The Sound of Computational Exoticism: Measuring Cultural Misrepresentation in Music Generation

Anonymous ACL submission

Abstract

Generative audio models (GAMs) produce music from text prompts, but their representation of underrepresented musical traditions remains poorly understood. We introduce a framework for measuring *exoticism* in music generation using two embedding-based metrics. *Genre Alignment* (r) captures proximity to a target genre versus Global North mainstream, whilst *Cultural Ambiguity* (m) captures ambiguous positioning between these poles. Expert annotation of generated Samba tracks validates the ambiguity metric against human exoticism judgements. We apply our framework to 8,400 tracks across four GAMs and three genres (Brazilian Samba, Brazilian Marchá, Turkish Makam). The analysis reveals systematic mainstream bias, a genre-label paradox, and two opposing generative mechanisms whereby models either overlay genre markers on mainstream foundations (*superficial layering*) or move decisively towards the genre pole (*authentic reproduction*). Our results show that GAMs often render underrepresented genres exoticised, motivating fairness evaluation beyond acoustic similarity. We publish all code, data and generated tracks publicly at [REDACTED].

1 Introduction

Generative audio models (GAMs) such as MusicGen (Copet et al., 2023), and Suno[1] enable users to generate music from text prompts. The growing societal impact of such systems is illustrated by the band *"The Velvet Sundown"*,[2] whose AI-generated music accumulated millions of streams on Spotify and sparked public debate about authenticity, transparency, and authorship (Bakare, 2025). Beyond these concerns, GAMs also raise questions of cultural representation and appropriation, particularly for underrepresented musical traditions. Existing training corpora exhibit severe representational imbalance, with, for instance, Latin American music accounting for only 2.9% of music in commonly used datasets (Mehta et al., 2025a). As a result, models may learn and reproduce stylistic elements of marginalised cultures without appropriate attribution or benefit-sharing, thereby reinforcing concerns about cultural appropriation (Zhao et al., 2020; Mehta et al., 2024; Barnett, 2023).

One manifestation of such appropriation may be the production of culturally distorted approximations rather than faithful reproductions of musical traditions. The musicological concept of *exoticism* provides a framework for this distinction. Exoticism refers to the deliberate adoption of sonic features, instruments, and structures from traditions perceived as non-Western in order to stage foreignness within a dominant musical framework (Locke, 2001). Rather than reproducing a genre authentically, exoticised music occupies an ambiguous position between authentic cultural expression and mainstream expectations.

Despite growing awareness of dataset bias in music generation (Mehta et al., 2025a; Wang et al., 2024), the extent to which GAMs generate culturally grounded versus exoticised representations of underrepresented musical traditions remains unexplored. Existing work focuses on fine-tuning to improve the generation quality for underrepresented genres, and employs quality metrics that measure acoustic similarity to reference corpora but lack the dimension to detect cultural misrepresentation (Kilgour et al., 2019). In contrast, we ask: ***To what extent do GAMs render underrepresented musical traditions as culturally grounded reproductions versus exoticised approximations?***

To address this question, we propose a novel measurement framework positioning generated outputs along the spectrum between authentic genre expression and Global North mainstream. We validate the framework through expert annotation.

[1] https://suno.com

[2] https://open.spotify.com/intl-de/artist/2GRtyAXWUiisGYub5SGMrb

Four experts evaluate 244 model-generated Samba tracks through pairwise exoticism comparisons, confirming that our metrics reliably predict exoticism judgements ($r_{pb} = 0.30$, $p < 0.001$). We apply our approach to three underrepresented genres (Brazilian Samba, Brazilian Marchá, Turkish Makam) across four text-to-music models (MusicGen Medium, MusicGen Large, Stable Audio 1.0, Suno V4.5). We design three prompt types of increasing specificity: neutral prompts with no genre cues, basic prompts naming the genre, and detailed prompts specifying musicological parameters.

Our analysis on 8,400 generated tracks reveals that GAMs systematically render underrepresented genres exoticised rather than authentic. All models default to Global North mainstream under genre-neutral prompts. We identify two opposing mechanisms. *Superficial layering* adds genre markers to mainstream foundations without displacing them, whilst *authentic reproduction* moves outputs decisively towards the genre pole. Only Suno achieves the latter, and exclusively for culturally distant genres. Critically, explicit cultural specification can amplify rather than mitigate bias, highlighting the need for frameworks assessing cultural misrepresentation beyond acoustic similarity.

Contributions We contribute (**1**) the first operationalisation of musicological exoticism for audio generation with validated metrics for cultural fairness assessment, (**2**) a systematic analysis revealing default mainstream bias and a genre label paradox in less capable models, and (**3**) an identification of two opposing generative mechanisms, superficial layering versus authentic reproduction.

2 Related Work

Music Generation from Natural Language Neural music generation has progressed from symbolic MIDI sequences (Huang et al., 2018) and label-conditioned raw audio (Dhariwal et al., 2020) to free-form text-to-music systems (Agostinelli et al., 2023; Copet et al., 2023). Current architectures fall into three paradigms, namely autoregressive transformers predicting discrete codec tokens (Copet et al., 2023; Agostinelli et al., 2023), diffusion-based systems denoising continuous latent tensors (Evans et al., 2025; Liu et al., 2023; Chen et al., 2024a; Melechovsky et al., 2024), and proprietary systems with undisclosed architectures (Suno, Udio). Our evaluation spans all three through MusicGen, Stable Audio, and Suno, since architectural differences may influence how models encode culturally specific features. Despite this progress, evaluation has remained tethered to acoustic similarity and perceptual preference (Kilgour et al., 2019), leaving cultural dimensions largely unexamined.

Cultural Bias in NLP Cultural bias in NLP denotes systems' tendency to privilege Western, educated, industrialised, rich, and democratic (WEIRD) cultures over non-Western perspectives (Tao et al., 2024; Hershcovich et al., 2022; Adilazuarda et al., 2024). Early work probed pretrained models with cloze-style and discriminative prompts (Nadeem et al., 2021; Nangia et al., 2020; Arora et al., 2023). With the shift to generative LLMs, the focus moved to social-science frameworks such as Hofstede's dimensions and the World Values Survey (Hofstede, 2005; Haerpfer et al., 2022; Zhao et al., 2024; Kiesel et al., 2022) and to perspectivist accounts tracing bias to annotator composition (Lee et al., 2024; Bui et al., 2025b). Across these strands, Western, English-speaking perspectives dominate model behaviour, with concrete effects on model outputs across tasks (Naous et al., 2024; Naous and Xu, 2025), traceable to training-corpus composition and the cultural positioning of development teams.

Bias in Music Generation Research on bias in generative audio has converged on the finding that Global North music dominates training corpora (Mehta et al., 2025a) and propagates into model outputs (Chen et al., 2024b; Zhao et al., 2020; Wang et al., 2024). Systemic factors amplify these biases, including symbolic formats that cannot represent microtonality and coarse metadata that collapse distinct traditions under labels such as "World Music" (Mehta et al., 2025b). Documented harms include uncompensated appropriation, economic marginalisation of Global South musicians, and the risk that distorted outputs circulate as authentic (Barnett, 2023; Mehta et al., 2024, 2025b). Existing mitigation efforts focus on fine-tuning to improve output quality (Mehta et al., 2025a), yet no framework assesses whether outputs achieve authentic reproduction or remain culturally distorted.

3 Exoticism in Music Generation

3.1 The Musicological Concept of Exoticism

Exoticism constitutes a concept rooted in musicology. Locke (2001) defines it as follows:

Definition. *The evocation of a place, people, or social milieu that is (or is perceived or imagined to be) profoundly different from accepted local norms in its attitudes, customs, and morals.*

In musical practice, this evocation manifests through the deliberate adoption of sonic features, instruments, and structures from traditions perceived as non-Western to stage a sense of foreignness within one's own musical framework. Elements such as pentatonic scales or unusual rhythms serve less to represent authentically and more as aesthetic colouration and dramaturgical coding (Locke, 2001). As Locke (2001) notes, these practices "form only a stage in a long and complex history of appropriation and 'borrowings' within popular and other heavily commercial streams of American and European music". Exoticism can thus be positioned on a spectrum between authentic Global South traditions and Global North mainstream. Historical examples illustrate this practice across periods. One prominent case is Bizet's opera *Carmen*, which integrates Spanish musical idioms whilst staging Andalusia as an exoticised locale (Penzlin, 2021). The example captures the central mechanism of musical exoticism, as selected stylistic markers are used to evoke cultural difference without providing authentic representation.

3.2 Limitations of Existing Metrics

Existing evaluation metrics for generated music primarily measure acoustic similarity or distributional alignment. Fréchet Audio Distance (FAD) compares generated audio with a reference corpus in VGGish embedding space and is widely used as a proxy for perceptual quality (Kilgour et al., 2019; Forsgren and Martiros, 2022; Copet et al., 2023; Evans et al., 2025; Batlle-Roca et al., 2024). Kullback-Leibler divergence (KL) complements this by measuring overlap between predicted content-label distributions like instruments or timbral features (Kullback and Leibler, 1951; Batlle-Roca et al., 2024).

These metrics are insufficient for assessing cultural misrepresentation. They evaluate similarity to a target genre in isolation and therefore cannot determine whether an output represents the genre authentically or merely adds stereotyped genre markers to a Global North mainstream foundation. A model may thus obtain favourable FAD or KL scores whilst still producing exoticised approximations of underrepresented musical traditions.

3.3 A Framework for Measuring Exoticism

Building on Locke's conceptualisation of exoticism as situated between culturally grounded genre expression and Global North mainstream conventions (Locke, 2001), we model generated music relative to both poles. Let μ_M and μ_G denote the embedding centroids of a mainstream corpus $\mathcal{C}_M$ and a target-genre corpus $\mathcal{C}_G$, respectively.

$$\mu_* = \frac{1}{|\mathcal{C}_*|} \sum_{x \in \mathcal{C}_*} \phi(x), \quad * \in \{M, G\}, \tag{1}$$

where $\phi(x) \in \mathbb{R}^d$ denotes the embedding of a track x. For each generated track with embedding z, we compute distances to the centroids $d(z, \mu_M)$ and $d(z, \mu_G)$. We then measure *Genre Alignment* (r) as the ratio of these distances:

$$r = \frac{d(z, \mu_M)}{d(z, \mu_G)}, \tag{2}$$

where $r < 1$ indicates proximity to mainstream and $r > 1$ to the target. As a ratio, r scales asymmetrically around the threshold $r = 1$. Equal arithmetic deviations from $r = 1$ do not correspond to equal degrees of ambiguity. We therefore define *Cultural Ambiguity* (m):

$$m = 1 - \frac{|d(z, \mu_M) - d(z, \mu_G)|}{d(z, \mu_M) + d(z, \mu_G)}, \tag{3}$$

where $m \in [0, 1]$. Values close to 1 indicate near-equidistance between centroids, whereas lower values indicate clearer assignment to one pole. Jointly, r and m distinguish mainstream-oriented ($r < 1$, low m), genre-oriented ($r > 1$, low m), and culturally ambiguous outputs ($m \approx 1$), which we treat as candidates for exoticised approximation.

4 Experiments

4.1 Data

We evaluate our framework across three Global South genres, four text-to-music models, and three prompt types of increasing specificity.

4.1.1 Reference Datasets

Data Collection and Corpus Construction The application of the proposed exoticism framework requires authentic corpora comprising a mainstream reference corpus and target genre corpora. We select three target genres from the Global South for investigation: Brazilian Samba, Brazilian Marchá, and Turkish Makam. We choose

Samba as the primary genre because we have access to Brazilian music experts for human validation (§4.2). We include Marchá as a complementary Brazilian genre to enable analysis of regional variation within the same cultural context, whilst Turkish Makam serves as a culturally and geographically distant genre to test the framework's regional independence and generalisability.

Mainstream Corpus We construct the mainstream reference corpus from the German year-end singles charts (2005–2024), sourced from the official German charts website.[3] German charts include both international and German-language titles, providing a more representative cross-section than US-only charts. From 2,000 chart positions corresponding to 1,733 distinct titles, we retrieve 1,613 tracks from YouTube.

Target Genre: Samba We derive the Samba corpus from SAMBASET (Maia et al., 2019),[4] which comprises historical *samba de enredo* recordings, a specific urban subgenre composed for annual carnival parades by Brazilian samba schools (Béhague, 2001b). Of the 486 catalogued titles, we retrieve 233 tracks from YouTube, reflecting the limited availability of marginalised genres.

Target Genre: Marchá The Marchá corpus is sourced from the phonographic archive Discografia Brasileira,[5] which catalogues 78 RPM recordings from 1902–1964. We download 216 public domain titles from the Marchá category, constituting authentic historical recordings.

Target Genre: Turkish Makam We obtain the Turkish Makam corpus from the curated Dunya corpus (Uyar et al., 2014),[6] developed within the CompMusic project, and download 4,424 tracks. For each of the four corpora, the mean embedding across all tracks constitutes the corpus centroid.

Embedding Methodology We represent all audio tracks as 512-dimensional vectors using LAION-CLAP (Wu et al., 2023), embedding only the central 10-second segment of each track. Full methodology, justification of this choice, pairwise cosine distances between genre centroids, and a robustness check with MuQ-MuLan (Zhu et al., 2025) are reported in Appendix A (Table 1).

[3] https://www.offiziellecharts.de
[4] https://www02.smt.ufrj.br/~starel/sambaset/
[5] https://discografiabrasileira.com.br/
[6] https://dunya.compmusic.upf.edu/

4.1.2 Prompt Design and Music Generation

To investigate how the degree of genre specification in prompts affects positioning along the mainstream-genre spectrum, we design three prompt types with increasing levels of detail. See Appendix B for prompt construction details.

Neutral Genre-neutral prompts contain no genre cues but specify general musical characteristics, e.g., *"Create a 30-second instrumental track, upbeat and joyful, suitable for a summer morning."*

Basic Basic genre prompts name a marginalised music genre without further elaboration, such as instrumentation or structural details, e.g., *"Create a Samba track."*

Detailed Detailed genre prompts specify a marginalised music genre alongside additional musical parameters including metre, tempo, harmony, instrumentation, form, and stylistic markers, e.g., *"Create a Samba track built around call-and-response verses and refrains set to a lively percussion ensemble of bombo, caixa, tambourine, cuíca, reco-reco, and guaiá."*

Model Choice and Generation We use GPT-4o (OpenAI, 2024) to generate 300 paraphrases per prompt type for each target genre, plus 300 neutral paraphrases shared across genres. One author manually validates them. Each prompt set is applied to four text-to-music models (MusicGen Medium, MusicGen Large, Stable Audio 1.0, and Suno V4.5; see Appendix C for model specifications). Model choice follows Grötschla et al. (2025), combining a top-performing proprietary system (Suno V4.5) with open-source models that show strong performance in human preference evaluations (MusicGen Large, Medium), and adding Stable Audio 1.0 as a diffusion-based counterpart to the transformer-based MusicGen variants. After generation, we embed all tracks and compute centroids for each prompt type per model, analogous to the reference corpus centroids described above. The complete dataset comprises 8,400 tracks. Since neutral prompts contain no genre cues, each neutral track is evaluated against all three genre centroids, yielding 10,800 track-genre evaluations (7,200 basic and detailed tracks $\times$ 1 evaluation + 1,200 neutral tracks $\times$ 3 evaluations).

4.2 Human Validation

To validate our exoticism metrics, we conduct a two-stage validation for Samba. Four experts par-

ticipate in both stages, comprising two professional Brazilian musicians (compensated at €50 each), a hobby musician, and the first author (musicologist specialising in exoticism). Survey 1 identifies tracks producing recognisable music, whilst Survey 2 assesses exoticism through pairwise comparisons of tracks.

4.2.1 Survey 1: Music Quality Assessment

Sampling Strategy From each model-prompt combination (4 models $\times$ 3 prompt types), we select 30 tracks using stratified sampling across the r-value spectrum, ensuring uniform coverage from mainstream-oriented to genre-authentic outputs rather than clustering around modal values.

Evaluation Design Each of the four experts evaluates 120 tracks (40 shared inter-rater stimuli plus 80 unique per survey, totalling 360 tracks). Annotators rate 20-second centre excerpts on a 5-point Likert scale from (1) *no recognisable musical structure* to (5) *clearly identifiable as music* (see Figure 4 in Appendix D). The number of inter-rater stimuli reflects the cognitive demands of audio annotation. Unlike text, audio requires simultaneous attention to harmony, rhythm, melody, tempo, and instrumentation, often necessitating repeated listening. The annotation load is comparable to Mehta et al. (2025a)'s first evaluation phase, which used 36 comparisons of 10-second excerpts per genre.

Quality Threshold We consider tracks rated $\geq$ 4.0 to have produced recognisable music. For unique tracks, we apply this threshold to the single annotator's rating. For inter-rater tracks, the four-annotator average must reach the threshold. Inter-rater reliability yields Krippendorff's $\alpha = 0.587$, moderate agreement (Landis and Koch, 1977). In total, 244 of 360 tracks (67.8%) pass the threshold; per-condition results are reported in Table 4.

4.2.2 Survey 2: Expert Exoticism Annotation

Pairwise Comparison Design To manage cognitive load, we randomly pair the 244 qualified tracks across two surveys, presenting each pair as a 40-second sequence (20 seconds per track). Annotators select which track sounds more exoticised, or indicate equal exoticism, no exoticism, or uncertainty (see Figure 5 in Appendix D). This approach yields 247 judgements (50 unique pairs evaluated by all experts plus 197 single judgements), with each track appearing in at least two pairings.

Annotation Reliability Inter-annotator agreement for the 50 shared pairs yields Krippendorff's $\alpha = 0.386$ (nominal). Whilst below conventional thresholds for strong reliability, this value must be interpreted in light of the task complexity. Auditory assessment requires simultaneous attention to multiple musical dimensions, and exoticism adds a culturally conditioned layer of judgement. Mehta et al. (2025a) report that inter-rater agreement for music evaluation varies substantially by criterion, ranging from $\alpha = 0.89$ for objective dimensions like instrument accuracy down to $\alpha = 0.02$ for subjective dimensions like creativity, the category within which exoticism falls. Annotator feedback further highlights ambiguous cases such as big band instrumentation in Samba complicating authenticity judgements. We therefore interpret the observed agreement as partial but meaningful convergence under a highly subjective task, illustrating that exoticism is culturally contested and difficult to operationalise consistently.

4.2.3 Metric Validation

Point-Biserial Correlation To assess whether m aligns with human exoticism judgements, we correlate pairwise differences $\Delta m = m_1 - m_2$ with binary judgements ($n = 184$ after excluding ambiguous responses). The point-biserial analysis reveals a significant positive association ($r_{pb} = 0.30$, $p < 0.001$), indicating that tracks with higher relative m-values are more likely to be perceived as exoticised. Given the subjective nature of exoticism, this moderate alignment is expected. Appendix F.1 (Table 5) reports filtering criteria and detailed results.

Logistic Regression A complementary logistic regression confirms Δm as a significant predictor ($\beta = 6.16$, $p < 0.001$). An increase of 0.1 in Δm raises the odds of being judged more exoticised by approximately 85% (OR = 1.85). Full regression results, predicted probabilities, and a supplementary model with annotator fixed effects ($p = 0.86$) are reported in Appendix F.2 (Table 6).

4.2.4 Cross-Genre Generalisation

The human evaluation provides evidence supporting the m-metric for Samba ($r_{pb} = 0.30$, $p < 0.001$). Whilst direct validation was conducted only for Samba, we argue that the observed relationship plausibly generalises to Marchá and Makam based on three converging observations: (1) all genres are embedded within the same representation

space, making consistent metric behaviour across genres plausible; (2) prompt effects replicate coherently across all three genres (Appendix §H); and (3) Samba occupies an intermediate position within the cultural distance spectrum, allowing it to serve as a bridge for both interpolation and extrapolation. We provide a detailed discussion in Appendix G.

5 Analysis of the Results

5.1 General Patterns: Model, Genre and Prompt Performance

We analyse model and genre patterns aggregated across all conditions, computed over condition-level means (300 tracks per genre–model–prompt).

Model Patterns Figure 1 reveals three distinct generative profiles (Table 10). The two MusicGen variants produce sharply peaked r-distributions centred well below the threshold $r = 1.0$ ($\bar{r} = 0.847$ and 0.861, $\sigma_r = 0.074$ and 0.098), with maximum values barely reaching it ($r_{\max} \leq 1.031$). Stable Audio centres directly at the threshold ($\bar{r} = 1.016$), whilst Suno produces a broad distribution extending substantially beyond it ($\bar{r} = 1.026$, $r_{\max} = 1.340$, $\sigma_r = 0.232$).

The critical distinction emerges in the m-value distributions. Despite nearly identical mean r-values (1.016 vs. 1.026), Stable Audio exhibits the highest mean Cultural Ambiguity across all models ($\bar{m} = 0.932$), whilst Suno achieves the lowest ($\bar{m} = 0.874$). This contrast demonstrates that **genre proximity alone does not determine cultural authenticity**. The *manner* of approaching the genre pole matters. Stable Audio reaches the threshold but remains maximally ambiguous, whereas Suno traverses beyond it with decreasing exoticism. Suno's architecture is undisclosed, but the consistent performance gap between open-source and proprietary systems suggests that differences in training data composition, model capacity, or proprietary optimisation techniques contribute to this divergence.

Genre Patterns Table 9 reveals three distinct profiles, with pairwise centroid distances (Table 2) providing the reference frame (see Figure 6 in Appendix K for KDE distributions). Samba, the genre closest to mainstream ($d = 0.169$), never crosses the threshold ($\bar{r} = 0.851$, $r_{\max} = 0.977$). No model reaches authentic reproduction for Samba. Marchá, despite the greatest mainstream distance ($d = 0.379$), achieves only moderate genre proximity ($\bar{r} = 0.938$). Makam ($d = 0.272$) shows a bimodal r-distribution with peaks near $r \approx 0.9$ and $r \approx 1.3$ and the highest maximum overall ($r_{\max} = 1.340$). **Mainstream distance does not monotonically predict generative performance**, as Marchá achieves weaker authenticity than the closer Makam. Corpus size likely contributes to this discrepancy, as the Marchá corpus ($n = 216$) is substantially smaller than Makam ($n = 4{,}424$), potentially producing less stable centroid estimates.

The m-distributions reveal complementary patterns. Marchá exhibits the highest mean Cultural Ambiguity ($\bar{m} = 0.902$, $\sigma_m = 0.027$), indicating consistently elevated exoticism, whilst Makam shows the lowest ($\bar{m} = 0.889$, $\sigma_m = 0.038$). Combined with Makam's bimodal r-distribution, this divergence indicates the presence of two distinct generative mechanisms, which we analyse in §5.2.

Prompt Effects Figure 2 traces the progression from neutral through basic to detailed prompts across all model-genre combinations (Table 7). Three key findings emerge. First, **all models default to Global North mainstream** under neutral prompts: 11 of 12 model-genre combinations yield $r < 1.0$, with an aggregate mean of $\bar{r} = 0.862$ (Table 11). Genre-neutral generation is mainstream-oriented, preventing organic discovery of under-represented genres. Second, **explicit genre labels produce a genre-distance-dependent paradox**. They improve outputs for the mainstream-proximate genre Samba (all four models increase r-values, up to $\Delta r = +0.27$), but can actively harm cultural representation for culturally distant genres. MusicGen Medium shifts Makam outputs *towards* mainstream ($\Delta r = -0.25$), producing outputs further from the target genre than uninstructed generation. The effect is capacity-dependent. Under the same labels, Suno achieves substantial improvements ($\Delta r = +0.38$), suggesting that less capable models default to stereotyped "World Music" associations, whilst more capable models resolve labels to culturally specific characteristics. Third, **detailed musical specifications partially correct the label paradox**, with benefits scaling by genre distance and bounded by model capacity. For Samba, additional detail adds noise. For Makam, MusicGen Medium recovers from $r = 0.70$ to $r = 1.03$ but only reaches the threshold, whilst Suno traverses substantially beyond ($r = 1.34$). Marchá shows weaker patterns, likely reflecting its small corpus size ($n = 216$). Prompt engineer-

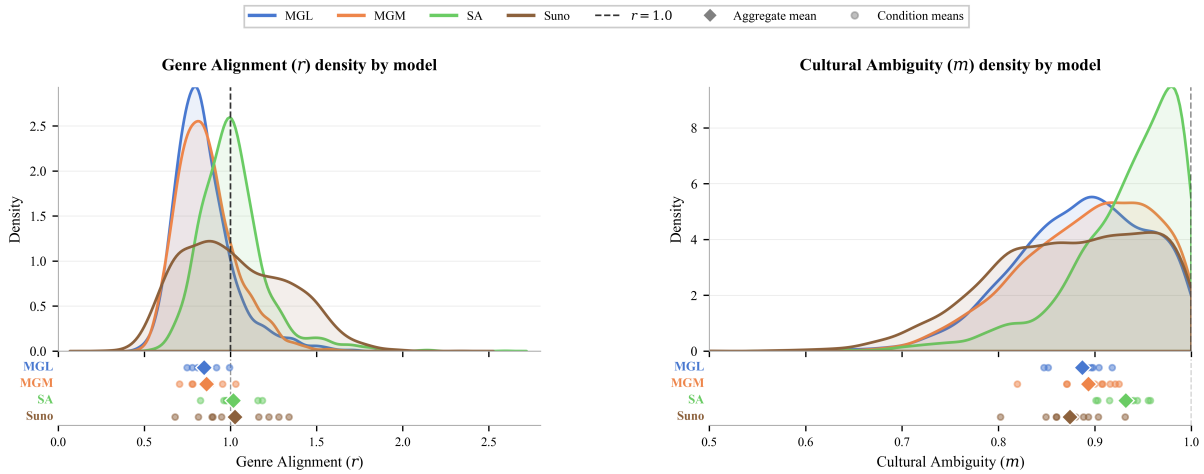

Figure 1: Kernel density estimates of r-values (left) and m-values (right) by model. The strip below each panel shows condition-level means (circles) and aggregates (diamonds). The dashed line ($r = 1.0$) marks equidistance between mainstream and genre centroids. MGL=MusicGen Large, MGM=MusicGen Medium, SA=Stable Audio.

ing thus shifts outputs along the mainstream-genre spectrum but cannot overcome representational deficiencies (see Appendix I for full analysis).

5.2 Exoticism Analysis

The preceding analyses demonstrate that prompt engineering shifts outputs along the mainstream-genre axis, with the achievable range bounded by model capacity. Given the asymmetric scaling of r as defined in §3.3, we examine m-values jointly with r-values to determine whether models traverse decisively beyond the threshold or merely approach it (Table 7, Table 8).

This joint analysis reveals two opposing generative mechanisms. ***Superficial layering*** produces outputs that converge towards the threshold ($r \approx 1.0$) with elevated Cultural Ambiguity ($m \to 1.0$). Genre-specific elements are added to mainstream foundations without displacing them, creating maximally ambiguous outputs. ***Authentic reproduction*** produces outputs that traverse substantially beyond the threshold ($r > 1.0$) with decreasing Cultural Ambiguity, moving decisively towards the genre pole. Figure 3 visualises these trajectories for MusicGen Medium and Suno, which best instantiate the contrast.

Makam illustrates both mechanisms most clearly. MusicGen Medium progresses from $r = 0.70$ to $r = 1.03$ under detailed prompts, yet Cultural Ambiguity rises ($m : 0.82 \to 0.91$). Both genre proximity and ambiguity rise, the hallmark of superficial layering. Suno progresses from $r = 0.90 \to 1.28 \to 1.34$ with Cultural Ambiguity decreasing throughout ($m : 0.89 \to 0.86 \to 0.85$, Figure 3, Makam panel), demonstrating authentic reproduction. The bimodal r-distribution observed in Figure 6 thus reflects these two distinct mechanisms.

Marchá shows a mixed pattern. Suno's r-values cross beyond the threshold ($0.81 \to 1.16 \to 1.22$), suggesting movement towards authentic reproduction, but m-values increase initially before stabilising ($0.86 \to 0.89 \to 0.88$), indicating that this process remains incomplete. MusicGen Medium improves only modestly ($r : 0.78 \to 0.86$) with elevated Cultural Ambiguity ($m : 0.87 \to 0.91$), characteristic of superficial layering.

Samba does not adhere to these two mechanisms. No model-prompt combination achieves $r > 1.0$, so all outputs remain on the mainstream side of the threshold. Detailed prompts reduce both r- and m-values (e.g., Suno's $r : 0.95 \to 0.89$ and $m : 0.93 \to 0.90$), reflecting movement within the mainstream region rather than differentiation between superficial and authentic reproduction.

These findings directly address our research question. **Overall, GAMs systematically favour exoticised approximations over culturally grounded representations of underrepresented musical traditions.** The majority of model-genre combinations position outputs in the ambiguity zone ($r \approx 1.0$, elevated m), exemplified by MusicGen's superficial layering for Makam and Marchá. Among the evaluated models, only Suno achieves authentic reproduction, and only for the most culturally distant genres, suggesting that the available

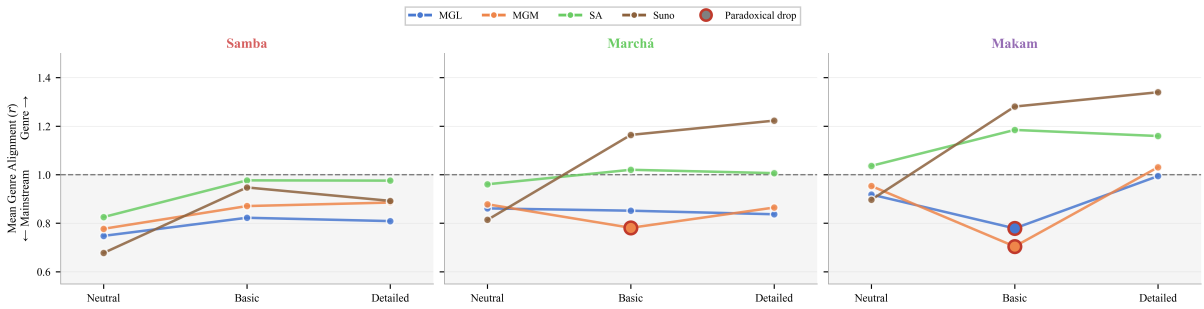

Figure 2: Mean r-values across prompt types (Neutral → Basic → Detailed) per model and genre. Shading marks the mainstream region ($r < 1.0$). Red-circled markers denote paradoxical drops where genre labels decrease genre proximity ($\Delta r < -0.05$). MGL = MusicGen Large, MGM = MusicGen Medium, SA = Stable Audio.

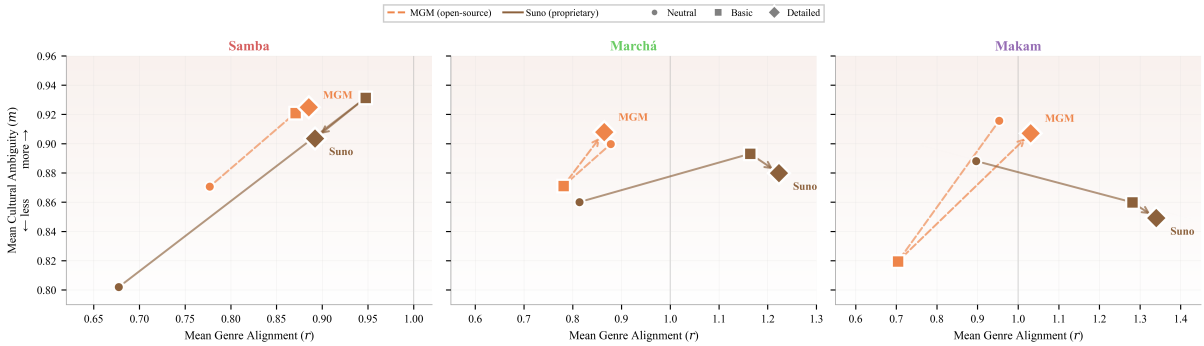

Figure 3: Trajectories of MusicGen Medium (dashed) and Suno (solid) in r-m space across prompt types (circles = Neutral, squares = Basic, diamonds = Detailed). The vertical line at $r = 1.0$ marks the genre proximity threshold.

mainstream-genre distance constrains the achievable range of movement. For the mainstream-proximate genre Samba, neither mechanism produces outputs that cross the authenticity threshold.

6 Discussion and Conclusion

Systemic Bias and Structural Discrimination The default mainstream bias has implications beyond model performance. GAMs increasingly serve as creative tools whose outputs shape which musical traditions gain visibility in AI-generated media. A model that never produces underrepresented genres under neutral prompts effectively excludes these traditions, amplifying the distributional imbalance already present in training corpora (Mehta et al., 2025a). This creates a feedback loop in which underrepresented genres generate less AI-produced content, receive less exposure, and remain underrepresented in future training data. Interventions at the training data level, such as deliberate oversampling or culturally informed curation, are necessary. The consistency of the default across architectures suggests that model design alone cannot mitigate this.

The Genre Label Paradox The finding that explicit genre labels can harm cultural representation in less capable models points to stereotyped internal representations. Genre labels may activate generic features closer to mainstream than culturally specific patterns. Critically, this effect does not scale with cultural distance alone. Marchá is furthest from mainstream yet shows smaller prompt-effect sizes than the closer Makam, indicating a non-monotonic relationship in which cultural distance interacts with training data composition, labelling accuracy, and corpus size. Future work should try to disentangle these factors.

Implications for Prompt Engineering Detailed prompts can worsen cultural representation for less capable models, contradicting the intuitive "add more details" advice. Since the same prompt may improve Suno's outputs whilst harming MusicGen's, optimal strategies depend on genre-distance × model-capacity interactions. Model cards should provide genre-specific prompting guidelines indicating when detailed specifications help versus harm, and acknowledge genres for which authentic reproduction cannot be achieved.

Limitations

Metric Validation Scope We validate Cultural Ambiguity through human evaluation only for Samba. Although the shared CLAP embedding space, coherent statistical patterns (§H), and Samba's intermediate position support cross-genre application, direct validation with Marchá and Makam experts remains absent. Experts from different traditions may judge exoticism through different genre-specific cues, so the relationship between m-values and perceived exoticism may vary. We therefore treat the framework as a preliminary assessment tool rather than a universally validated measure.

Corpus Size and Measurement Stability The Marchá corpus comprises only $n = 216$ tracks, substantially smaller than Makam ($n = 4{,}424$). This manifests in metric inconsistencies specific to Marchá, where statistical effect sizes indicate substantial improvements ($d = 0.74$, $d = 0.67$, §H) whilst r-values show minimal changes, suggesting unstable centroid estimates. This reflects genuine scarcity of digitised recordings rather than a methodological flaw, highlighting that the most marginalised traditions often lack extensive corpora, so measurement precision degrades for precisely those genres most requiring fairness protection.

Genre Selection Constraints The study examines only three genres, namely Brazilian Samba, Brazilian Marchá, and Turkish Makam, limiting generalisability across Global South musical diversity. African, East Asian, and many other marginalised traditions are not represented. The findings should therefore be interpreted as evidence for the examined genres rather than universal conclusions about all musical traditions.

Absence of Absolute Thresholds The framework enables robust pairwise comparison but lacks thresholds for standalone judgements. We lack empirical basis for absolute thresholds such as "$m > 0.85$ constitutes problematic exoticism" or "$r < 0.5$ indicates unacceptable bias". Thresholds are likely genre-specific, as broader stylistic ranges may tolerate higher m-values. The framework thus supports comparative evaluation (track comparison, ranking models, assessing prompt effects) but cannot judge individual outputs without genre-specific benchmarking.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and 1 others. 2023. Musiclm: Generating music from text. *ArXiv preprint*, abs/2301.11325.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP)*, pages 114–130.

Lanre Bakare. 2025. An AI-generated band got 1m plays on Spotify. Now music insiders say listeners should be warned. *The Guardian*. Accessed: 2025-02-02.

Julia Barnett. 2023. The ethical implications of generative audio models: A systematic literature review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 146–161.

Roser Batlle-Roca, Wei-Hsiang Liao, Xavier Serra, Yuki Mitsufuji, and Emilia Gómez. 2024. Towards assessing data replication in music generation with music similarity metrics on raw audio. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*, pages 1004–1011.

Gerard Béhague. 2001a. Brazil. In *Grove Music Online*. Oxford University Press. Published in print: 20 January 2001. Accessed: 2026-02-20.

Gerard Béhague. 2001b. Samba. In *Grove Music Online*. Oxford University Press. Published in print: 20 January 2001. Accessed: 2025-11-15.

Minh Duc Bui, Carolin Holtermann, Valentin Hofmann, Anne Lauscher, and Katharina von der Wense. 2025a. Large language models discriminate against speakers of German dialects. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8212–8240, Suzhou, China. Association for Computational Linguistics.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025b. Multi^{3}Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9714–9731, Albuquerque, New Mexico. Association for Computational Linguistics.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024a. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1206–1210.

Yanxu Chen, Linshu Huang, and Tian Gou. 2024b. Applications and advances of artificial intelligence in music generation: A review. *ArXiv preprint*, abs/2409.03715.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *Advances in neural information processing systems*, 36:47704–47720.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *ArXiv preprint*, abs/2005.00341.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2025. Stable audio open. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Seth Forsgren and Hayk Martiros. 2022. Riffusion – stable diffusion for real-time music generation. Accessed: 2026-02-20.

Florian Grötschla, Ahmet Solak, Luca A Lanzendörfer, and Roger Wattenhofer. 2025. Benchmarking music generation models and metrics via human preference studies. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World values survey: Round seven – country-pooled datafile version 5.0. Dataset.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Geert Hofstede. 2005. Culture's recent consequences. In *Designing for Global Markets 7, IWIPS 2005: Proceedings of the Seventh International Workshop on Internationalisation of Products and Systems*, pages 3–4, Amsterdam, The Netherlands.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music transformer. *ArXiv preprint*, abs/1809.04281.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of Interspeech 2019*, pages 2350–2354.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. 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 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

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. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21450–21474. PMLR.

Ralph P. Locke. 2001. Exoticism. In *Grove Music Online*. Oxford University Press. Published in print: 20 January 2001; this version: 30 March 2020. Accessed: 2025-11-15.

Lucas S Maia, Magdalena Fuentes, Luiz W P Biscainho, Martín Rocamora, and Slim Essid. 2019. SAMBASET: a dataset of historical samba de Enredo recordings for computational music analysis. In *The 20th International Society for Music Information Retrieval Conference*, Delft, Netherlands.

Atharva Mehta, Shivam Chauhan, and Monojit Choudhury. 2024. Missing melodies: Ai music generation and its" nearly" complete omission of the global south. *ArXiv preprint*, abs/2412.04100.

Atharva Mehta, Shivam Chauhan, Amirbek Djanibekov, Atharva Kulkarni, Gus Xia, and Monojit Choudhury. 2025a. Music for all: Representational bias and cross-cultural adaptability of music generation models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4569–4585, Albuquerque, New Mexico. Association for Computational Linguistics.

Atharva Mehta, Shivam Chauhan, Megha Sharma, Gus Xia, Kaustuv Kanti Ganguli, Nishanth Chandran, Zeerak Talat, and Monojit Choudhury. 2025b. Who gets heard? Rethinking fairness in AI for music systems. *ArXiv preprint*, abs/2511.05953.

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.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6423–6443, Albuquerque, New Mexico. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/. Accessed: 2025-10-01.

Dagmar Penzlin. 2021. Dem Toxischen auf der Spur. BR-Klassik, 21 January 2021. Accessed: 2025-03-05.

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.

Baris Uyar, Hasan Sercan Atli, Sertan Şentürk, Barış Bozkurt, and Xavier Serra. 2014. A corpus for computational research of Turkish makam music. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, London, UK. ACM.

Lei Wang, Ziyi Zhao, Hanwei Liu, Junwei Pang, Yi Qin, and Qidi Wu. 2024. A review of intelligent music generation systems. *Neural Computing and Applications*, 36(12):6381–6401.

Owen Wright, Christian Poché, and Amnon Shiloah. 2001. Arab music. In *Grove Music Online*. Oxford University Press. Accessed: 2026-02-20.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

Yizhou Zhao, Liang Qiu, Wensi Ai, Feng Shi, and Song-Chun Zhu. 2020. Vertical-horizontal structured attention for generating music with chords. *CoRR*, abs/2011.09078.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. 2025. Muq: Self-supervised music representation learning with mel residual vector quantization. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3653–3664.

Model	Params	License	Use
LAION-CLAP `laion/clap-htsat-fused`	153.5M	CC0-1.0	audio-text
MuQ-MuLan `OpenMuQ/MuQ-MuLan-large`	∼700M	CC-BY-NC 4.0	music-text

Table 1: Embedding models used for genre-centroid distance computation.

	Mainstream	Makam	Marchá	Samba
Mainstream	0.0000	0.2720	0.3787	0.1690
Makam	0.2720	0.0000	0.1738	0.2497
Marchá	0.3787	0.1738	0.0000	0.3746
Samba	0.1690	0.2497	0.3746	0.0000

Table 2: Pairwise cosine distances between genre centroids in CLAP embedding space.

A Embedding Methodology

We represent all audio tracks as 512-dimensional vectors using LAION-CLAP (Wu et al., 2023) (Table 1), a contrastive language-audio model selected over alternatives such as VGGish (Kilgour et al., 2019) for its higher-dimensional representation. The metric framework operates on distance relations between embeddings and is thus applicable to any embedding model preserving genre-relevant musical structure, which we verify through the human validation reported in §4.2.

Segment Selection Our initial experiments employ feature fusion for tracks exceeding 10 seconds, combining a downsampled global overview with three 10-second samples from the beginning, middle, and end (Wu et al., 2023). However, for Samba, the downsampling produces artefacts that render mainstream and Samba virtually indistinguishable in the embedding space. We therefore embed only the central 10-second segment of each track, yielding substantially improved discriminability.

Empirical Validation Pairwise cosine distances between corpus centroids (Table 2) confirm that all genres occupy distinct regions in the embedding space. Samba lies closest to mainstream ($d = 0.169$), followed by Makam ($d = 0.272$) and Marchá ($d = 0.379$). Makam and Marchá are relatively similar ($d = 0.174$) despite their distance from mainstream, whilst Samba is considerably separated from both ($d = 0.250$ to Makam, $d = 0.375$ to Marchá). This separation establishes the spectrum within which exoticism can manifest.

Robustness Across Embedding Models To assess whether our framework generalises beyond a single embedding, we additionally apply it using MuQ-MuLan (Zhu et al., 2025) (Table 1). The two models agree on 88% of r-value directions, confirming that genre-proximity findings generalise. However, only CLAP's Δm correlates significantly with human exoticism judgements, whilst MuQ-MuLan's Δm shows no predictive power. This suggests that CLAP's text-audio alignment captures culturally relevant dimensions that purely musical embeddings do not.

B Prompt Templates and Construction

We design three prompt types with increasing levels of genre specification. Neutral prompts are shared across all target genres, whilst basic and detailed prompts are genre-specific with sets for each tradition (Samba, Marchá, Makam). All paraphrases are generated with GPT-4o (OpenAI, 2024).

B.1 Neutral Prompts

Neutral prompts avoid genre-specific terminology and instead specify general musical characteristics through mood, texture, and temporal descriptors. Each prompt follows the template:

`<Mood> <Layer> of <Source> <Tempo>, <Dynamics>`

where `<Mood>` describes affective qualities (e.g., "airy", "vibrant"), `<Layer>` specifies textural elements (e.g., "audio bed", "melodic fragment"), `<Source>` indicates instrumental or timbral characteristics (e.g., "deep bass swells", "percussive accents"), `<Tempo>` provides rhythmic context (e.g., "mid-tempo pulse", "uptempo drive"), and `<Dynamics>` describes intensity or evolution (e.g., "subtle rise", "steady energy").

Example: *"Airy audio bed of deep bass swells, mid-tempo pulse, subtle rise"*

We generate 300 neutral prompt variants and apply them uniformly across all three target genres, enabling measurement of default model behaviour in the absence of cultural cues.

B.2 Basic Prompts

Basic prompts explicitly name the target genre without further musical specification. Each prompt follows the template:

`<Verb> <Genre>`

where <Verb> denotes a generative action (e.g., "craft", "create", "compose", "generate") and <Genre> specifies the target musical tradition.

Examples:

- **Samba:** *"Create a Samba track"*
- **Marchá:** *"Compose a Brazilian Marchá"*
- **Makam:** *"Produce a Turkish Makam piece"*

This minimal specification tests whether genre labels alone activate authentic cultural representations or stereotyped associations. We generate 300 distinct basic prompt variants per genre through paraphrasing of verb choices and genre names.

B.3 Detailed Prompts

Detailed prompts combine genre specification with musical parameters based on *Grove Music Online* (Béhague, 2001b,a; Wright et al., 2001), an established reference work in musicology. Each prompt follows the template:

```
<Verb> <Genre> <Musical Parameters>
```

where <Musical Parameters> specifies metre, tempo, rhythmic patterns, instrumentation, harmonic structures, or formal characteristics documented in the source literature.

Examples:

- **Samba:** *"Design a Samba beat featuring the guaiá shimmer and cuíca squeal, complemented by simple harmony evolving into richer progressions"*
- **Marchá:** *"Craft a Brazilian Marchá with an accelerated tempo and syncopated patterns"*
- **Makam:** *"Produce a compound Turkish Makam journey that modulates to a related mode via seyir and returns decisively to the original durak"*

We generate 300 distinct detailed prompt variants per genre, rotating different musical parameters across variants to cover documented genre characteristics. This tests whether musicologically grounded specifications enable culturally authentic outputs or trigger superficial layering of genre markers onto mainstream foundations.

B.4 Dataset Structure

The complete prompt dataset comprises 300 neutral prompts (shared across genres), 900 basic prompts (300 per genre), and 900 detailed prompts (300 per genre). Applied to four text-to-music models (MusicGen Medium, MusicGen Large, Stable Audio 1.0, Suno V4.5), this yields 2,100 tracks per model and **8,400 total tracks**.

C Generation Models

We evaluate four text-to-music systems spanning open-source and proprietary architectures (Table 3). MusicGen variants (Copet et al., 2023) are single-stage autoregressive transformers operating over EnCodec tokens at 32 kHz, trained on approximately 20,000 hours of licensed music from Meta's Sound Collection, Shutterstock, and Pond5. Stable Audio Open 1.0 (Evans et al., 2025) is a latent diffusion model with T5-based text conditioning, trained on audio from Freesound and the Free Music Archive licensed under CC0, CC-BY, and CC Sampling+, at 44.1 kHz. Suno V4.5 is accessed via the official API; its parameter count, architecture, and training data are not publicly disclosed.

Model	Params	License	Access
MusicGen Medium `facebook/musicgen-medium`	1.5B	CC-BY-NC 4.0	open
MusicGen Large `facebook/musicgen-large`	3.3B	CC-BY-NC 4.0	open
Stable Audio 1.0 `stabilityai/stable-audio-open-1.0`	1.1B	SA Community*	open
Suno V4.5 Released 2025-05	n/d	proprietary	API

Table 3: Text-to-music models used for track generation. * Stability AI Community License (non-commercial research use). n/d = not disclosed.

D Survey Design

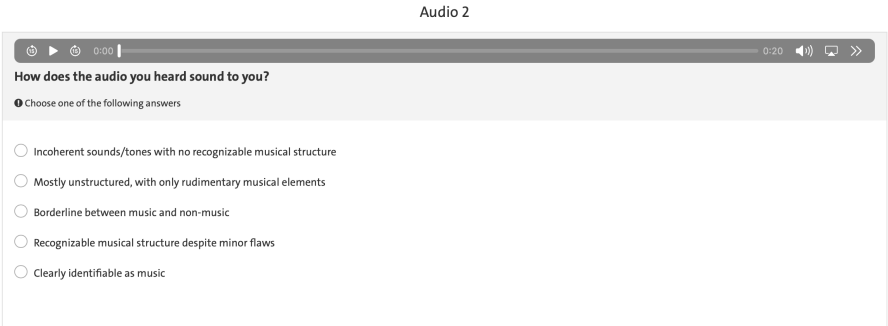

Figure 4: Survey 1: Music quality assessment. Annotators rate 20-second excerpts on a 5-point Likert scale from (1) no recognisable musical structure to (5) clearly identifiable as music.

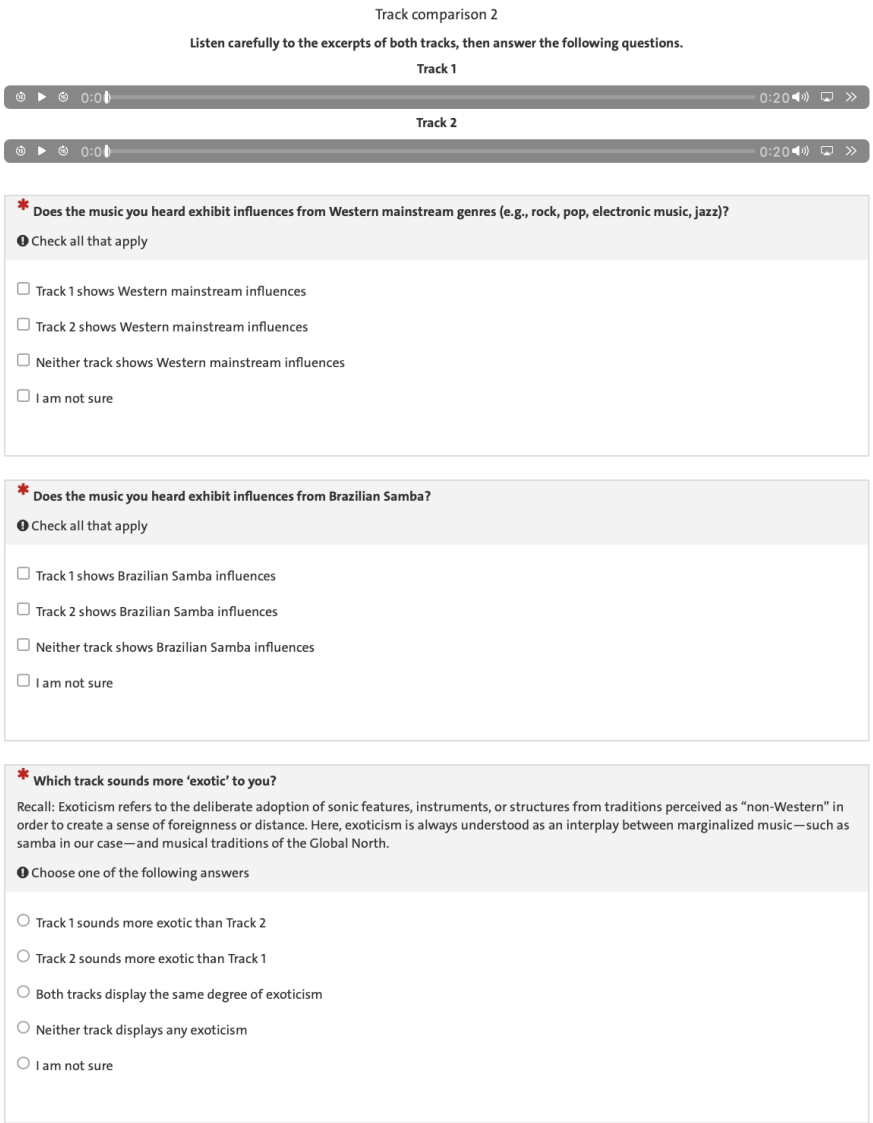

Figure 5: Survey 2: Expert exoticism annotation. Annotators listen to paired 20-second excerpts and select which track sounds more exoticised, or indicate equal/no exoticism.

E Survey 1: Quality Assessment Results

Table 4 reports the per-condition results of Survey 1, which assessed whether generated tracks produce recognisable music. Ratings on a 5-point Likert scale are aggregated as mean ($\bar{x}$) and standard deviation (σ) across $n = 30$ tracks per condition. Tracks rated ≥ 4.0 are classified as recognisable music and proceed to Survey 2 ($n_{\geq 4}$). In total, 244 of 360 tracks (67.8%) pass.

Model–Prompt	$\bar{x}$	σ	$n_{\geq 4}$	%
Suno Neutral	5.00	0.00	30	100.0
Suno Basic	5.00	0.00	30	100.0
Suno Detailed	5.00	0.00	30	100.0
MusicGen Large Neutral	3.76	1.16	18	60.0
MusicGen Large Basic	4.29	0.66	26	86.7
MusicGen Large Detailed	3.83	1.11	20	66.7
MusicGen Medium Neutral	3.53	1.03	15	50.0
MusicGen Medium Basic	3.74	1.24	20	66.7
MusicGen Medium Detailed	3.53	1.07	17	56.7
Stable Audio Neutral	3.12	1.29	10	33.3
Stable Audio Basic	3.84	1.00	18	60.0
Stable Audio Detailed	3.00	1.31	10	33.3
Total	**3.97**	**1.17**	**244**	**67.8**

Table 4: Survey 1 music quality assessment results across all 12 model-prompt combinations ($n = 30$ tracks each). $\bar{x}$ and σ are mean and standard deviation of 5-point Likert ratings. $n_{\geq 4}$ counts tracks meeting the quality threshold (≥ 4.0).

F Metric Validation: Detailed Results

F.1 Point-Biserial Correlation: Detailed Results

We compute point-biserial correlation between $\Delta m = m_1 - m_2$ and binary exoticism judgements, where m_1 and m_2 are the Cultural Ambiguity values of the two tracks in each Survey 2 pair. The outcome variable Y encodes whether Track 1 ($Y = 1$) or Track 2 ($Y = 0$) is judged more exoticised. We exclude ambiguous responses ("both tracks exotic", "neither track exotic", or unclear) and discard ties in the inter-rater aggregation, yielding 184 analysable observations (44 aggregated inter-rater pairs and 140 single judgements) from 247 raw judgements.

Table 5 reports results across three configurations. The inter-rater analysis yields the strongest effect ($r_{pb} = 0.411$, $n = 44$), benefiting from noise reduction through majority aggregation. Single judgements produce a weaker but still significant effect ($r_{pb} = 0.259$, $n = 140$, $p = 0.002$). The combined analysis ($r_{pb} = 0.299$, $p < 0.001$, 95% CI $[0.16, 0.43]$) maximises sample size whilst avoiding double-counting.

	Inter-rater	**Single**	**Combined**
n	44	140	184
r_{pb}	.411	.259	.299
p	.006	.002	<.001
R^2	16.9%	6.7%	9.0%
95% CI	[.13, .63]	[.10, .41]	[.16, .43]
$\overline{\Delta m}$ (Y=1)	+.082	+.033	+.045
$\overline{\Delta m}$ (Y=0)	−.011	−.021	−.019

Table 5: Point-biserial correlation between Δm and binary exoticism judgements. Y=1: Track 1 judged more exoticised; Y=0: Track 2 judged more exoticised.

F.2 Logistic Regression: Detailed Results

To quantify the relationship between Δm and exoticism judgements, we fit a logistic regression model with Y (1 = Track 1 judged more exoticised, 0 = Track 2 judged more exoticised) as the outcome and Δm as the predictor. As with the point-biserial analysis, we include only clear binary exoticism decisions (Track 1 > Track 2 or Track 2 > Track 1) and exclude ambiguous responses ("both", "neither", "not sure"). The model converges successfully ($n = 184$) and demonstrates significant predictive power (likelihood ratio test vs. null model: $p < 0.001$). The coefficient for Δm is $\beta = 6.16$ (SE $= 1.59$, $z = 3.88$, $p < 0.001$), indicating that higher Δm values substantially increase the probability that Track 1 is perceived as more exoticised. For practical interpretation, an increase of 0.1 in Δm raises the odds of Track 1 being selected by approximately 85% (OR = 1.85). Predicted probabilities illustrate a clear monotonic relationship: At $\Delta m = -0.2$, $P(\text{Track 1 more exoticised}) = 19.5\%$; at $\Delta m = 0.0$, $P = 45.4\%$; at $\Delta m = +0.2$, $P = 74.0\%$. Full regression results and predicted probabilities are reported in Table 6.

A supplementary model with annotator fixed effects (fitted only on single judgements, $n = 140$) reveals no significant interindividual differences ($p = 0.86$, likelihood ratio test), suggesting that the Δm effect is consistent across annotators and not driven by individual baseline tendencies.

	Base	+ Annotator
n	184	140
β_0 (intercept)	-0.185	-0.140
β_1 (Δm)	6.155***	5.387**
SE (β_1)	1.588	1.804
OR (per 0.1 Δm)	1.85	1.71
Pseudo-R^2	.068	.054
AIC	241.3	192.9
LR test (annotator)	—	$p = .864$

Δm	P(T1 more exoticised)
-0.20	19.5%
-0.10	31.0%
0.00	45.4%
+0.10	60.6%
+0.20	74.0%

Table 6: Logistic regression predicting exoticism judgements from Δm (top) and predicted probabilities at selected Δm values (bottom). OR = odds ratio. Significance: $^{**}p < .01$, $^{***}p < .001$.

G Cross-Genre Generalisation: Extended Discussion

We motivate the extension of the Samba-validated relationship to Marchá and Makam through three converging arguments.

Shared embedding geometry All genres occupy the same 512-dimensional CLAP embedding space. Distance-based metrics within a shared geometric space apply consistently across embedded entities regardless of cultural origin, and the human validation (§4.2) demonstrates that these distance relations capture perceptually meaningful differences for at least one genre.

Cross-genre statistical replication The WEAT analysis (§H) demonstrates significant and coherent distributional shifts across all three genres. This robust replication provides independent statistical evidence that positioning between mainstream and genre poles operates consistently across cultural contexts, even when effect magnitudes vary.

Strategic validation position Samba's mainstream distance ($d = 0.169$, Table 2) places it between the mainstream pole and the more distant genres Makam ($d = 0.272$) and Marchá ($d = 0.379$). Validating the metric at this intermediate position enables both interpolation towards mainstream-proximate genres and extrapolation towards culturally distant ones, providing empirical coverage across the examined range.

H Statistical Validation of Prompt Effects

We apply the Word Embedding Association Test (WEAT; Caliskan et al., 2017), adapted for audio embeddings, to assess the statistical significance of prompt-type differences. WEAT projects embeddings onto the mainstream–genre axis and tests differential associations through permutation testing (10,000 iterations), returning Cohen's d effect sizes and p-values.

For the majority of genre–model combinations, WEAT confirms statistically significant distributional shifts ($p < 0.001$), with effect sizes ranging from $d = 0.13$ to $d = 1.47$. One systematic discrepancy emerges for Marchá, where WEAT indicates substantial basic-to-detailed improvements for MusicGen ($d = 0.74$ and $d = 0.67$) whilst r-values show minimal corresponding changes. This divergence, absent for Samba and Makam, likely reflects unstable centroid estimates due to the small corpus size ($n = 216$). Condition-specific effect sizes are reported alongside the corresponding r-value analyses in §I.

I Prompt Effects: Detailed Analysis

Default Mainstream Bias When no genre is specified, do models produce culturally diverse outputs or default to Global North mainstream? Under neutral prompts, 11 of 12 model-genre combinations yield $r < 1.0$, indicating systematic mainstream proximity (Table 7). The sole exception is Stable Audio generating Makam ($r = 1.036$), which barely exceeds the threshold. All models begin in the shaded mainstream region of Figure 2. Aggregated across models and genres, neutral prompts produce $\bar{r} = 0.862$ (Table 11), confirming that genre-neutral generation is mainstream-oriented rather than culturally neutral.

This default operates independently of user intent. Prompts such as "30-second instrumental track, upbeat and joyful" contain no genre cues, yet outputs cluster near the mainstream centroid. The consistency across all four architectures suggests that the bias originates in training data composition, where Global North music dominates (Mehta et al., 2025a), rather than in architectural choices.

The strength of the default varies systematically with genre distance. Samba, the genre closest to mainstream ($d = 0.169$), exhibits the strongest mainstream pull, with Suno producing $r = 0.677$ under neutral prompts, the lowest value in the study. Marchá ($d = 0.379$) and Makam ($d = 0.272$)

show weaker but consistent defaults. Mean neutral-prompt r-values across models are $\bar{r} = 0.878$ for Marchá and $\bar{r} = 0.951$ for Makam.

This pattern carries direct fairness implications. Genre-neutral prompts never produce outputs resembling underrepresented traditions, preventing exposure through unprompted use. Users who are not already familiar with an underrepresented genre cannot encounter it through generative models, reinforcing the invisibility of marginalised traditions and perpetuating a cycle of underrepresentation in future training data.

Genre Label Effects Given the default mainstream bias, explicit genre labels should plausibly correct this default by directing generation towards the target genre. Comparing neutral and basic prompts across all model-genre combinations (Table 7) reveals a genre-distance-dependent paradox. Genre labels improve outputs for the mainstream-proximate genre but can actively harm cultural representation for culturally distant genres.

For Samba ($d = 0.169$), the expectation holds. All four models increase r-values, with Suno showing the largest shift ($\Delta r = +0.27$, from $r = 0.68$ to $r = 0.95$). This consistent improvement suggests that models possess adequate internal representations of mainstream-proximate genres that genre labels successfully activate.

For culturally distant genres, the open-source transformer models exhibit a paradoxical inverse effect. MusicGen Medium shifts Makam outputs *towards* mainstream ($\Delta r = -0.25$, from $r = 0.95$ to $r = 0.70$), producing the lowest r-value for any basic prompt condition (red-circled markers in Figure 2). MusicGen Large exhibits a comparable pattern ($\Delta r = -0.14$, from $r = 0.92$ to $r = 0.78$). For Marchá, MusicGen Medium similarly decreases genre proximity ($\Delta r = -0.10$), whilst MusicGen Large remains largely unchanged ($\Delta r = -0.01$). Genre labels here actively increase distance to the target genre.

The paradox is model-dependent. Under the same labels that harm MusicGen's outputs, Suno achieves substantial improvements for Makam ($\Delta r = +0.38$) and Marchá ($\Delta r = +0.35$), crossing the $r = 1.0$ threshold in both cases. Stable Audio shows consistent but smaller gains ($\Delta r = +0.15$ for Makam, $\Delta r = +0.06$ for Marchá).

The genre specificity of this paradox, absent for Samba, pronounced for Makam, and moderate for Marchá, extends a finding from Bui et al. (2025a), who demonstrate that explicit demographic labels amplify bias in LLMs more than implicit cues. Whilst their work shows explicit labelling increasing existing bias in a consistent direction, our results reveal a stronger effect. For culturally distant genres in less capable models, explicit genre labels can reduce rather than increase genre proximity. We hypothesise that genre labels activate stereotyped internal representations, such as generic "oriental" or "World Music" features situated closer to mainstream than to the authentic genre, rather than culturally specific musical patterns. This mechanism is capacity-dependent. Models with greater capacity or more diverse training data (Suno) may resolve genre labels to culturally specific characteristics, whilst less capable models (MusicGen) default to generic associations.

Detail Specification Effects Can detailed musical specifications such as metre, tempo, instrumentation, and harmonic structure compensate where genre labels alone fail? Comparing basic and detailed prompts across all conditions (Table 7) reveals a distance-scaling mechanism. The benefit of detailed specifications increases with genre distance from mainstream, yet remains bounded by model capacity.

For Samba ($d = 0.169$), detailed prompts provide negligible or negative benefit. Three of four models decrease in r-value, with Suno exhibiting the largest regression ($\Delta r = -0.06$, from $r = 0.95$ to $r = 0.89$). For mainstream-proximate genres, genre labels alone activate adequate internal representations, and additional specifications introduce noise rather than signal.

For Makam ($d = 0.272$), detailed prompts produce a striking correction of the label paradox. MusicGen Medium recovers to $r = 1.03$ ($\Delta r = +0.33$ from basic prompts; Figure 2, Makam panel), and MusicGen Large shows a comparable recovery ($\Delta r = +0.22$, from $r = 0.78$ to $r = 0.99$). The detailed musical parameters thus bypass the stereotyped representations that genre labels activated. Crucially, however, this correction brings MusicGen outputs only to the threshold ($r \approx 1.0$) without crossing substantially beyond it, compensating for representational deficiencies without enabling authentic reproduction. Suno, which already crossed the threshold under basic prompts ($r = 1.28$), gains modestly ($\Delta r = +0.06$, $r = 1.34$). Stable Audio remains essentially unchanged ($\Delta r = -0.03$).

For Marchá ($d = 0.379$), the pattern is less

consistent. MusicGen Medium improves ($\Delta r = +0.08$), and Suno achieves the second-highest r-value in the study ($r = 1.22$, $\Delta r = +0.06$). However, MusicGen Large ($\Delta r = -0.01$) and Stable Audio ($\Delta r = -0.01$) show slight decreases. The weaker consistency likely reflects the small corpus size ($n = 216$).

J Supplementary Tables: Complete Metric Results

This appendix provides comprehensive tabulations of r-values (Genre Alignment) and m-values (Cultural Ambiguity) across all experimental conditions, alongside aggregated statistics by genre, model, and prompt type.

J.1 Mean r-Values by Genre, Model, and Prompt Type

Table 7 presents mean Genre Alignment (r) values across all conditions. Values $r < 1$ indicate proximity to mainstream, whilst $r > 1$ indicates proximity to the target genre. $r = 1$ denotes equidistance between mainstream and genre poles.

Genre	Model	Neutral	Basic	Detailed
Makam	MusicGen Large	0.9182	0.7786	0.9942
	MusicGen Medium	0.9533	0.7041	1.0308
	Stable Audio	1.0364	1.1846	1.1594
	Suno	0.8967	1.2809	1.3399
Marchá	MusicGen Large	0.8611	0.8515	0.8369
	MusicGen Medium	0.8780	0.7810	0.8646
	Stable Audio	0.9605	1.0203	1.0064
	Suno	0.8139	1.1641	1.2229
Samba	MusicGen Large	0.7478	0.8224	0.8087
	MusicGen Medium	0.7770	0.8706	0.8852
	Stable Audio	0.8253	0.9767	0.9755
	Suno	0.6774	0.9475	0.8921

Table 7: Mean r-values (Genre Alignment) by genre, model, and prompt type. Higher values indicate greater proximity to target genre relative to mainstream.

J.2 Mean m-Values by Genre, Model, and Prompt Type

Table 8 presents mean Cultural Ambiguity (m) values across all conditions. Values approaching $m = 1$ indicate high exoticism (equidistant placement between mainstream and genre), whilst values approaching $m = 0$ indicate low exoticism (clear assignment to either pole).

Genre	Model	Neutral	Basic	Detailed
Makam	MusicGen Large	0.9179	0.8470	0.8846
	MusicGen Medium	0.9157	0.8195	0.9071
	Stable Audio	0.9553	0.9028	0.9151
	Suno	0.8881	0.8599	0.8491
Marchá	MusicGen Large	0.8961	0.9041	0.8923
	MusicGen Medium	0.8997	0.8712	0.9080
	Stable Audio	0.9440	0.9399	0.9374
	Suno	0.8601	0.8931	0.8799
Samba	MusicGen Large	0.8515	0.8978	0.8906
	MusicGen Medium	0.8706	0.9210	0.9249
	Stable Audio	0.9010	0.9575	0.9339
	Suno	0.8020	0.9314	0.9036

Table 8: Mean m-values (Cultural Ambiguity) by genre, model, and prompt type. Higher values indicate greater exoticism (ambiguous positioning between mainstream and genre).

J.3 Aggregated Statistics by Genre

Tables 9–11 present aggregated statistics computed over condition-level means (i.e., mean, standard deviation, minimum, and maximum of the per-condition r- and m-values reported in Tables 7 and 8) by genre, model, and prompt type.

	$\bar{r}$	σ_r	$r_{\min}$	$r_{\max}$	$\bar{m}$	σ_m	$m_{\min}$	$m_{\max}$
Makam	1.023	.192	.704	1.340	.889	.038	.820	.955
Marchá	0.938	.140	.781	1.223	.902	.027	.860	.944
Samba	0.851	.092	.677	0.977	.899	.042	.802	.958

Table 9: Aggregated statistics by genre: mean, standard deviation, minimum, and maximum computed over condition-level means (4 models $\times$ 3 prompt types $=$ 12 conditions per genre).

J.4 Aggregated Statistics by Model

Table 10 presents aggregated r-value and m-value statistics across all genres and prompt types for each model. These statistics enable comparison of overall model-specific behaviour independent of genre or prompt type effects.

	$\bar{r}$	σ_r	$r_{\min}$	$r_{\max}$	$\bar{m}$	σ_m	$m_{\min}$	$m_{\max}$
MG-Large	0.847	.074	.748	0.994	.887	.023	.847	.918
MG-Medium	0.861	.098	.704	1.031	.893	.034	.820	.925
Stable	1.016	.107	.825	1.185	.932	.021	.901	.958
Suno	1.026	.232	.677	1.340	.874	.037	.802	.931

Table 10: Aggregated statistics by model: mean, standard deviation, minimum, and maximum computed over condition-level means (3 genres $\times$ 3 prompt types $=$ 9 conditions per model). MG = MusicGen, Stable = Stable Audio.

J.5 Aggregated Statistics by Prompt Type

Table 11 presents aggregated r-value and m-value statistics across all genres and models for each prompt type. These statistics enable comparison of overall prompt-type effects independent of genre or model.

	$\bar{r}$	σ_r	$r_{\min}$	$r_{\max}$	$\bar{m}$	σ_m	$m_{\min}$	$m_{\max}$
Neutral	0.862	.101	.677	1.036	.892	.042	.802	.955
Basic	0.949	.183	.704	1.281	.895	.040	.820	.958
Detailed	1.001	.165	.809	1.340	.902	.025	.849	.937

Table 11: Aggregated statistics by prompt type: mean, standard deviation, minimum, and maximum computed over condition-level means (3 genres $\times$ 4 models $= 12$ conditions per type).

K Additional Visualisations

Figure 6 complements the model-level KDE analysis (Figure 1) by aggregating distributions across genres rather than models.

Figure 7 presents violin plots complementing the kernel density estimates in the main text (Figures 1 and 6) by showing the full distributional shape of r- and m-values per model within each genre.

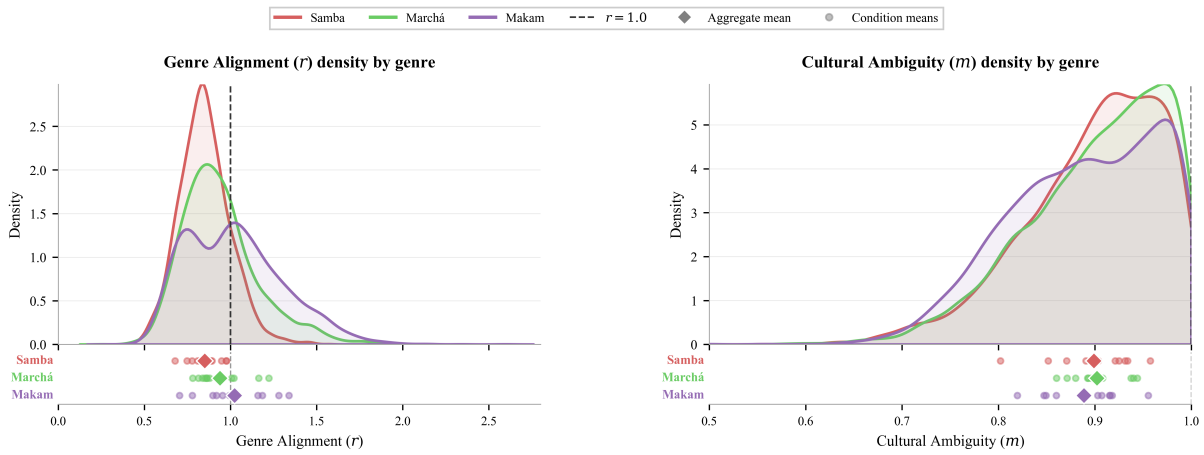

Figure 6: Kernel density estimates of r-values (left) and m-values (right) by genre. The strip below each panel shows condition-level means (circles) and aggregates (diamonds). The dashed line at $r = 1.0$ marks equidistance between mainstream and genre centroids. Pairwise centroid distances d are reported in Table 2.

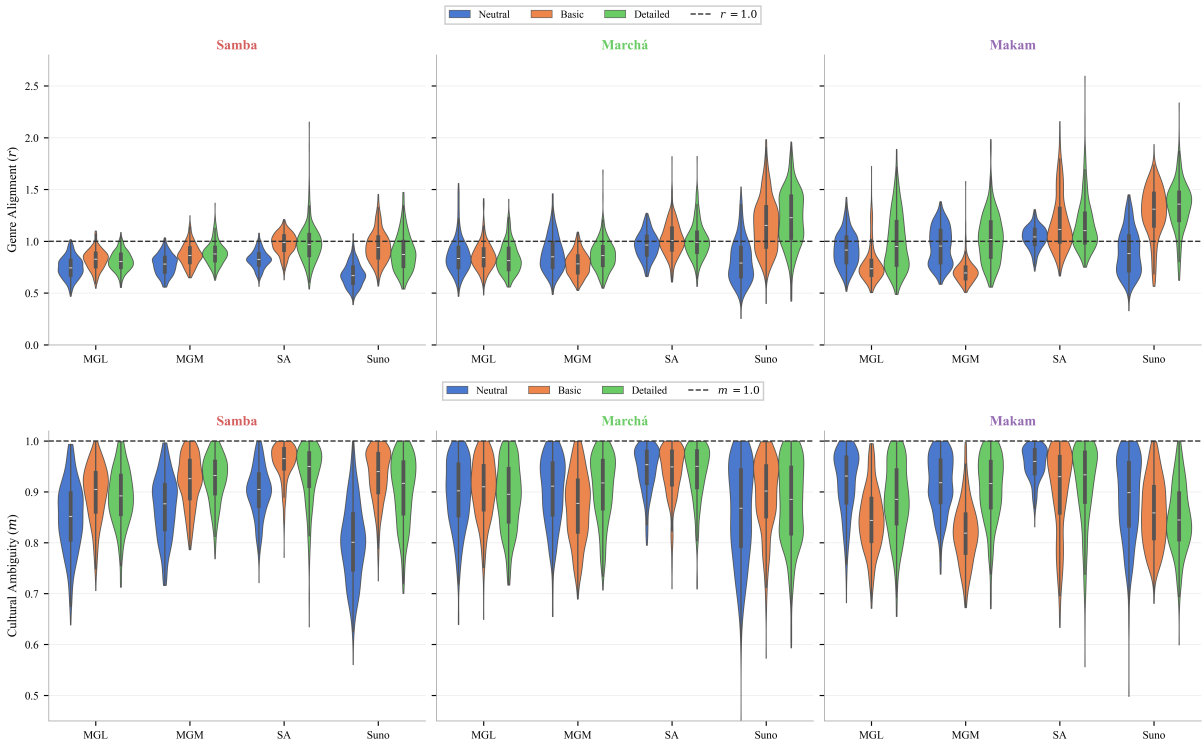

Figure 7: Violin plots of genre-wise distributions for the ratio metric r (top) and the Cultural Ambiguity metric m (bottom). The plots illustrate systematic differences across musical genres in the relationship between model- and genre-centric distances.