prompt-specific Variation

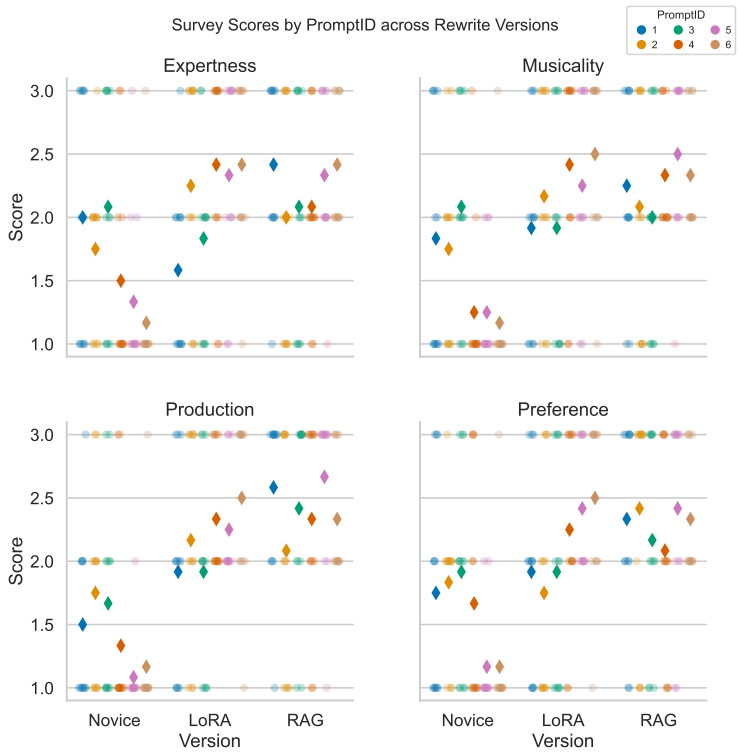

Figure A3: Survey scores (Questions 2-5) for four evaluation metrics for music generated using Novice Baseline, LoRA, RAG prompts across PromptIDs. Each circle represents a participant rating. Diamonds indicate the mean score for each rewrite method within each PromptID.

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 Model 3: OLS with PromptID interaction in Table 1, 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.