# 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 platforms such as Stable Audio [Evans et al., 2024], Suno[2], and Riffusion[3] enable users to express creative intent through text prompts. However, models trained on prompts with domain-specific semantics [Agostinelli et al., 2023] often struggle with underspecified real-world queries [Doh et al., 2024]; such *out-of-distribution* prompts often cannot exploit the full capability of music generation models and can lead to subpar outputs during inference.

To address this description gap, we propose a *Retrieval-Augmented Prompt Rewrite* system (RAG) that helps novices craft precise, expressive prompts without requiring extensive musical training. Our approach uses CLAP-based retrieval [Elizalde et al., 2022] to preserve and enrich user intent. Pre-computed CLAP embeddings from MusicCaps [Agostinelli et al., 2023] enable retrieval of audio and captions most similar to the user's novice query. Users then select keywords to guide GPT-3.5 [Brown et al., 2020] in generating an expert-level rewrite. For example, a novice prompt such as "*Calming classical music similar to Bach with harp*" becomes "*Heavenly, melancholic ballads with harp arpeggios, similar to calming classical Bach*" (See Figure 1 for selected keywords).

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

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

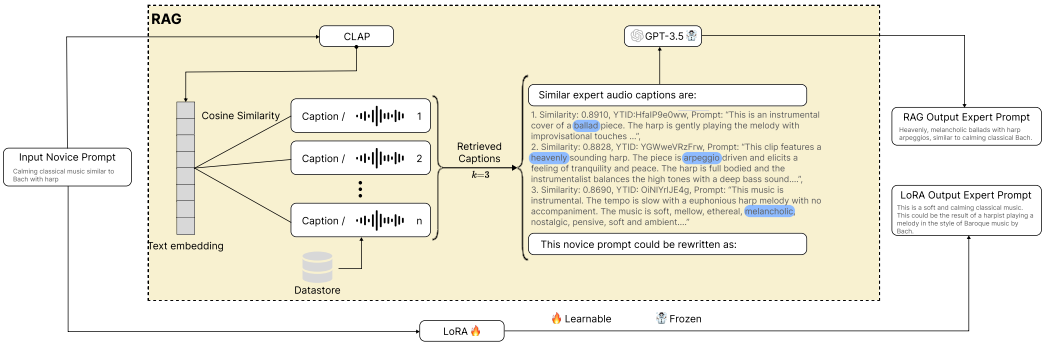


Figure 1: Overview of two novice-to-expert prompt rewrite methods: (1) **RAG**, a retrieval-augmented generation method that uses CLAP-based similarity to retrieve the top-$k = 3$ most relevant audio captions; participants then select keywords (highlighted in blue) to guide GPT-3.5 in generating a custom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

## 2 Related Work

**Challenges in Text-to-Music Prompt Construction.** Underspecified prompts often yield generic outputs. Zang and Zhang [2024] identify this "one-to-many mapping" problem between a vague prompt and its many valid interpretations, proposing the use of LLMs for aligning model outputs with user intent. Other efforts include rank-based alignment [Chang et al., 2023] and intent taxonomies for retrieval scenarios [Doh et al., 2024]. These approaches emphasize cross-modal similarity scores or retrieval, often at the expense of expressive generation.

**Retrieval-Augmented Generation.** Retrieval-Augmented Generation (RAG) [Lewis et al., 2021] combines a retriever 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 pre-trained generator, which are fine-tuned jointly for task adaptation.

Ghosh et al. [2024] extend RAG to audio captioning by incorporating retrieved captions as contextual input. We extend this idea in the reverse direction: text-to-music generation. We treat novice prompts as out-of-distribution inputs and enrich them with retrieved textual descriptions from a CLAP-based index, relying on pre-trained components (CLAP retriever + GPT-3.5 generator).

Yuan et al. [2024] address diffusion-based models' poor performance on rare events in audio generation by introducing retrieval of top-$k$ text–audio pairs (via CLAP) and incorporating features from AudioMAE and T5 encoders into a latent diffusion model through cross-attention. In contrast, our system prioritizes user interaction and prompt enrichment, improving generation quality without requiring model fine-tuning.

**Contrastive Language-Audio Pretraining.** CLAP [Elizalde et al., 2022] aligns audio and text in a shared embedding space using contrastive learning. We adopt the `music_audioset_epoch_15_esc_90.14.pt` checkpoint, trained on Music + Audioset + LAION-Audio-630k, because of its strong performance on music-related tasks.

## 3 Method

**Baseline: LoRA Model.** We fine-tuned LLaMA-3.1-8B-Instruct [Grattafiori et al., 2024] 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 enriched novice prompts by selecting keywords from retrieved captions on a StreamLit user interface and each novice prompt is passed through GPT-3.5

Table 1: NLP metrics evaluating rewrite adherence to novice prompts: BLEU-1 to BLEU-4 (B1–B4), lexical diversity (TTR, MTLD), and complexity (FRE).

| 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 2: Effects of rewrite versions on survey scores (OLS) and Audiobox metrics (mixed-effects model with PromptID random intercept). $\dagger p < 0.001$, $^*p < 0.1$.

| **Score** | **Intercept** | **LoRA** | **RAG** | **Adj.** $R^2$ |
|---|---|---|---|---|
| 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**† | – |

alongside the keywords to generate the RAG expert rewrite; experts prompts were used to generate audio via Stable Audio 2.0.

**Counterbalanced Design.** 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.

## 4 Results

**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.

**Survey Results.** Across all analyses, both RAG and LoRA significantly outperform novice prompts in all four metrics (See Table 2, Figure 2). 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. OLS with PromptID interaction reveals prompt-specific variation for LoRA, whereas RAG remains robust across different prompt context. Mixed Effects model with Participant as a random intercept finds negligible participant-level variance, indicating the consistency of the version effects across listeners.

**Audiobox Results.** We used the Meta Audiobox Aesthetics model [Tjandra et al., 2025] to evaluate generation quality across four perceptual dimensions: Content Usefulness (CU), Production Complexity (PC), Production Quality (PQ), and Content Enjoyment (CE). Mixed-Effects model treating PromptID as random intercept shows that both RAG and LoRA improve CU, PQ, and CE over novice prompts (Figure 3), with no gain in PC—consistent with the fact that our rewrite method do not inherently favor more audio components. In contrast, the survey's *musicality* dimension captures broader artistic qualities like genre fit and emotion, where both methods showed strong gain. 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.

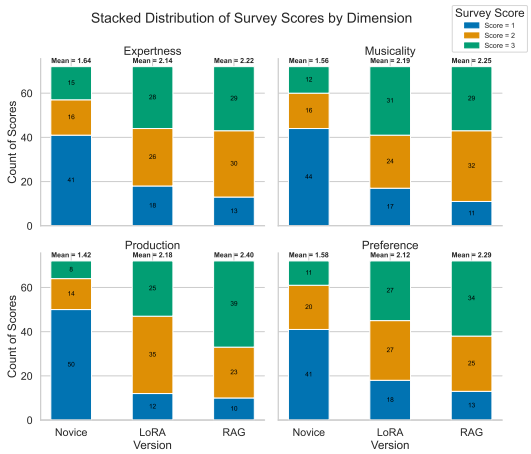


Figure 2: Stacked bar plots showing the distribution of survey scores (1–3) across rewrite versions for each evaluation dimension. Mean scores of each version are annotated above each bar. RAG yields a higher proportion of top ratings (score = 3, shown in green) compared to LoRA and Novice prompts across all evaluation dimensions.

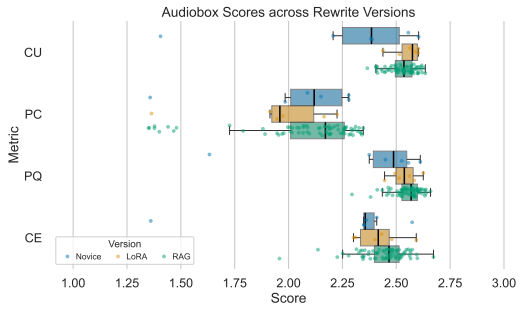


Figure 3: Audiobox evaluation scores (CU, PC, PQ, CE) for Novice, LoRA, and RAG prompts. Each point represents one audio clip; boxplots summarize distributions.

## 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.

We did not model diffusion randomness explicitly, but multiple generations from the same prompt show that RAG/LoRA rewrites yield higher mean and lower variance in CU, PQ, and CE than Novice prompts. This suggests rewrites better leverage model capacity and produce more consistent outputs and improved handling of underspecified prompts despite stochasticity. RAG also shows robust performance across a stylistically diverse prompt set (R&B, classical, pop, soul, indie, jazz).

## 6 Conclusion

Our findings show that both rewrite methods generates music that consistently score higher across subjective and objective evaluations. 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 with greater robustness across different prompt contents. 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


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.

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.

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.

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.

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

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.

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

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.

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.

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.

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

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