Music for All: Representational Bias and Cross-Cultural Adaptability of Music Generation Models

Atharva Mehta Shivam Chauhan Amirbek Djanibekov
Atharva Kulkarni Gus Xia Monojit Choudhury
Mohamed bin Zayed University of Artificial Intelligence
{atharva.mehta, shivam.chauhan, amirbek.djanibekov,
atharva.kulkarni, gus.xia, monojit.choudhury}@mbzuai.ac.ae

Abstract

The advent of Music-Language Models has greatly enhanced the automatic music generation capability of AI systems, but they are also limited in their coverage of the musical genres and cultures of the world. We present a study of the datasets and research papers for music generation and quantify the bias and under-representation of genres. We find that only 5.7% of the total hours of existing music datasets come from non-Western genres, which naturally leads to disparate performance of the models across genres. We then investigate the efficacy of Parameter-Efficient Fine-Tuning (PEFT) techniques in mitigating this bias. Our experiments with two popular models – MusicGen and Mustango, for two underrepresented non-Western music traditions – Hindustani Classical and Turkish Makam music, highlight the promises as well as the non-triviality of cross-genre adaptation of music through small datasets, implying the need for more equitable baseline music-language models that are designed for cross-cultural transfer learning. The code for the paper is available at our Github Repository and the model adapters are available at Huggingface.

1 Introduction

Music, as a powerful expression of cultural identity, is deeply embedded in traditions (Swain, 1995; Chung, 2006). Recent advancements in AI, powered by deep learning models (Schneider et al., 2024; Copet et al., 2023; Tal et al., 2024), have led to significant improvements in automatic music generation technologies. This progress has led to several music generation playgrounds such as Jukebox (Radford et al., 2020), Suno[1], and Udio[2] offering users the ability to generate music according to their specifications. However, these models often reflect biases, particularly towards Western musical traditions (Tao et al., 2024; Copet et al., 2023), in their training data.

This lack of diversity in datasets, as outlined by Copet et al. (2023); Melechovsky et al. (2024); Radford et al. (2020), is also evident in the disparate performance of the music generation models across genres. More specifically, the models tend to rely on Western tonal and rhythmic structures when generating music for non-Western genres, such as Indian or Middle Eastern music. The situation is comparable to the lack of cultural and linguistic diversity (Joshi et al., 2020; Bender and Friedman, 2018; Bender et al., 2021) in NLP research.

In order to quantify the severity of this problem in music generation research landscape, we conduct a comprehensive analysis of existing music datasets and music generation papers, which reveals a stark disparity in the representation of non-Western music. Particularly noteworthy is the scarcity of non-Western music data, with merely 5.7% of the total hours of the available datasets. This finding highlights the need for more diverse musical datasets and methods to adapt state-of-the-art models to low-resource genres.

However, it remains unclear whether cross-genre music adaptation, similar to cross-lingual adaptation, can be effectively achieved using lightweight computational techniques such as parameter-efficient fine-tuning (PEFT) (Houlsby et al., 2019). In this paper, we explore this question by adapting two open-source models, MusicGen (Copet et al., 2023) and Mustango (Melechovsky et al., 2024) for two low-resource non-Western genres - *Hindustani Classical* [3] music of India and *Makamat* [4] music of the Middle East.

[1] https://suno.com/
[2] https://www.udio.com/

[3] Hindustani Classical music is a traditional system of music that emphasizes melodic development based on ragas (melodic frameworks) and talas (rhythmic cycles).

[4] Makam, in traditional Arabic music, is a melodic mode system defining pitches, patterns, and improvisation, central to Arabian art music, with 72 heptatonic scales

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 4569–4585

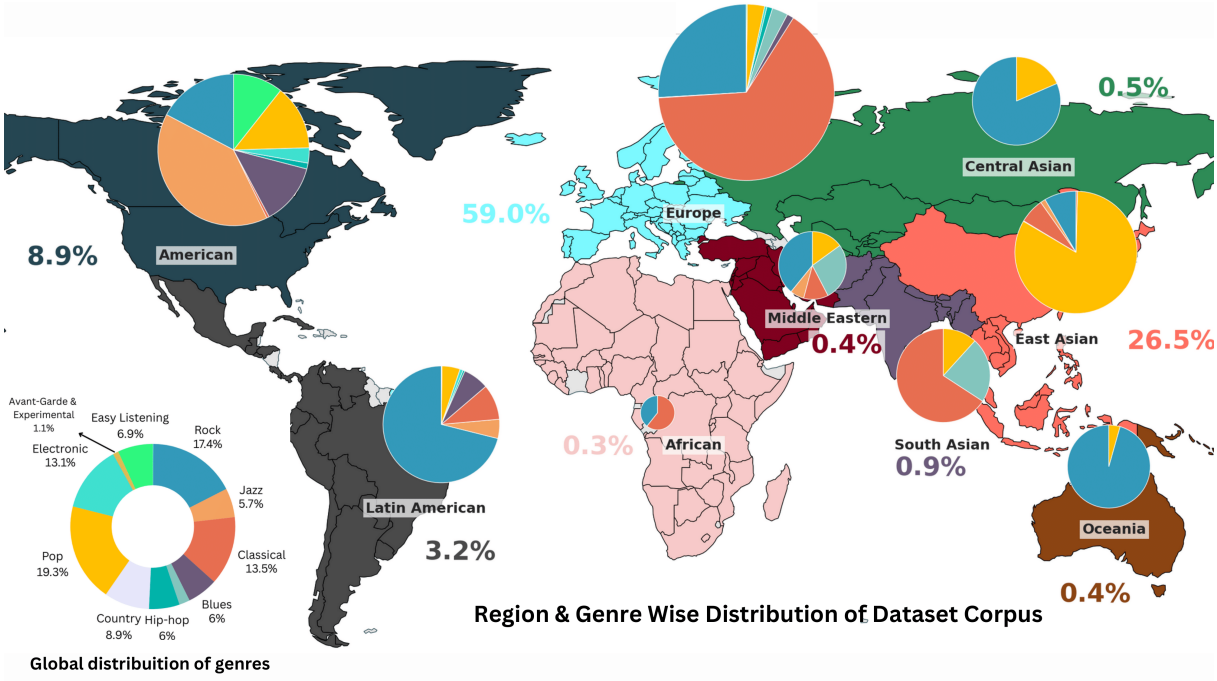

Figure 1: The bottom left piechart shows the global distribution of genre. Each piechart in the map shows the distribution of genres in different regions with the size of each piechart being proportional to their contribution to the data corpus.

We hypothesize that explicit fine-tuning with a small number of additional parameters (less than 1% of the pre-trained model) as adapters (Bapna and Firat, 2019), will lead to a far better performance for the under-represented music style. We conduct a series of experiments comparing the performance of baseline models and our adapter-enhanced models on objective metrics. For human evaluations, each model is tested using our novel evaluation framework roughly based on Bloom's Taxonomy (Armstrong, 2010): **Recall, Analysis**, and **Creativity** evaluate a model's audio generation. Recall reproduces trained entities, Analysis forms new combinations of them, and Creativity blends different entities across genres in novel, unseen ways. Evaluations are conducted in a play-arena style, ranking models based on their adherence to the text prompt in terms of *rhythm*, *instrument*, *melody* and *creativity*. Mustango shows improvement when finetuned on Hindustani Classical music by 8% and MusicGen shows improvement by 4% on Turkish Makam in ELO ratings from their respective baselines.

Our results show that while PEFT techniques are effective in improving the overall quality of generated music for the under-represented genre over the baseline models, not all models are adaptable to all genres. This implies that the various design choices made in the architecture, and the training datasets and recipes for the base model are crucial determinants of the adaptability of a model to certain musical genres.

The contributions of this paper are threefold:

1. We provide a detailed analysis of the current state of musical datasets, highlighting the under-representation of non-Western music.
2. We present the first application of parameter-efficient training with adapters for cultural adaptation of under-represented genres in music generation models.
3. We introduce a novel arena-style evaluation framework based on Bloom's Taxonomy to assess the text to music generation capabilities of models using a play-arena style, ranking models on on their adherence to the text prompt on *rhythm*, *instrument*, *melody* and *creativity*.
4. We demonstrate that while adapting base models to different genres is possible, it is a non-trivial challenge.

The rest of the paper is organized as follows: In Section 2, we discuss global disparities in music representation, followed by Section 3, which details our approach to adapting genres. Section 4

presents our evaluation methodology and the results obtained from our analysis. We conclude our findings in Section 5.

2 The Disparity in Music Generation Research

AI music generation has evolved rapidly with techniques such as autoregressive (Agostinelli et al., 2023; Copet et al., 2023; Ziv et al., 2024), diffusion-based (Schneider et al., 2024; Huang et al., 2023; Li et al., 2024) and GAN-based (Dong et al., 2018; Li and Sung, 2021) producing high-quality music. Some of the works include adapter-based settings which proved effective for music editing and inpainting (Lin et al., 2024; Zhang et al., 2024). Moreover, Lan et al. (2024) used adapters for rhythm and chord conditioning. Tan et al. (2020) showcased how visual emotions from images can be effectively translated into music using deep learning techniques.

Drawing inspiration Joshi et al. (2020), which systematically analyzes the under-representation of languages spoken by the global majority, we conduct a survey of the datasets and research papers on music generation.

2.1 Data Collection

To get our initial pool of papers, we implemented an efficient, automated data collection method.

We employed a multi-stage, keyword-based selection method, leveraging the Scholarly package Cholewiak et al. (2021) to gather approximately 5000 papers. This included up to 1000 papers per query, using broad search terms such as "music," "music generation," "non-Western music," "MIDI," and "symbolic music." We then refined our selection by focusing on papers presented at 10 major conferences including *IJCAI, AAAI, ICML, EURASIP, EUSIPCO, ISMIR, NeurIPS, SMC, NIME* and *ICASSP*, chosen based on their popularity and prestige in the area of computational processing of music, narrowing our pool to around 800 papers. Conferences such as ISMIR and NIME specialize in music information retrieval and musical expression, frequently showcasing work related to generative AI. Additionally, conferences like ICASSP, AAAI, and NeurIPS are known for their focus on cutting-edge AI technologies, such as GANs and transformers, which are crucial for music generation.

2.1.1 Dataset Papers

To identify papers proposing datasets, we read through the title and abstract of each paper. This led to a set of **152** papers proposing new datasets with a total of **1 million+** hours of music. These datasets were manually annotated for the region and genres covered, total hours of music data, and whether the dataset is annotated with other details (such as, instruments, genre, and style). Papers that directly provided details of the distribution of data points across genres and regions, were analyzed with the already available statistics. Unfortunately, several datasets did not offer substantial details necessary for our study. If such a dataset had more than ***10,000*** hours of audio data, we analyzed each sound file's metadata to collect genre and region information. However, when the genre and region were not explicitly mentioned in either the paper or the metadata, we did not make any assumptions; thus, **7.9%** of the datasets totaling **5,772** hours were excluded from our analysis.

2.2 Findings

Our findings are summarized in Figure 1. The results reveal an almost complete omission of musical genres from non-Western countries, especially those from the Global South. Approximately 94% of the total hours in available datasets are dedicated to music from the Western world, while only 5.7% are devoted to *South Asian, Middle Eastern, Oceanian, Central Asian, Latin American*, and *African* music combined. This imbalance is likely to cause poor-quality music generation for genres from the Global South. For detailed analysis, please refer to Appendix A.

3 Genre Adaptation: Data, Models and Experiments

For our genre-adaptation experiments, we selected two distinct non-Western genres — Hindustani Classical (Jairazbhoy, 1971) and Turkish Makam (Signell, 2008) — both significantly underrepresented in music generation research and datasets, and two open source models – MusicGen (Copet et al., 2023) and Mustango (Melechovsky et al., 2024). We begin by describing the dataset creation, followed by prompt generation, the models, adapter architectures and finally, the training process.

3.1 Dataset Creation

Our study necessitated diverse corpus of non-Western music with detailed metadata. The Dunya (Porter et al., 2013) which is part of the CompMusic project (Serra, 2014), emerged as the ideal choice, offering an extensive collection of over 1,300 hours of music across multiple non-Western genres. This corpus includes Carnatic, Hindustani, Turkish Makam, Beijing opera, and Arab Andalusian music, providing a broad spectrum of cultural music. We focused specifically on Hindustani Classical and Turkish Makam genres as both genres possess complex culturally specific melodic and rhythmic structures different from Western music & we had easier access to listeners familiar with Indian and Turkish music. For Hindustani Classical, we chose the MTG Saraga (Srinivasamurthy et al., 2021) annotated dataset which is built on CompMusic offering 50 hours of audio. For Turkish Makam, we use the Dunya dataset API for accessing the metadata and audio samples leading to 405 hours of audio.

To ensure consistency and manage computational resources effectively, we implemented several pre-processing steps. We standardized the audio sample length by truncating longer recordings to 30 seconds. We utilized the accompanying metadata from the Dunya corpus without modification. These descriptions, rich in genre-specific details, served as valuable inputs for creating prompt templates. Finally, to accommodate the differing requirements of our chosen models, we performed audio resampling. Specifically, for MusicGen we resampled the audio to a 32 kHz sampling rate and for Mustango 16 kHz sampling rate.

The metadata from the dataset provides genre-specific information for each audio clip, including three key details critical to our study: melodic line, rhythmic pattern, and instrumentation. For the melodic line, we extracted the *raga* (a melodic framework in Hindustani Classical music) and *Makam* (a system of melodic modes in Turkish music). For rhythmic patterns, we identified *laya* (tempo) in Indian music and *usul* (a sequence of rhythmic strokes) in Turkish music. Additionally, we extracted the meta-data for the instruments (including voice) played in each audio sample. Details of the dataset can be found in Appendix F.

After pre-processing, we collected a total of 23.24 hours of audio for Hindustani Classical music and 121.16 hours for Turkish Makam music.

Query Type	Example
Recall	Imagine a traditional ★ **Makam** performance that brings together ▷ Clarinet, ▷ Darbuka, ▷ Kanun, ▷ Oud, ▷ Voice, ♯ Aksak makam, and ♭ Hicaz usul, flowing effortlessly.
Analysis	Imagine a traditional ★ **Makam** performance that brings together ▷ Tanbur, ▷ Oud, ▷ Cello, with the flowing essence of ♯ Aksak makam and ♭ Fahte usul, flowing effortlessly.
Creativity	Imagine a modern ★ **Western Electronic Dance Music (EDM)** performance infused with the soulful sound of ▷ Tanbur, rich vocals blending with ♯ Acem makam and ♭ Fahte usul.

Table 1: Recall, Analysis & Creativity Queries: Recall uses known combinations, while Analysis introduces novel combinations to test analytical capability and Creativity introduces cross-genre combinations. Refer to Section 4. **Genre**:★, melodic line:♯, rhythmic pattern:♭, and instrumentation:▷.

The dataset was then divided using an 80-20% split for training and testing, allowing us to evaluate the final model performance effectively. We ensured that audio clips for training and testing come from different songs to prevent distribution overlap in train and test. This split resulted in 18.91 hours of Hindustani Classical music and 97.23 hours of Turkish Makam music for training, and the remaining portions reserved for testing.

3.2 Prompt Generation

To create effective prompts for model training, we created three distinct templates that describe each musical piece based on sample metadata from the selected genres.

For each audio sample, we randomly selected one of the three templates and populated it with relevant metadata attributes as shown in Table 1. This process ensures that each prompt captures the unique musical elements of the sample. By maintaining this metadata-specific structure across prompts, we help the model learn to identify and respond to key attributes within each genre, enabling it to generate more accurate and culturally informed outputs during training.

3.3 Models

We utilize two state-of-the-art models, MusicGen (Copet et al., 2023) and Mustango (Melechovsky et al., 2024), to explore cross-genre adaptation. MusicGen is a transformer-based model,

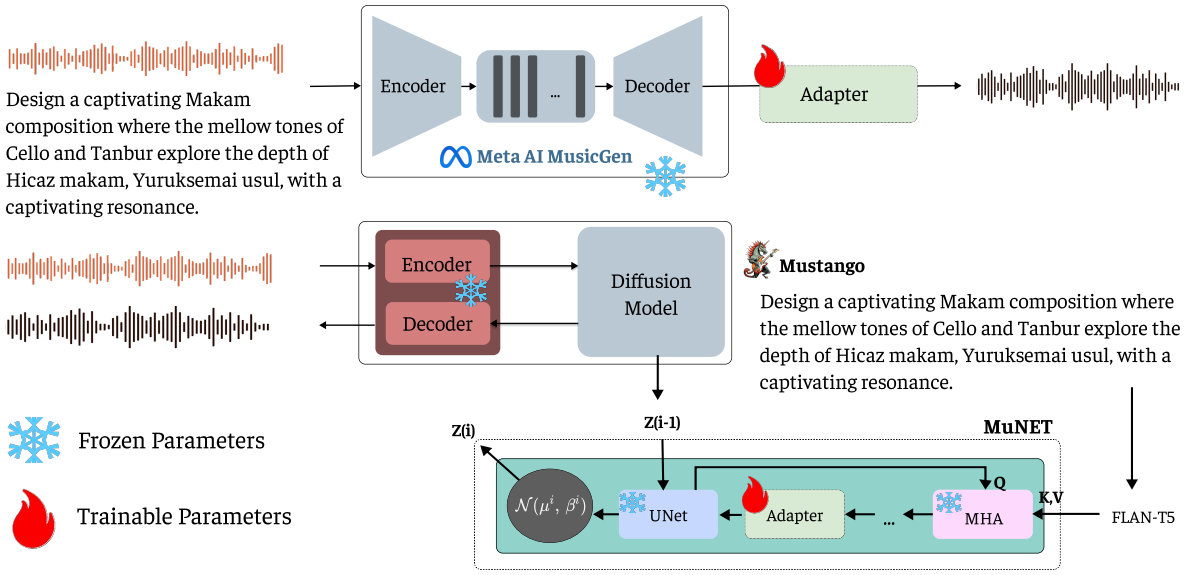

Figure 2: Mustango & MusicGen settings for low-resource fine-tuning.

while Mustango integrates both diffusion and transformer architectures. We introduce adapters (Pfeiffer et al., 2020) that enable low-resource fine-tuning. We also considered Moûsai (Schneider et al., 2024) and MusicLM (Agostinelli et al., 2023), but Moûsai and MusicLM lack open-source training codes.

3.3.1 MusicGen

In MusicGen, we enhance the model with an additional 2 million parameters by integrating **Bottleneck Residual Adapter** after the transformer decoder within the MusicGen architecture after thorough experimentation with other placements. The total parameter count of MusicGen is 2 billion, making the adapter only 0.1% of the total size. The adapter, as shown in Figure 2, consists of a linear layer that compresses the embedding to a very small dimension, followed by a non-linear activation and projection back to the original size.

MusicGen leverages the Encodec (Défossez et al., 2023) framework, which compresses audio into latent representations. These latent representations are processed through a transformer model, which generates new music based on input prompts. By placing adapters at the end of the decoder, we achieve a lightweight adaptation mechanism that enhances the model's ability to generate music in specific styles or regions, such as Hindustani Classical music and Makam, without modifying the fundamental Encodec structure.

3.3.2 Mustango

In Mustango, we enhance the model with an additional 2 million parameters, which represents only 0.1% of the model's total parameter count, by integrating a **Bottleneck Residual Adapter**.

While Mustango supports chord and beat embeddings, we opted not to use them here due to the distinct focus of Hindustani Classical and Turkish Makam on melodic lines rather than harmonic progressions. Unlike Western classical music, these genres feature complex, rhythms with accents often within a single beat, making fixed beat and chord embeddings difficult to apply.

The adaptation process in Mustango begins with the FLAN-T5 (Chung et al., 2024) model, which converts the input text into embeddings. These embeddings are then incorporated into the UNet architecture (Ronneberger et al., 2015) through a cross-attention mechanism, aligning the text and audio components. To refine this process, a Bottleneck Residual Adapter with convolution layers is incorporated into the up-sampling, middle, and down-sampling blocks of the UNet, positioned immediately after the cross-attention block at the end of each stage (Figure 2). The adapters reduce channel dimensions by a factor of 8, using a kernel size of 1 and GeLU activation after the down-projection layers to introduce non-linearity. Various adapter configurations and placements were explored to preserve the musical structure while adapting stylistic elements, with this setup yielding the best output quality. This design facilitates cultural adaptation while preserving computational efficiency.

Objective Metrics				
Hindustani Classical Music				
Model	**FAD ↓**	**FD ↓**	**KLD ↓**	**PSNR ↑**
MGB	40.05	75.76	6.53	16.23
MGF	40.04	72.65	6.12	16.18
MTB	6.36	45.31	2.73	16.78
MTF	**5.18**	**22.03**	**1.26**	**17.70**
Turkish Makam				
Model	**FAD ↓**	**FD ↓**	**KLD ↓**	**PSNR ↑**
MGB	39.65	57.29	7.35	14.60
MGF	39.68	56.71	7.21	14.46
MTB	8.65	75.21	6.01	**16.60**
MTF	**2.57**	**20.56**	**4.81**	16.17

Table 2: Objective Evaluation Metrics for Hindustani Classical Music and Turkish Makam.

3.4 Training settings

For MusicGen fine-tuning, we used two RTX A6000 GPUs over a period of around 10 hours. The adapter block was fine-tuned, using the AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate of 5e-5 and a weight decay of 0.05 using MSE based Reconstruction Loss. The training spanned 20 epochs, with a patience threshold of 5 epochs for early stopping based on validation loss. We utilized a batch size of 4 and applied gradient clipping with a maximum norm of 1.0. The training data was split into 90% for training and 10% for validation.

For Mustango model fine-tuning, we used one RTX A6000 GPU over a period of 12 hours. The adapter block was fine-tuned, using the AdamW optimizer with a learning rate of 4.5e-5 and a weight decay of 0.01 using MSE based Reconstruction Loss. The training spanned 25 epochs for both genres, with a patience threshold of 5 epochs for early stopping based on validation loss. The training data was split into 80% for training and 20% for validation.

4 Results

We evaluated four models, Mustango Baseline (MTB), Mustangto Fine-tuned (MTF), MusicGen Baseline (MGB), and MusicGen Finetuned (MGF), on two genres using both objective metrics and human evaluation, providing both objective and subjective insights into model performance.

4.1 Automatic Metrics

We sample 400 audio samples from the test set to form our test prompt corpus. For capturing the distance between generated audio and the test corpus we compute Fréchet Audio Distance(FAD), Fréchet Distance(FD) and Kullback-Leibler(KL) with Sigmoid activation. We utilize the AudioLDM (Liu et al., 2023) toolkit for implementation of FAD, FD, and KL, with distributions computed using PANN-CNN14 (Kong et al., 2020) as the backbone model for extracting features for each audio sample.

Table 2 presents the performance metrics for models trained on Hindustani Classical music. Both finetuned versions of MusicGen and Mustango demonstrate superior performance compared to their baseline counterparts across all evaluated metrics. Notably, Mustango exhibits significant improvement after finetuning, whereas MusicGen shows only marginal gains. This disparity suggests that Mustango possesses a greater capacity for dataset-specific adaptation.

Adapter finetuning for Mustango model better incorporates domain-specific nuances of Hindustani Classical music, resulting in generated outputs that more closely align with the target style.

Table 2 additionally presents results for Turkish Makam generation. The performance trends mirror those observed in Hindustani Classical music. Mustango demonstrates strong improvement with one notable exception: the PSNR metric. For PSNR, the baseline Mustango model performs better.

4.2 Human Evaluation

To complement our objective metrics, we designed a rigorous human evaluation process, recognizing the crucial role of human perception in assessing music quality and authenticity. We begin by generating prompts for drawing audio inferences from the models based on Bloom's taxonomy criteria. Then we present the outputs to human judges to compare them in an arena setup (Chiang et al., 2024; mrfakename et al., 2024).

We divided our process into two phases. In first phase, two annotators independently judged a portion of the same set of data points. This allowed us to compute inter-annotator agreement, a crucial measure of evaluation reliability. Disagreements were systematically discussed and resolved, refining our evaluation criteria. In phase two, annotators transitioned to single annotations per data point continuing evaluation of the rest audios. Finally, we compute ELO ratings of the models based on second phase annotations.

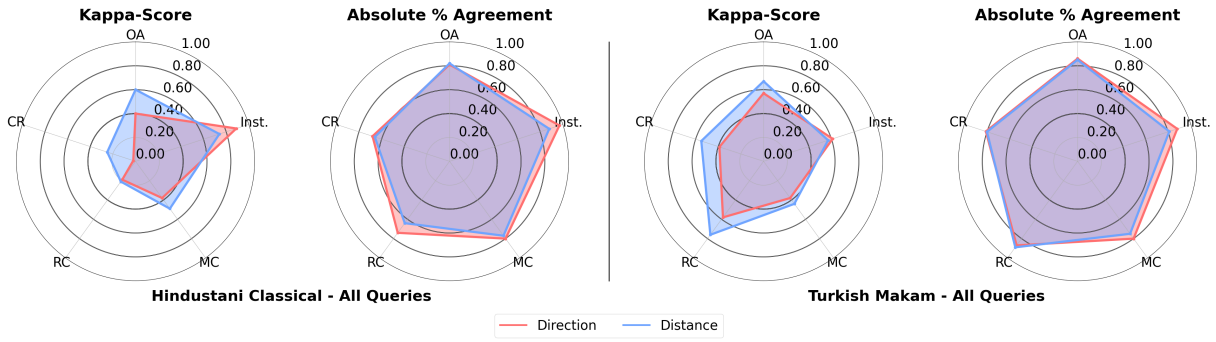

Figure 3: Inter Annotator Agreement metrics for Hindustani Classical & Turkish Makam.

4.2.1 Material

We introduce novel evaluation criteria based on Bloom's Taxonomy to assess a model's understanding of musical elements in text and their alignment with the generated audio using arena-style evaluation.

We evaluate the models under three conditions: **recall**, **analysis**, and **creativity**, by manually generating 10 prompts in each category (Table 1).

- **Recall**: Tests the model's ability to reproduce combinations of melody, instrument, and rhythm from the fine-tuning data, testing effective memorization and recall.
- **Analysis**, We create novel combinations by substituting melodies, rhythms, or instruments, testing the model's adaptability beyond the training data.
- **Creativity**, We combine genres, blending melodies, rhythms, and instruments across styles to test the model's integration of underrepresented and over-represented genres.

For each case, we generate model responses from all models, creating 120 total music samples. Since Mustango generates 10-second inferences at 16kHz, we process MusicGen outputs by clipping them to 10 seconds and downsampling to 16kHz to ensure uniform evaluation conditions.

4.2.2 Method

We decided to go for a comparative evaluation of pieces instead of absolute judgments of pieces in isolation to control the subjectivity so that, the shorter, lower-sampling-rate music clips, are more effectively evaluated through comparison. For each comparison, the user receives a reference prompt and two anonymous audio samples (with the comparisons ordered randomly), followed by five comparative evaluation questions comparing the two audio generations on each criterion: Overall Aesthetic*(OA)*, Instrument Accuracy*(Inst.)*, Rhythm Capture*(RC)*, Melody Capture*(MC)*, and Creativity*(CR)* since we provide these entities in the prompt and we are trying to assess the alignment of the text to the audio generated. For each criteria, we provide the annotator with 7 options: A $\gg$ B, A $>$ B, A = B, A $<$ B, A $\ll$ B, None, and Not Applicable (NA). Please refer to Appendix B for questions.

In first phase, we conduct four types of comparisons: *baseline* vs. *baseline*, *baseline* vs. *fine-tuned* (for both models), and *fine-tuned* vs. *fine-tuned*. We request two avid listeners of each genre, who are aware of the nuances but not themselves professional musicians(demography details in Appendix I), to annotate these samples. The annotation process begins by evaluating 36 comparisons for each genre—9 generations from each model per genre—compared across all models based on the five evaluation criteria. After the completion of first phase we compute the **Inter-Annotator Agreement(IAA)**, using **distance** and **direction-based kappa** scores. The distance-based Kappa quantifies the absolute differences in annotations by both the annotators whereas direction-based Kappa assesses consistency in preference order rather than the extent of preference. Detailed kappa-score calculation methods are provided in the Appendix E.

Figure 3 presents the kappa score and average agreement for each criterion. In evaluating Hindustani Classical and Turkish Makam music, these metrics reveal distinct patterns across assessment criteria. As we can see from the figure, *OA* scores range from 0.40 to 0.67, while *Inst.* shows high agreement, with scores up to 0.89 due to its objectivity (Figure 3). *MC* achieves moderate agreement, and *RC* scores are generally lower, particularly for Hindustani Classical (0.19 and 0.21), likely due

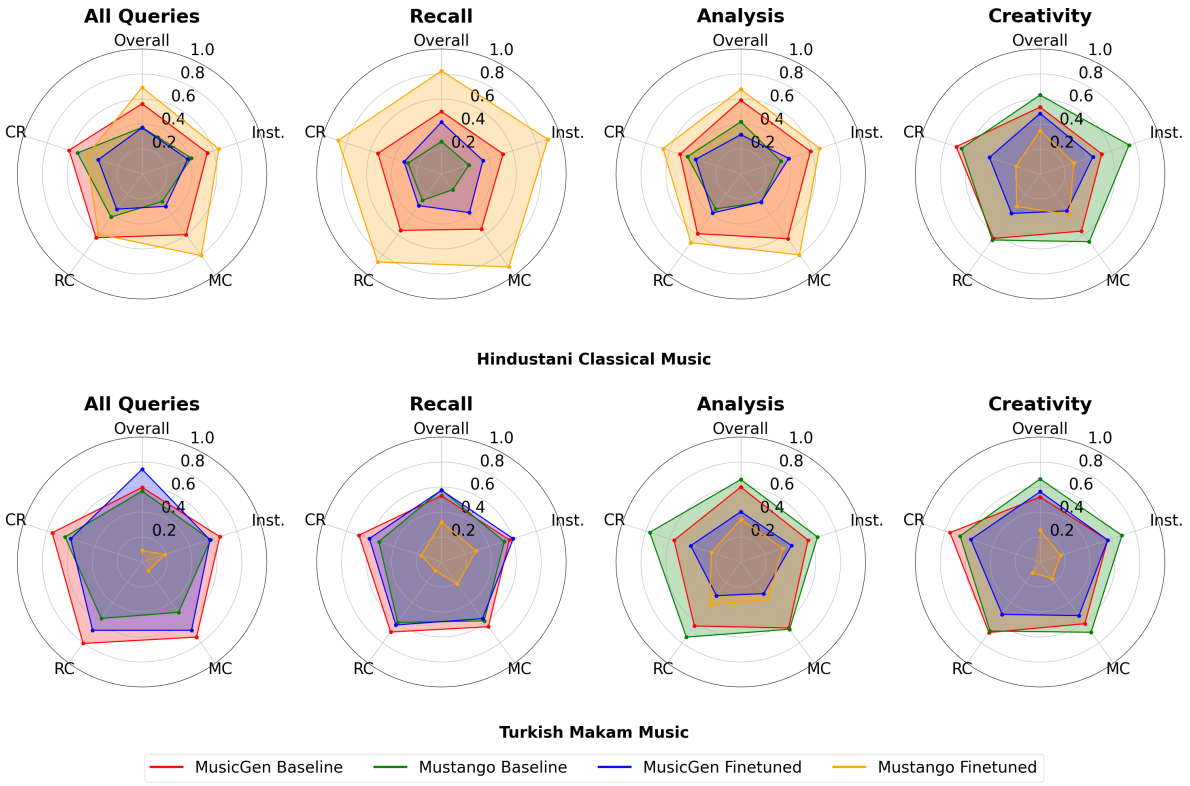

Figure 4: Scaled ELO ratings for each model in Hindustani Classical and Turkish Makam music. Categories include OA: Overall Assessment, Inst.: Instrument Accuracy, MC: Melody Capture, RC: Rhythm Capture, and CR: Creativity. Ratings are provided for all query types and individual query categories.

to the complexity of rhythmic patterns within the short 10-second evaluation span. *CR* consistently records the lowest scores, reflecting the subjective nature of this criterion(For more genre-wise and query-type-wise details refer to AppendixC). After phase-I, we ask the annotators to discuss and re-annotate music samples for Inst. and RC criteria where disagreement is higher.

In the second phase, for Hindustani Classical, we removed the MGB vs MGF comparison since the trend made it clear that MGB is better with agreement from both the annotators. For Turkish Makam, we removed MTF vs MTB since MTB proved to be better(see Figure 4). After filtering we are left with 3 sets of comparisons, with 7 queries for each model, across 3 query types leading to 63 more comparisons for each genre. The annotations from both rounds are combined to compute each model's ELO rating.

4.2.3 ELO Ratings

After comparing the model outputs, we compute ELO ratings (usually used to calculate the relative skill levels of players in a two-player game) for each model across all query types for each evaluation criterion. For each criterion, we consider a single annotation as a single match between the models. If the annotator marks it as NA, then we omit it from the calculation, if A=B or None is marked we consider it as a draw and A$\gg$B, A$>$ is considered a win for A and vice-versa for the remaining cases. The details of computing ELO ratings are given in Appendix G.

The normalized ELO ratings are shown in Figure 4. **For Hindustani Classical music**, over all queries, MTF outperforms all models. Interestingly, MGB is better than MTF, but MGF is judged least favourably, implying that while fine-tuning significantly improves Mustango, MusicGen's performance regresses considerably. These trends hold for all aspects (melody, rhythm, instrument) except creativity. The trends are most prominent for Recall queries, but also hold for Analysis queries, but completely reverses for creativity queries, where MTF performs the worst while MTB performs the best. Qualitative analysis of the generated pieces confirm this finding and shows that there was a strong effect of adaptation on Mustango which led to knowledge attrition and resultant poor performance on creative queries, which required the model to utilize previ-

ous knowledge.

For Turkish Makam music, MTF regresses significantly from MTB, for types of queries as well as on all aspects. While MGF performs slightly better than MGB on all queries for overall rating, the trends are not consistent across different aspects, or different types of queries. In fact, for Analysis queries, even MusicGen's performance significantly regresses for all aspects on finetuning. Thus, we can conclude that the PEFT technique explored here did not help boost the performance of the models for Turkish Makam music.

5 Conclusion

In this paper, for the first time, we systematically explored and established the skewed distribution of musical genres from around the world in datasets used for training Music-Language Models. Non-Western musical traditions are severely underrerpresented which naturally leads to disparate performance across genres in these models. We also demonstrate that PEFT-based techniques vary in effectiveness across different genres and models, further aggravating the challenges of overcoming the data scarcity problem.

As generative models continue to gain traction in the field of music generation and are expected to be used even more widely in the coming years, the misrepresentation and under-representation of the musical genres of the "global majority" poses a significant threat to the inclusion of musical cultures from around the world. The skewed distribution in datasets, reflected in model outputs, can lead to several issues, including cultural homogenization, reinforcement of Western culture dominance (Crawford, 2016), misrepresentation of musical styles, and most importantly, gradual decline leading to the disappearance of many musical genres (Tan, 2021; Lund, 2019; Team, 2023). Therefore, it is critically important to prioritize the creation of inclusive music datasets and models, with an emphasis on under-represented musical genres.

Limitations

Our work relies on adapter-based techniques for cross-cultural adaptation but there is a need to explore additional architectural configurations to further optimize low-resource fine-tuning such as LoRA (Hu et al., 2021) or Compacter (Davison, 2021) approaches.

Additionally, our approach only focused on a few genres, and future work should aim to incorporate a broader range of musical styles. Our investigation involves only Hindustani classical and Turkish Makam traditions, leaving other genres from the Dunya dataset unexplored. This narrow focus stems not from a lack of curiosity, but from our limited cultural expertise - a constraint we acknowledge upfront.

We also trained separate models for Hindustani Classical and Turkish Makam music; combining these into a single model could offer greater generalization across genres.

Another limitation lies in the evaluation process. Human evaluations were conducted on a limited number of samples with a duration of 10 seconds, and more genre-specific assessments are necessary. We also believe that computing objective metrics for underrepresented genres may obscure the full picture because the backbone models used to compute these metrics may not have been trained on various underrepresented genres, resulting in an erroneous portrayal of genres.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

P. Armstrong. 2010. Bloom's taxonomy. https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/. Retrieved October 12, 2024.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science.

Transactions of the Association for Computational Linguistics, 6:587–604.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Steven A. Cholewiak, Panos Ipeirotis, Victor Silva, and Arun Kannawadi. 2021. SCHOLARLY: Simple access to Google Scholar authors and citation using Python.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Sheng-Kuan Chung. 2006. Digital storytelling in integrated arts education. *The International Journal of Arts Education*, 4(1):33–50.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. 2023. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 47704–47720. Curran Associates, Inc.

Kate Crawford. 2016. Artificial intelligence's white guy problem. *The New York Times*. Accessed: 2024-10-16.

Joe Davison. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. In *Neural Information Processing Systems*.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. High fidelity neural audio compression. *Transactions on Machine Learning Research*. Featured Certification, Reproducibility Certification.

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Havnø Frank, Jesse Engel, Quoc V. Le, William Chan, and Wei Han. 2023. Noise2music: Text-conditioned music generation with diffusion models. *ArXiv*, abs/2302.03917.

N.A. Jairazbhoy. 1971. *The Rāgs of North Indian Music: Their Structure and Evolution*. Wesleyan University Press.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Yun-Han Lan, Wen-Yi Hsiao, Hao-Chung Cheng, and Yi-Hsuan Yang. 2024. Musicongen: Rhythm and chord control for transformer-based text-to-music generation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*.

Peike Patrick Li, Boyu Chen, Yao Yao, Yikai Wang, Allen Wang, and Alex Wang. 2024. Jen-1: Text-guided universal music generation with omnidirectional diffusion models. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 762–769. IEEE.

Shuyu Li and Yunsick Sung. 2021. Inco-gan: variable-length music generation method based on inception model-based conditional gan. *Mathematics*, 9(4):387.

Liwei Lin, Gus Xia, Yixiao Zhang, and Junyan Jiang. 2024. Arrange, inpaint, and refine: Steerable long-term music audio generation and editing via content-based controls. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7690–7698. International Joint Conferences on Artificial Intelligence Organization. AI, Arts Creativity.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. AudioLDM: Text-to-audio generation with latent diffusion models. *Proceedings of the International Conference on Machine Learning*, pages 21450–21474.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Holger Lund. 2019. Decolonizing pop music.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8293–8316, Mexico City, Mexico. Association for Computational Linguistics.

mrfakename, Vaibhav Srivastav, Clémentine Fourrier, Lucain Pouget, Yoach Lacombe, main, and Sanchit Gandhi. 2024. Text to speech arena. `https://huggingface.co/spaces/TTS-AGI/TTS-Arena`.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Alastair Porter, Mohamed Sordo, and Xavier Serra. 2013. Dunya: a system for browsing audio music collections exploiting cultural context.

Alec Radford, Ilya Sutskever, et al. 2020. Jukebox: A generative model for music. *OpenAI Blog*, 1(5).

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. 2024. Moûsai: Efficient text-to-music diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8050–8068.

Xavier Serra. 2014. Creating research corpora for the computational study of music: the case of the compmusic project. In *AES 53rd International Conference: Semantic Audio; 2014 Jan 27-29; London, UK. New York: Audio Engineering Society; 2014. Article number 1-1 [9 p.].* Audio Engineering Society.

K.L. Signell. 2008. *Makam: Modal Practice in Turkish Art Music*. Usul Editions.

Ajay Srinivasamurthy, Sankalp Gulati, Rafael Caro Repetto, and Xavier Serra. 2021. Saraga: open datasets for research on indian art music. *Empirical Musicology Review*, 16(1):85–98.

Joseph P. Swain. 1995. The Concept of Musical Syntax. *The Musical Quarterly*, 79(2):281–308.

Or Tal, Alon Ziv, Itai Gat, Felix Kreuk, and Yossi Adi. 2024. Joint audio and symbolic conditioning for temporally controlled text-to-music generation. *Preprint*, arXiv:2406.10970.

Shzr Ee Tan. 2021. Special issue: Decolonising music and music studies. *Ethnomusicology Forum*, 30(1):4–8.

Xiaodong Tan, Mathis Antony, and H Kong. 2020. Automated music generation for visual art through emotion. In *ICCC*, pages 247–250.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Intercontinental Music Awards Team. 2023. Music and ai: The pros, cons, and ethical implications. *Intercontinental Music Awards*.

Yixiao Zhang, Yukara Ikemiya, Woosung Choi, Naoki Murata, Marco A Martínez-Ramírez, Liwei Lin, Gus Xia, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. 2024. Instruct-musicgen: Unlocking text-to-music editing for music language models via instruction tuning. *arXiv preprint arXiv:2405.18386*.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. 2024. Masked audio generation using a single non-autoregressive transformer. In *International Conference on Learning Representations (ICLR)*.

A Research Landscape

Region	Papers (count)	Duration (hrs.)
European	66	**6127.92**
East Asian	**71**	2746.73
South Asian	1	88.78
Central Asian	0	57.01
American	72	921.84
Latin American	5	323.25
Oceania	3	41.99
African	0	27.50
Middle Eastern	5	37.86

Table 3: Distribution of Papers and Duration in hours by Region.

A.1 Genre Distribution Analysis

Genre-wise analysis we can observe in Table 4 and that *Pop* music forms 200K+ hours of the corpus while *Folk* music constitutes only 20K hours of the corpus which is a 10 times between the two as shown in Table 4. *Pop, Rock, Classical* and *Electronic* music genres each constitute more than 10% of the total corpus and more than 100K hours in the total corpus. As shown in Figure 1, *Pop* music has the highest (19.3%) representation followed by *Rock* (17.4%) and *Classical* (13.5%) genres. *Country, hip-hop, blues* and *jazz* have a moderate (more than 5%) representation in the corpus. *Folk* and *experimental* music contribute to only 2.1% of the corpus. The other genres receive minimal attention($\leq$ 1%) which includes music for Children, *Indie-music*, and region-specific genres.

A.2 Regional Distribution Analysis

In region-wise analysis, from analyzing the research space we find that more than 6k hours of music in the research belongs to *European* music and only 28 hours of music belong to *African* music as shown in Table 3. *European*, *East Asian* and *American* music are well explored forming 84.5% of the corpus. On the other hand, *South Asian, Middle Eastern, Central Asian* and *African* music each contribute less than 1% to the whole corpus as depicted in Figure 1.

Genre	Papers (count)	Duration (hrs.)
Pop	**24**	**206.89**
Rock	7	186.67
Electronic	36	140.25
Classical	91	144.64
Country	-	95.77
Hip-hop	3	64.35
Jazz	15	60.62
Blues	-	64.01
Easy Listening	2	74.39
Folk	3	22.802
Experimental	26	11.310
Others	15	0.94

Table 4: Distribution of Hours and Papers by Genre. Duration (Dur.) is represented as 10^3 hours.

B Annotation Details

We asked annotators to choose between two audio samples, based on their preference, to select which better represents the prompted culture in both the inter-annotator agreement scenario and human evaluation. For both inter-annotator agreement and human evaluation, we relied on the same set of questions outlined below.

- Overall, which piece do you like more?
- Which piece captures the instrument (if mentioned the prompt) better?
- Which piece captures the melodic line/scale (if mentioned the prompt) better?
- Which piece captures the rhythm/tempo (if mentioned the prompt) better?
- Which piece is more creative (ignore audio quality while answering this question)?

C Evaluation of Inter Annotator Agreement Results

The inter-annotator agreement (IAA) results, measured using Cohen's Kappa, reveal interesting patterns across genres, metrics, and query types.

In Table 5 Turkish Makam consistently showed higher agreement (0.57-0.67) than Hindustani Classical (0.40-0.60), suggesting potentially clearer structural elements. This trend is particularly pronounced in Rhythm (RC) annotations, where Turkish Makam exhibits substantially higher agreement (0.58-0.76) compared to Hindustani Classical (0.19-0.21).

Instrument identification (Inst.) showed high agreement across both genres (0.57-0.89), with Hindustani Classical scoring particularly well (0.89 for direction). Creativity (CR) exhibited the lowest overall agreement (0.02-0.55), reflecting the inherent subjectivity in assessing creativity.

Examining query types in Table 6 reveals that Recall queries generally yielded higher agreement, particularly in Turkish Makam (0.74-0.75). This indicates strong consistency in factual recall tasks. Analysis queries showed mixed results, with some categories in Hindustani Classical even showing negative agreement, pointing to potential confusion or divergent interpretations in analytical tasks. Interestingly, Creativity queries showed perfect agreement (1.0) in Melody for both genres, suggesting a strong consensus in perceiving creative aspects of melody.

D Evaluation of Human Evaluation Results

The human evaluation results in Table 7 and Table 8, measured using ELO ratings, also reveal intriguing patterns across genres, models, and query

Inter Annotator Agreement (Kappa Score, ↑)					
Hindustani Classical - All Queries					
Metric Type	**OA**	**Inst.**	**MC**	**RC**	**CR**
Direction	0.40	0.89	0.38	0.19	0.02
Distance	0.60	0.74	0.49	0.21	0.25
Turkish Makam - All Queries					
Metric Type	**OA**	**Inst.**	**MC**	**RC**	**CR**
Direction	0.57	0.61	0.38	0.58	0.39
Distance	0.67	0.57	0.44	0.76	0.55

Table 5: Inter Annotator Agreement for Hindustani Classical and Turkish Makam genres. The IAA is calculated using both Direction and Distance-based metrics.

Inter Annotator Agreement (Kappa Score, ↑)					
Hindustani Classical Music					
Recall Queries					
Metric Type	**OA**	**Inst.**	**MC**	**RC**	**CR**
Direction	0.38	1	0.07	0.72	-0.04
Distance	0.48	0.63	0.33	0.39	0.32
Analysis Queries					
Metric Type	**OA**	**Inst.**	**MC**	**RC**	**CR**
Direction	0.43	1	0.48	-0.56	0.15
Distance	0.66	1	0.63	-0.11	0.33
Creativity Queries					
Metric Type	**OA**	**Inst.**	**MC**	**RC**	**CR**
Direction	0.38	0.65	1	0.48	-0.14
Distance	0.63	0.59	0.63	0.38	0.06
Turkish Makam					
Recall Queries					
Metric Type	**OA**	**Inst.**	**MC**	**RC**	**CR**
Direction	0.74	1	0.38	0.65	0.74
Distance	0.75	0.79	0.63	0.84	0.75
Analysis Queries					
Metric Type	**OA**	**Inst.**	**MC**	**RC**	**CR**
Direction	0.48	0.38	-0.04	0.72	0.48
Distance	0.63	0.33	0.18	0.80	0.45
Creativity Queries					
Metric Type	**OA**	**Inst.**	**MC**	**RC**	**CR**
Direction	0.48	0.55	1	0.38	-0.04
Distance	0.63	0.68	0.51	0.63	0.45

Table 6: Inter Annotator Agreement (IAA) Metrics across Recall, Analysis, and Creativity Queries for Hindustani Classical Music and Turkish Makam. Higher Kappa Scores (↑) indicate better agreement.

types. Comparing the two genres, we observe distinct performance profiles for each model. In Hindustani Classical Music, the Mustango finetuned model emerges as the clear leader (OA: 1577), outperforming other models across most categories, particularly excelling in Melodic Contour (MC: 1623). This suggests a strong grasp of the melodic structures specific to Hindustani music. Conversely, for Turkish Makam, the MusicGen finetuned model takes the lead (OA: 1597), with both MusicGen and Mustango baseline models also performing well.

The MusicGen Baseline shows remarkable consistency across both genres, often scoring above 1500 in various categories. This suggests a robust general understanding of musical elements that transcends genre boundaries. The Mustango Baseline, while competitive, generally scores lower than MusicGen Baseline, especially in Hindustani Classical Music.

Finetuning yields mixed results across the two models. For Mustango, it significantly improves performance in Hindustani Classical but drastically

Human Evaluation (ELO Ratings, ↑)					
Hindustani Classical Music - All Queries					
Model	**OA**	**Inst.**	**MC**	**RC**	**CR**
MusicGen Baseline	1525	1520	1540	1552	1546
Mustango Baseline	1449	1466	1409	1470	1518
MusicGen Finetuned	1448	1454	1428	1439	1448
Mustango Finetuned	1577	1559	1623	1538	1487
Turkish Makam - All Queries					
Model	**OA**	**Inst.**	**MC**	**RC**	**CR**
MusicGen Baseline	1539	1562	1597	1622	1603
Mustango Baseline	1527	1531	1499	1523	1560
MusicGen Finetuned	1597	1529	1570	1570	1541
Mustango Finetuned	1337	1377	1334	1286	1297

Table 7: Overall Evaluation Metrics for Hindustani Classical Music and Turkish Makam. ELO ratings (human evaluation) have higher values as better (↑).

Human Evaluation (ELO Ratings, ↑)					
Hindustani Classical Music					
Recall Queries					
Model	**OA**	**Inst.**	**MC**	**RC**	**CR**
MusicGen Baseline	1500	1508	1518	1523	1514
Mustango Baseline	1404	1393	1362	1404	1413
MusicGen Finetuned	1466	1440	1453	1425	1426
Mustango Finetuned	1630	1659	1668	1648	1648
Analysis Queries					
Model	**OA**	**Inst.**	**MC**	**RC**	**CR**
MusicGen Baseline	1536	1535	1556	1536	1505
Mustango Baseline	1467	1435	1412	1438	1480
MusicGen Finetuned	1426	1462	1411	1454	1452
Mustango Finetuned	1571	1566	1620	1572	1562
Creativity Queries					
Model	**OA**	**Inst.**	**MC**	**RC**	**CR**
MusicGen Baseline	1514	1508	1526	1555	1583
Mustango Baseline	1553	1600	1568	1561	1565
MusicGen Finetuned	1494	1478	1446	1456	1471
Mustango Finetuned	1439	1414	1460	1428	1381
Turkish Makam					
Recall Queries					
Model	**OA**	**Inst.**	**MC**	**RC**	**CR**
MusicGen Baseline	1512	1529	1556	1577	1578
Mustango Baseline	1531	1512	1533	1539	1511
MusicGen Finetuned	1530	1542	1524	1549	1543
Mustango Finetuned	1428	1417	1387	1334	1368
Analysis Queries					
Model	**OA**	**Inst.**	**MC**	**RC**	**CR**
MusicGen Baseline	1540	1527	1561	1553	1525
Mustango Baseline	1564	1559	1566	1597	1607
MusicGen Finetuned	1461	1471	1425	1433	1469
Mustango Finetuned	1436	1442	1448	1469	1399
Creativity Queries					
Model	**OA**	**Inst.**	**MC**	**RC**	**CR**
MusicGen Baseline	1508	1528	1544	1579	1604
Mustango Baseline	1566	1576	1578	1573	1570
MusicGen Finetuned	1525	1528	1512	1507	1534
Mustango Finetuned	1402	1369	1366	1342	1291

Table 8: Model Evaluation Metrics across Recall, Analysis, and Creativity Queries for Hindustani Classical Music and Turkish Makam. ELO ratings (human evaluation) have higher values as better (↑).

reduces its effectiveness in Turkish Makam. Conversely, MusicGen's finetuning slightly lowers its performance in Hindustani Classical but enhances it in Turkish Makam. This divergence underscores the complexity of adapting models to specific musical traditions without losing generalizability.

Examining performance across query types reveals further insights. In Recall queries for Hindustani Classical, Mustango finetuned significantly outperforms other models (OA: 1630), particularly in Melody (1668). For Turkish Makam, MusicGen Baseline leads in Recall queries (OA: 1512), with MusicGen finetuned close behind (OA: 1530). This suggests that finetuning can enhance a model's ability to accurately reproduce genre-specific musical elements.

Creativity queries yield particularly interesting results, with baseline models outperforming their finetuned counterparts in both genres. In Hindustani Classical, Mustango Baseline leads (OA: 1553), while in Turkish Makam, it shares the top position with MusicGen Baseline (OA: 1508 and 1566 respectively). This suggests that finetuning, while beneficial for recall and analysis, might constrain the model's creative capabilities.

E Kappa Score Computation

	A$\gg$B	A>B	A=B	A<B	A$\ll$B
A$\gg$B	1	1	0.33	0	0
A>B	1	1	0.67	0.33	0
A=B	0.33	0.67	1	0.67	0.33
A<B	0	0.33	0.67	1	1
A$\ll$B	0	0	0.33	1	1

Table 9: Matrix representation of distance-based agreement score for Inter Annotator Agreement. Column represents Annotator-1's preference and Row represents Annotator-2's preference.

E.1 Distance-based Computation Matrix

The **distance-based Kappa** quantifies the absolute differences in annotations by both the annotators. Each option(except None & NA) is assigned a value between 2 to -2 in order; A $\gg$ B, A > B, A = B, A < B, A $\ll$ B. After assigning values we calculate absolute distances between annotator preferences while excluding all cases which are annotated None or NA by annotators. The distance values are clipped to a maximum of 3, with agreement computed as follows :

$$p_o^i = \frac{3-d}{3}$$

Table 9 represents the annotator preferences and the agreement score for each combination.

E.2 Direction-based Computation Matrix

	A$\gg$B	A>B	A=B	A<B	A$\ll$B
A$\gg$B	1	1	1	0	0
A>B	1	1	1	0	0
A=B	1	1	1	1	1
A<B	0	0	1	1	1
A$\ll$B	0	0	1	1	1

Table 10: Matrix representation of direction-based agreement score for Inter Annotator Agreement. Column represents Annotator-1's preference and Row represents Annotator-2's preference.

The **direction-based Kappa** assesses consistency in preference order rather than the extent of preference. A disagreement is defined as only when the preference orders are reversed between the two annotators (i.e., when one annotator chooses A<B or A$\ll$B and the other annotator chooses B<A or B$\ll$A) Agreement is calculated as follows :

$$p_o^i = 1 - d$$

Table 10 shows the agreement scores between different annotations.

Observed agreement is averaged per criterion, with expected agreement(p_e) and Kappa(κ) calculated as follows:

$$\kappa = \frac{\frac{1}{n}\sum_{i=1}^{n} p_o^i - p_e}{1 - p_e}$$

F Dataset Details

For Hindustani Classical, the dataset includes five instrument types—sarangi, harmonium, tabla, violin, and tanpura—along with voice. It comprises 41 ragas across two laya types: Madhya and Vilambit.

For Turkish Makam, the dataset features 15 makam-specific instruments, including the oud, tanbur, ney, davul, clarinet, kös, kudüm, yaylı tanbur, tef, kanun, zurna, bendir, darbuka, classical kemençe, rebab, and çevgen. It encompasses 93 different makams and 63 distinct usuls.

G ELO Ratings Computation

For phase I, the total number of evaluations are 36 by each annotator and we consider each annotation as a single match. In Phase II, 63 additional annotations are conducted making a total of 135 matches for computing the ELO ratings. For every match the new rating :

$$R_i = R_i + K * (S_i - E_i)$$

R_i: Player's current Elo rating.
K: Weighting factor that determines how much a single game affects the rating.
S_i: Outcome of the game for the player: 1 for a win, 0.5 for a draw, and 0 for a loss.

$$E_i = \frac{1}{(1 + 10^{\frac{(R_j - R_i)}{400}})}, E_j = \frac{1}{(1 + 10^{\frac{(R_i - R_j)}{400}})}$$

R_i: Player- 1 current Elo rating.
R_j: Player- 2 current Elo rating.
E_i: Expected Elo rating of Player-1.

We use a K value of 15 for calculations due to the limited number of matches; a higher K would disproportionately weight each match and skew the ELO ratings.

H Annotation Tool

We deployed LabelStudio, a versatile and user-friendly annotation tool, on HuggingFace Spaces. Figure 5 provides a visual representation of our annotation tool interface, illustrating the layout and features that our human evaluators used to assess the generated music samples.

I Annotator Demographics

The annotators for our music generation task using adapter models include three individuals from India and one from Uzbekistan, all of whom are avid music listeners with diverse cultural backgrounds and a keen interest in music technology.

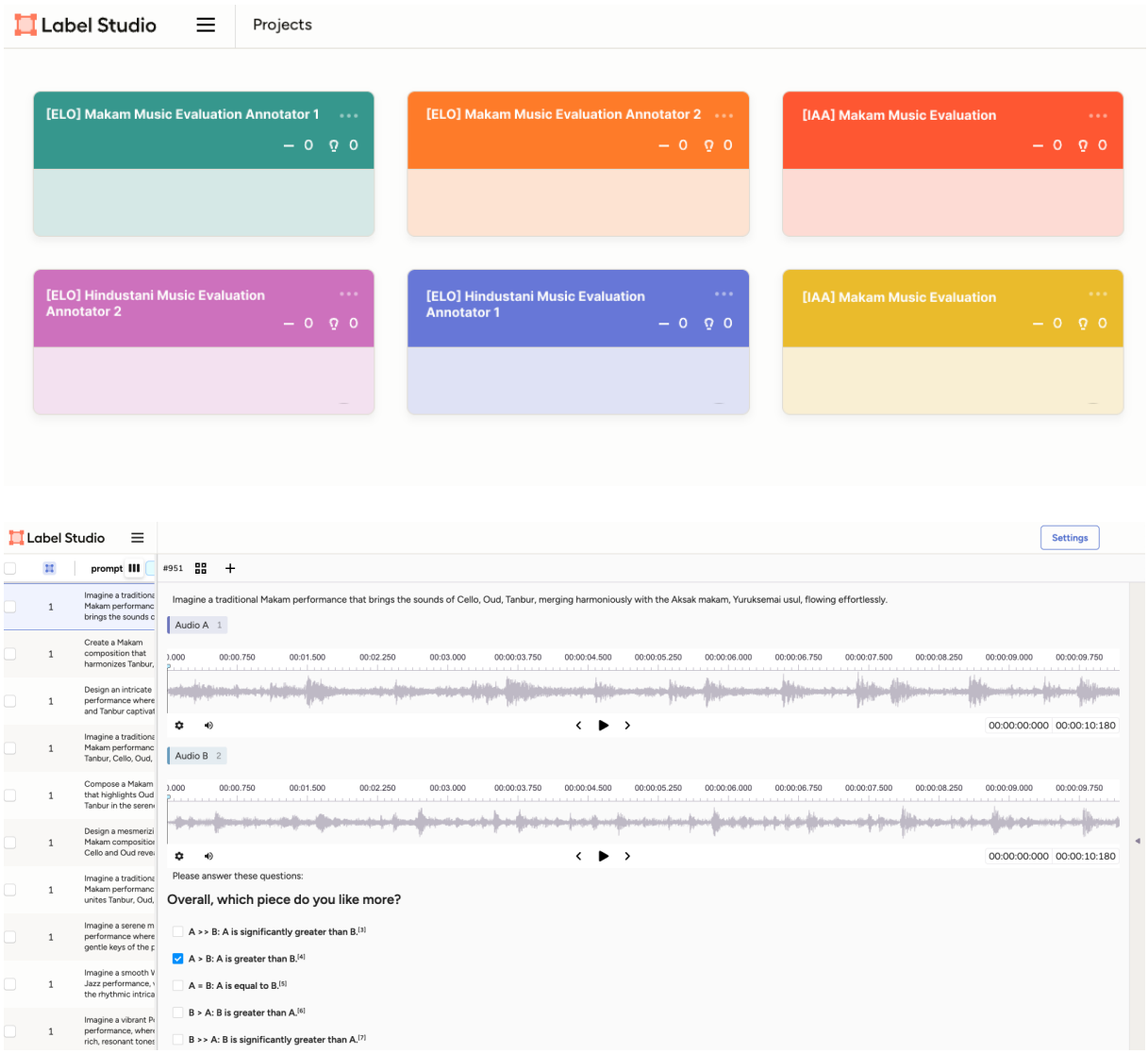

Figure 5: Screenshots of Label Studio, annotation tool for Inter Annotator Agreement and ELO rating comparison