LoRA Training and the Fragmentation of Style in Text-to-Music Systems

Tugrul Orkun Akyol
Department of Mathematics and Computer Science
Freie Universität Berlin
tugruloa92@zedat.fu-berlin.de

Abstract

Text-to-music systems promise personalization through techniques such as Low-Rank Adaptation (LoRA), suggesting that individual musical style can be embedded into generative models. This paper examines that premise through a practice-based study using ACE-Step-1.5, training a LoRA model on a dataset derived from the author's work as a contemporary art music composer. Results indicate that stylistic features are unevenly learned and recombined, resisting recognition as a stable sonic fingerprint. This study suggests that personalization may be conditioned by the alignment between a target dataset and the model's prior stylistic distribution, with less-represented musical features remaining unstable. In response, a dataset curation pipeline based on embedding extraction, dimensionality reduction, and clustering is proposed, treating style as an emergent, fragmentable property of the dataset. This reframes model training as both compositional practice and a form of self-analysis, engaging with questions of stylistic coherence in AI-mediated music.

1 INTRODUCTION

As musical common sense continues to be explored by AI researchers, adaptivity and individualized stylistic behavior remain open challenges (Singh et al., 2024). State-of-the-art text-to-music tools such as Suno and Udio offer highly accessible interfaces that lower the threshold for music generation; however, these interfaces and their underlying apparatus are aesthetically biased and resist personalization. While features such as audio prompting enable more direct forms of interaction, they do not fully circumvent these systemic inclinations.

From the perspective of a contemporary art music composer, these limitations become particularly pronounced. The affordances of current systems often privilege metrically regular, tonally stable, and production-oriented musical forms. As a result, practices that emphasize timbral exploration, irregular phrasing, and non-standard instrumental techniques are not easily represented or recoverable through prompting alone. This creates a gap between the expressive possibilities of the interface and the compositional approaches it can meaningfully support, and raises the question of how such systems might be adapted to support more idiosyncratic compositional practices, particularly when these fall outside the statistical norms of their training data.

To address these limitations, this paper explores the use of low-rank adaptation (LoRA) as a means of personalizing a text-to-music model through a dataset derived from the author's own compositional practice. Working with the open-source ACE-Step-1.5 model, an 80-minute corpus of contemporary art music is segmented, embedded, and clustered in order to construct a more internally consistent training set. Particular attention is given to the role of captioning as a mediating layer between sound and model.

Three experimental conditions are developed: (1) a manually captioned dataset derived from the author's own compositions, segmented into longer excerpts; (2) a clustering-based approach using shorter overlapping segments and embedding-driven dataset refinement; and (3) a stylistically distinct corpus (Olivier Messiaen's *Catalogue d'oiseaux*) treated as a unified training set. Across these conditions, particular attention is given to the role of captioning, dataset consistency, and training configuration in shaping model behavior. Within this specific context, the results suggest that LoRA-based personalization remains constrained by the underlying model's learned distribution: while occasional outputs appear to reflect aspects of the training material, these effects are inconsistent and often indistinguishable from baseline generations without adaptation. The model frequently reverts to metrically regular, tonally oriented structures, indicating a strong pull toward its prior training regime. These findings do not preclude the possibility that LoRA may be more effective in stylistically aligned domains, but they highlight the difficulty of adapting current systems to compositional practices that fall outside their dominant aesthetic distributions.

2 BACKGROUND

Research on generative AI in music has expanded rapidly across several distinct trajectories. A survey of recent proceedings from the International Conference on AI and Musical Creativity reveals a strong emphasis on the technical development of large-scale generative systems, particularly text-to-music and text-to-audio models. Parallel to this, a growing body of work has

addressed legal and ethical concerns, including copyright, dataset provenance, and sustainability (Wilson et al., 2025; Cooper, 2025; Christodoulou & Jensenius, 2024). Within musicology, digital humanities, and emerging "AI Music Studies," researchers have begun to examine AI-mediated authorship, authenticity, platform infrastructures, and the socio-economic ecosystems surrounding services such as Suno and Udio (Zhao et al., 2025; Sturm et al., 2024; Oehler et al., 2025). Some studies have also explored the semiotic and cognitive dimensions of text-to-audio generation, investigating how linguistic prompts are translated into sonic form and how such systems reshape listening practices and musical signification (Coelho, 2024).

However, comparatively less attention has been paid to personalization within text-to-music systems as a structural dimension of artistic production. While developments such as audio-conditioned generation and parameter-efficient fine-tuning techniques (e.g., LoRA) enable users to adapt large models toward specific stylistic profiles, these mechanisms are typically framed in terms of controllability or output refinement. Their implications for aesthetic divergence, stylistic ownership, and the modularization of artistic identity remain underexamined. Personalization introduces a distinct layer of mediation in which style becomes trainable, transferable, and technically parameterized. This paper addresses this gap by examining personalization not as a refinement of generative performance, but as an experimental site in which questions of creative labor and aesthetic constraint are negotiated within AI-mediated musical practice.

3 METHODOLOGY

3.1 Artistic Context and Dataset

The experimental process is grounded in the author's compositional practice, which is situated within contemporary art music. The dataset consists of approximately 80 minutes of original works spanning twelve pieces for varying instrumentations, including solo violin, chamber ensemble, piano and percussion, and mixed acoustic–electronic configurations. Across these works, musical materials are characterized by the absence of stable meter and tonal hierarchy, with an emphasis instead on timbral exploration, extended instrumental techniques, and temporally expansive processes. Notable features include the use of string harmonics and extreme bow positions in solo violin works, polyrhythmic structures in ensemble settings, and sustained, slowly evolving textures shaped by dynamic swells and sparse density. While these characteristics provide a degree of aesthetic coherence, the dataset as a whole remains internally diverse in terms of instrumentation, gesture, and formal development.

3.2 Model and Training Configuration

Experiments were conducted using the ACE-Step-1.5 text-to-music model, a transformer-based system combining a language model planner with a diffusion-based audio generator. Personalization was implemented by LoRA, which introduces trainable parameter matrices into a frozen base model (Hu et al., 2021). Training was carried out on cloud infrastructure (RunPod, NVIDIA A100 80GB), as local hardware proved insufficient for the required memory demands.

Across experiments, LoRA configurations were varied to test the extent of stylistic imprinting. Models were trained with ranks of 32 and 64, learning rate of 0.0001 to 0.0003, and extended training durations (minimum 200 epochs, adjusted based on convergence behavior). Captioning strategies combined a consistent activation tag with minimal semantic descriptors intended to anchor timbral and structural features. These choices were designed to balance consistency with sufficient semantic grounding for text–audio alignment.

3.3 Experimental Design

Three experimental conditions were developed to evaluate the role of dataset structure and consistency in LoRA-based personalization.

Experiment 1: Full Dataset, Manual Captioning
The initial experiment employed the full dataset segmented into approximately one-minute excerpts. Each segment was manually labeled with descriptive captions and a shared activation tag. This approach prioritized semantic richness but retained the internal variability of the source material. It is worth noting that ACE-Step-1.5 provides an auto-captioning feature, but it unfortunately fails to perform in the context of contemporary art music.

Experiment 2: Clustering-Based Dataset Refinement
To address potential inconsistencies in the first experiment, a second condition introduced a clustering-based curation pipeline. Audio was segmented into shorter overlapping excerpts (12-second duration with 8-second hop size), embedded using LAION-CLAP, and reduced via PCA (18 components). K-means clustering (k=8) was applied, and subsets of three clusters were selected for training, both individually and in combination. In this condition, each cluster was treated as a more internally consistent dataset and assigned a shared caption consisting of a single activation tag with minimal descriptors. This approach tested the hypothesis that increased internal coherence would improve stylistic learnability.

Experiment 3: External Stylistic Corpus
A third experiment was conducted using Olivier Messiaen's Catalogue d'oiseaux as a comparatively large and stylistically consistent dataset. The corpus was treated as a unified training set with a single activation tag, allowing for comparison with the smaller and more heterogeneous datasets derived from the author's work. This condition was intended to evaluate whether limitations observed in previous experiments were attributable to dataset size and variability rather than model constraints.

3.4 Structured Comparison Protocol

To test the paper's central claims via a structured comparison, seven generation conditions were defined from combinations of three factors: adapter (none; the Experiment 1 full-dataset LoRA; the Experiment 2 Cluster 2 LoRA, selected for its internal coherence as described in Section 4.1), prompt content (activation tag alone; activation tag plus the same descriptive vocabulary used during captioning), and chain-of-thought (CoT) captioning (on/off). Four fixed

descriptor prompts, drawn from the instrumentation and texture vocabulary used across Experiment 1 adapter (e.g., "modern classical, string harmonics, bowed vibraphone, dynamic swell"), were used for all descriptor-based conditions; tag-only conditions used no additional descriptors. This vocabulary was not part of the Cluster 2 adapter's training captions, which used a single fixed tag-plus-universal-descriptor caption applied uniformly to every training segment. Repetition counts were 5 per prompt for descriptor conditions (20 generations per condition), 8 for tag-only conditions, and 3 per prompt for the two CoT conditions (12 generations per condition), for 100 generations in total. All generations used ACE-Step-1.5 with identical inference settings (8 steps, guidance scale 7) aside from the manipulated variables.

Each output was scored on audio-feature measures related to timbral similarity, periodicity, and pitch stability, although it should be acknowledged that they have limited implications in the context of contemporary art music (Section 5). Timbral/stylistic similarity was measured as cosine similarity between a generation's LAION-CLAP embedding and the centroid of (a) the full 80-minute training corpus and (b) the Cluster 2 subset used to train the Cluster 2 adapter, in the same embedding space used by the Experiment 2 clustering pipeline. Periodicity/onset density was measured with an autocorrelation-based pulse-clarity proxy (higher values indicate more regular, metered onset patterns). Pitch/tonal stability was measured with a Krumhansl-Schmuckler-style key-strength score: it measures the correlation of the mean chroma vector with major/minor key profiles, maximized over transposition, with higher values indicating a stronger tonal center. Group differences were assessed with two-sided Mann-Whitney U tests, chosen over parametric alternatives given the small, non-normal samples. Because the tag-only conditions have n = 8, associated p-values are treated as indicative rather than definitive.

4 RESULTS

4.1 Clustering and Dataset Selection

Clustering results are visualized by t-SNE projections of LAION-CLAP embeddings (Figure 1). While such projections are not strictly reliable representations of high-dimensional relationships, they provide a useful tool for assessing relative consistency within clusters. The distribution reveals several dense groupings alongside occasional outliers, suggesting partial but not absolute separation of musical materials.

Based on their internal coherence and correspondence to identifiable compositional characteristics, three clusters out of the total eight generated by K-Means were selected for further examination (Clusters 2, 3, and 5); and one of them (Cluster 2) is subjected to the structured comparison protocol. Cluster 2 is largely dominated by ensemble works featuring polytuplet writing over static harmonic fields, with sustained textures that culminate in gestural bursts and dynamic swells. Cluster 5 primarily contains excerpts from solo violin works characterized by harmonic trills, extreme bow positions, and timbral instability. Cluster 3 is centered around a trio for two pianos and percussion, marked by sparse textures, piano harmonics, and low dynamic intensity.

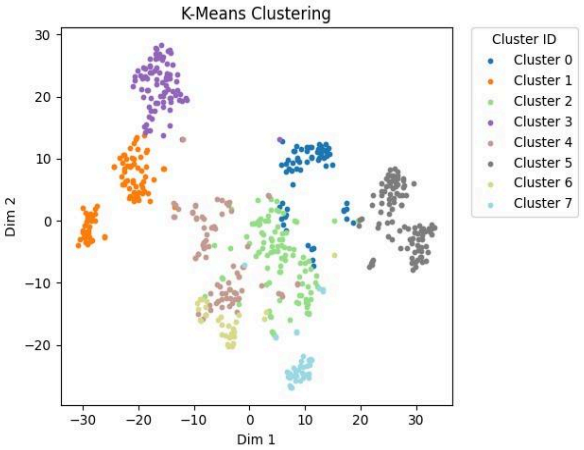

Figure 1: K-Means clustering of the dataset

Although each cluster is anchored by specific works, membership is not exclusive; excerpts from other pieces appear when timbral or gestural similarities are present. Clusters 2 and 5 exhibit relatively tighter distributions, whereas Cluster 3 appears more dispersed. Overall, the clustering process yields datasets that are more internally consistent than the full corpus, while still retaining a degree of heterogeneity.

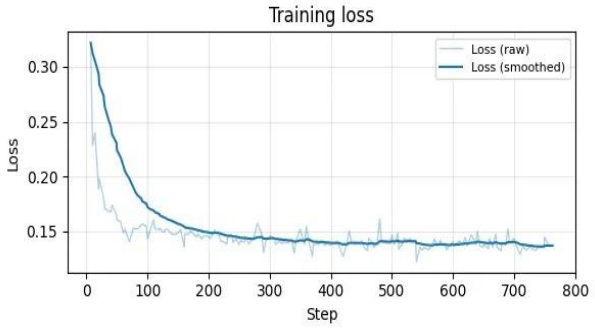

Figure 2: Training curve for Experiment 2, Cluster 5

4.2 Training Dynamics

Training loss curves for LoRA adaptation (Figure 2) show consistent convergence across configurations. For rank 64, loss values typically decrease from initial levels to approximately 0.13, whereas rank 32 models converge further to around 0.07.

This difference may be attributed to the reduced representational capacity of lower-rank adaptations, which can lead to faster or more stable minimization of the training objective. However, lower loss values do not necessarily correspond to improved perceptual or stylistic alignment, and no clear correlation between loss magnitude and qualitative output was observed.

4.3 Outputs

Outputs from the initial experiment demonstrate occasional alignment with isolated gestural and textural features present in the training material. They exhibit characteristics such as sustained sounds within sparse textures and internal fluctuations within relatively static timbral fields along with intermittent gestural bursts. In singular cases, percussive hits accenting phrase endings and amplitude-modulation-like behaviors emerge,

loosely recalling techniques used in both the electronic and instrumental components of the dataset.

Having said that, these correspondences remain partial and fragmentary. The generated outputs are generally crude in musical quality and lack the structural coherence, temporal sensitivity, and timbral precision that characterize the source material. Rather than constituting a recognizable stylistic imprint, the results suggest only a superficial approximation of certain features, often appearing as incomplete or unstable formations.

While some of these traits are difficult to elicit through prompting alone with the model showing strong tendency toward metrically regular, beat-driven, and tonally grounded outputs, the presence of such features after LoRA training does not translate into consistent or reliable stylistic behavior. In many cases, outputs remain dominated by generic patterns that bear limited relation to the training corpus. Section 4.4 reports a structured test of this pattern: in the two direct baseline comparisons, the LoRA adapters did not significantly increase similarity to the training corpus over baseline, giving the impression that apparent stylistic alignment is inconsistent rather than systematic.

Results from the clustering-based dataset (Experiment 2) and the external corpus (Experiment 3) follow similar patterns. In Experiment 2, cluster-specific training introduces a degree of stabilization at the level of instrumentation and timbre; for example, violin-based outputs become more frequent when training on Cluster 5. This suggests that certain descriptive anchors (e.g., instrument labels) are effectively learned and reproduced. However, this increased consistency does not extend to deeper structural or stylistic features. The model does not reliably capture the temporal, gestural, or formal characteristics that define the source material. Instead, outputs tend to revert toward familiar generative patterns, particularly those associated with tonal organization and periodic rhythm.

A comparable tendency is observed in the Messiaen dataset (Experiment 3). While the model produces piano-like timbres consistent with the training material, it does not reproduce the harmonic language, rhythmic complexity, or formal processes characteristic of the source works. Generated outputs instead gravitate toward simplified, tonally conventional textures.

4.4 Structured Comparison Results

Table 1 reports descriptive statistics for the four metrics across all seven conditions; Figures 3–6 show the corresponding distributions. Full pairwise test results for all four metrics across all eight comparisons are provided as supplementary material along with the generated audio[1]. Table 2 lists the five comparisons that reached significance (Mann-Whitney U, two-sided, $p < .05$).

Table 1: Mean ± SD for each metric, by condition

Condition	n	CLAP–corpus	CLAP–cluster2	Pulse clarity	Key strength
A- Baseline (no LoRA)	20	0.5254 ± 0.0865	0.4467 ± 0.1006	0.7909 ± 0.0788	0.9701 ± 0.013
B- Exp1: LoRA, tag only	8	0.4258 ± 0.0741	0.3785 ± 0.0995	0.8531 ± 0.0814	0.9693 ± 0.0099
C- Exp1: LoRA, tag + descriptor	20	0.4984 ± 0.1258	0.447 ± 0.1351	0.8468 ± 0.0949	0.9688 ± 0.0137
D- Cluster2: LoRA, tag only	8	0.4054 ± 0.0901	0.3363 ± 0.0948	0.7814 ± 0.0429	0.9503 ± 0.0353
E- Cluster2: LoRA, tag + Descriptor	20	0.5506 ± 0.0824	0.4501 ± 0.0954	0.7811 ± 0.0745	0.9647 ± 0.0151
F- Baseline, CoT on	12	0.4586 ± 0.0634	0.3899 ± 0.0763	0.7643 ± 0.1032	0.9667 ± 0.0251
G- Exp1: LoRA, CoT on	12	0.4473 ± 0.1693	0.4018 ± 0.181	0.8199 ± 0.0793	0.9729 ± 0.0168

Table 2: Comparisons reaching significance (Mann-Whitney U, two-sided)

Comparison	Metric	n (A/B)	Median A	Median B	p
A vs C	Pulse clarity	20 / 20	0.800	0.857	.036
C vs E	Pulse clarity	20 / 20	0.857	0.780	.022
D vs E	CLAP–corpus	8 / 20	0.413	0.556	.0005
D vs E	CLAP–cluster2	8 / 20	0.303	0.446	.010
A vs F	CLAP–corpus	20 / 12	0.552	0.441	.012

Two results support the paper's personalization claims. Neither LoRA adapter produced a significant increase in similarity to the training corpus relative to the unadapted baseline (A vs. C: $p = .474$; A vs. E: $p = .379$), which is consistent with the qualitative observation in Section 4.5 that adapter training does not reliably move outputs toward the training corpus. At the same time, adding descriptor text to the Cluster 2 activation tag produced a large, significant increase in similarity to both the full corpus and the Cluster 2 subset (D vs. E: $p = .0005$ and $p = .010$ respectively), while the equivalent comparison for the Experiment 1 tag was not significant (B vs. C, CLAP–corpus: $p = .150$). This asymmetry is informative because the two adapters were trained under different captioning systems. The Experiment 1 adapter was trained with the same kind of per-segment descriptive captions used to construct prompts 1–4, so both its tag-only condition (B) and its tag-plus-descriptor condition (C) are broadly consistent with what the adapter saw during training, which likely explains why adding descriptors made little further difference. The Cluster 2 adapter, by contrast, was trained with a single fixed caption applied uniformly across all training segments; its tag-only condition (D) exactly reproduces that training caption, while its tag-plus-descriptor condition (E) introduces text the adapter never encountered paired with this audio. Still, E nonetheless

[1] For supplementary material, see the folder below: https://drive.google.com/drive/folders/1NkUUSs3RZx4cE6lEUfxEvUuhYUz1GcHY?usp=sharing

shows significantly higher corpus similarity than D. Notably, E only recovers baseline-level similarity rather than exceeding it (A vs. E: p = .379), so the increase is most plausibly attributed to the base model's own prior association of this descriptive vocabulary with corpus-adjacent acoustic features, rather than to adapter-specific learning. The two remaining significant comparisons concern pulse clarity. Relative to baseline, the Experiment 1 adapter significantly increased pulse clarity (A vs. C: p = .036): training on a corpus that largely avoids periodic onset patterns yielded outputs that were more, not less, regularly pulsed, demonstrating the pull toward the base model's prior described in Section 4.5. Outputs under the Cluster 2 adapter were significantly less pulsed than under the Experiment 1 adapter (C vs. E: p = .022), but remained at baseline levels rather than approaching the aperiodicity of the source material.

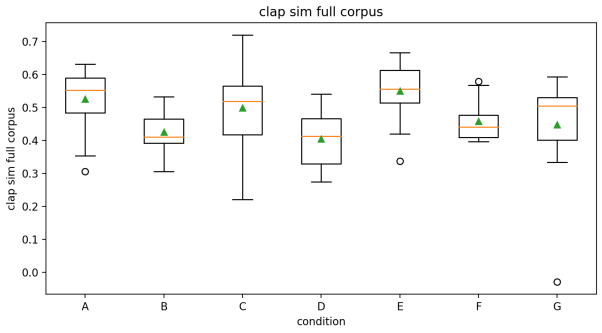

Figure 3: CLAP similarity to the full training-corpus centroid, by condition.

Figure 4: CLAP similarity to the Cluster 2 subset centroid, by condition.

The two designed tests of the chain-of-thought manipulation (A vs. F, C vs. G) did not produce the predicted shift toward a more regular meter or a stronger tonal center: neither pulse clarity nor key strength differed significantly with CoT on, in either the baseline or the Experiment 1-adapter condition. CoT did coincide with a significant drop in full-corpus similarity for the baseline condition (A vs. F: p = .012), an effect not anticipated by the original design and not evident under the Experiment 1 adapter (C vs. G: p = .521); it is reported here without a settled explanation and flagged as a direction for follow-up. Under the Experiment 1 adapter specifically, the author's own listening impression of the C-vs-G pairing is that turning CoT on yields more repetitive, beat-driven structuring. This is not reflected in the pulse-clarity proxy, which tracks onset-level periodicity rather than phrase- or motif-level repetition.

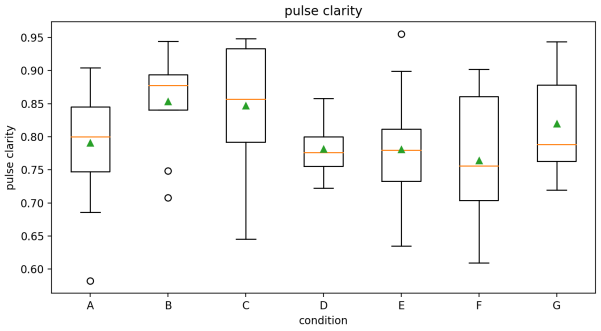

Figure 5: Pulse clarity (onset-periodicity proxy), by condition.

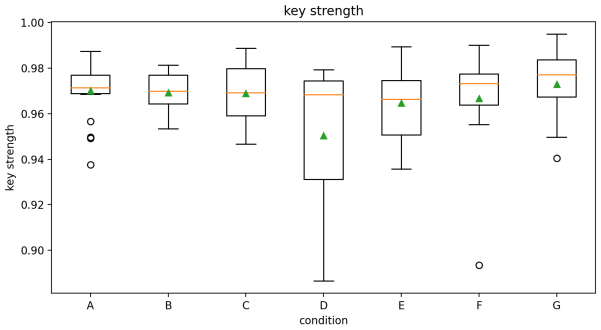

Figure 6: Key strength (tonal-center proxy), by condition.

These results should be read with two caveats. First, the tag-only conditions (B, D) have n = 8; comparisons resting on them (B vs. C, D vs. E) are likely under-powered even where nominally significant, and are reported as indicative rather than conclusive. Second, CLAP similarity, pulse clarity, and key strength are proxy measures validated primarily for popular/tonal music retrieval and rhythm analysis, not for the extended techniques, aperiodic gesture, and non-tonal pitch organization characteristic of the source corpus; as such they are reported alongside, rather than in place of, the qualitative evaluation of the author. Key strength in particular is effectively at ceiling across all conditions (0.95–0.97), and therefore carries little discriminative value in this comparison.

4.5 Summary of Findings

Across all experimental conditions, LoRA adaptation produces limited and inconsistent stylistic influence. While certain surface-level features like instrumentation and isolated timbral or gestural traits can be encouraged, these do not accumulate into a coherent or recognizable stylistic imprint.

Crucially, comparison with baseline generations (without LoRA) suggests that these features are not specific to the trained material. Similar outputs can be obtained through prompting alone, and the presence of such traits does not indicate that the model has internalized the compositional logic, temporal organization, or aesthetic

priorities of the dataset. In this sense, the observed correspondences remain superficial and non-distinctive, rather than constituting evidence of successful personalization.

Rather than producing a clear transformation of the model's output space, LoRA training appears to have a limited effect on the generation process in this context. The model continues to favor structurally familiar patterns such as tonal organization and metrically regular behavior, even when trained on material that systematically departs from these conventions.

5 CONCLUSION AND DISCUSSION

The results presented here suggest that LoRA-based personalization, as implemented in ACE-Step, has limited capacity to produce meaningful stylistic transfer in the context of contemporary art music. While certain surface-level features can be encouraged, these do not cohere into a recognizable or stable aesthetic profile. Instead, the model exhibits a persistent tendency to revert toward structurally familiar patterns, particularly tonal organization and metrically regular behavior, even when trained on material that systematically departs from these conventions.

At the same time, these findings should not be interpreted as a general limitation of LoRA as a technique. Rather, they point to the importance of the relationship between the target dataset and the model's prior distribution. It is plausible that LoRA-based adaptation is more effective when the desired style is already well represented within the model's training corpus, allowing fine-tuning to reinforce or specialize existing tendencies. In contrast, when the target material diverges significantly from those priors, the capacity for stylistic transfer appears limited. Under this interpretation, personalization does not operate as an open-ended mechanism for imprinting arbitrary styles, but as a modulation of tendencies already embedded within the model.

This raises broader questions regarding the aesthetic scope implicitly privileged by contemporary text-to-music systems. Earlier work in music AI often developed in dialogue with experimental and research-oriented musical practices, as exemplified by systems such as David Cope's EMI (Cope, 1989) and George Lewis' *Voyager* (Lewis, 2000). In contrast, many current systems emerge within commercial and platform-driven contexts, where accessibility, scalability, and market relevance play a central role. As a result, the musical assumptions embedded in these models may align more closely with widely circulating musical forms than with less standardized or structurally heterogeneous practices. This orientation is evident not only in the composition of training datasets but also in interface design. For instance, the auto-captioning functionality demonstrates a consistent misalignment with the datasets used in this study, revealing underlying biases in how musical content is interpreted and categorized.

Within this context, the pipeline proposed in this study that combines embedding-based clustering, dataset refinement, and LoRA training, should be understood as an exploratory framework rather than a failed solution. Although it did not yield strong results here, the approach remains conceptually viable and may prove effective when applied to models whose priors are more compatible with the target material. The results therefore reflect not only the limits of the method, but also the constraints imposed by the underlying system.

Finally, these experiments raise questions about the nature of musical style in AI-mediated contexts. While LoRA is often framed as a means of encoding a "personal sound," the findings suggest that style is not easily reducible to a set of transferable features. The clustering process, in particular, points toward a different understanding of style. In that context, style is not understood as a fixed and unified identity, but as an emergent and potentially fragmentable property of a corpus. By isolating subsets of material based on timbral and gestural similarity, clustering functions as a form of computational self-analysis, revealing latent consistencies and divergences within a body of work. The results indicate that such consistencies are more readily captured at the level of local attributes such as timbre, texture, and gesture, and fail at the level of higher-order musical organization, including phrasing, temporal structure, or harmonic language.

In this sense, the study highlights a tension between the conceptual flexibility of style as understood in compositional practice and its operationalization within current generative systems. While AI tools offer new forms of interaction and mediation, their capacity to engage with complex, non-standardized musical languages remains uneven. Further work is needed to explore how such systems might accommodate a broader range of aesthetic practices, and how personalization techniques can move beyond surface-level adaptation toward more substantive forms of stylistic engagement.

AUTHOR DECLARATION

The author used AI-assisted tools for grammar and phrasing improvement. All intellectual content, analysis, and conclusions are the author's own. The author takes full responsibility for the integrity of the work.

REFERENCES

Christodoulou, A.-M., & Jensenius, A. R. (2024). Navigating challenges in multimodal music data management for AI systems. In Proceedings of the 5th International Conference on AI and Musical Creativity (AIMC 2024). https://aimc2024.pubpub.org/pub/4y99xcm3/release/1

Coelho, G. (2024). Semantic and semiotic interplays in text-to-audio AI: Exploring cognitive dynamics and musical interactions. In Proceedings of the 5th International Conference on AI and Musical Creativity (AIMC 2024). https://aimc2024.pubpub.org/pub/44f6hknl/release/5

Cooper, Z. (2025). She's lost control again: The next generation of copyright challenges over next-generation interactive music. In Proceedings of the 6th International Conference on AI and Musical Creativity (AIMC 2025). Zenodo. https://doi.org/10.5281/zenodo.17041710

Cope, D. (1989). Experiments in musical intelligence (EMI): Non-linear linguistic-based composition. *Interface, 18*(1–2), 117–139. https://doi.org/10.1080/09298218908570541

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-rank adaptation of large language models* (arXiv:2106.09685) [Preprint]. arXiv. https://arxiv.org/abs/2106.09685

Lewis, G. E. (2000). Too many notes: Computers, complexity, and culture in *Voyager*. *Leonardo Music Journal, 10*,

33–39. https://doi.org/10.1162/096112100570585

Oehler, M., Schepmann, J., & Saurbier, B. (2025). Assessing expectancy bias in listener evaluations of AI and human music compositions. In Proceedings of the 6th International Conference on AI and Musical Creativity (AIMC 2025). Zenodo. https://doi.org/10.5281/zenodo.16955782

Singh, N., Mishra, M., & Machover, T. (2024). *AI for musical discovery: An MIT exploration of generative AI*. https://doi.org/10.21428/e4baedd9.8fa181e9

Sturm, B. L. T., Déguernel, K., Huang, R. S., Kaila, A.-K., Jääskeläinen, P., Kanhov, E., Cros Vila, L., Dalmazzo, D. C., Casini, L., Bown, O. R., Collins, N., Drott, E., Sterne, J., Holzapfel, A., & Ben-Tal, O. (2024). AI music studies: Preparing for the coming flood. In Proceedings of the 5th International Conference on AI and Musical Creativity (AIMC 2024). https://aimc2024.pubpub.org/pub/ej9b5mv1/release/2

Wilson, E., Wszeborowska, A., & Bryan-Kinns, N. (2025). A short review of responsible AI music generation. In Proceedings of the 6th International Conference on AI and Musical Creativity (AIMC 2025). Zenodo. https://doi.org/10.5281/zenodo.16946342

Zhao, Y., Kanhov, E., & Sturm, B. L. T. (2025). Methodological considerations of digital ethnographic studies in the AI music field. In Proceedings of the 6th International Conference on AI and Musical Creativity (AIMC 2025). Zenodo. https://doi.org/10.5281/zenodo.16946368