Exploring LLMs for South Asian Music Understanding and Generation

Anonymous ACL submission

Abstract

Recent advancements in Large Language Models (LLMs) have shown promising results in music understanding and generation tasks. However, existing works remain confined to Western tonal traditions, offering little insight into whether current LLMs can handle structurally distinct low-resource musical traditions. We present the first systematic evaluation of LLM competence in South Asian classical music, a tradition governed by raga, tala-based melodic constraints that impose fundamentally different structural principles from Western harmony-driven music. We ground our evaluation in Hindustani classical theory and Bengali classical forms, including *Rabindra* and *Nazrul* Sangeet – representative low-resource traditions within South Asian classical music. For music understanding evaluation, we introduce a **504-question-answer** benchmark spanning raga grammar, cultural knowledge, and symbolic notation reasoning, evaluating **33 LLMs** where frontier models such as **Gemini 2.5 Pro** achieve **85-90%** accuracy, while most open-source models remain in the **23-40%** range. For music generation, we design a five-level controlled prompting framework and find that even the strongest model produces stylistically faithful outputs only **40%** of the time. These results reveal that structural validity and stylistic faithfulness in music generation are distinct objectives and highlight an open challenge for culturally grounded music modeling.

1 Introduction

Large Language Models (LLMs) have demonstrated strong generative capabilities across a wide range of traditional NLP tasks (Li et al., 2025). Beyond NLP, they also perform well in creative and structured domains such as story generation (Zhao et al., 2023), poetry composition (Zhang and Eger, 2024), code synthesis (Chen et al., 2021), and mathematical reasoning (Lewkowycz et al., 2022), showing cross-domain generalization (Wei et al., 2022).

These capabilities have naturally extended to music. Because LLMs operate over discrete symbolic representations (i.e., tokens), symbolic music notation provides a natural interface for probing their musical capabilities. For example, ABC notation (Walshaw, 2021) encodes hierarchical pitch and rhythmic structure in discrete textual form, enabling language models to process musical structure in a manner analogous to natural language. This makes symbolic notation a natural testbed for evaluating cross-domain generalization. Consequently, LLMs have been applied to both music understanding (Yuan et al., 2024) and symbolic music generation tasks (Ding et al., 2025; Deng et al., 2024; Xing et al., 2025).

Despite such progress, most existing work focuses on Western musical traditions. As such, whether LLMs can understand and generate culturally specific, structurally distinct, and low-resource music (e.g., South Asian music) remains to be systematically evaluated. Indeed, existing music understanding evaluation benchmarks (Yuan et al., 2024; Zhao et al., 2025a) focus primarily on Western tonal traditions. Similarly, LLM-based music generation frameworks are designed and evaluated mostly on Western-specific properties such as harmonic chord progressions and tonal harmony (Deng et al., 2024; Xing et al., 2025). As a result, it remains unclear whether general-purpose LLMs can model musical traditions whose structure is defined by different musical principles, such as modal frameworks, cyclic rhythmic systems, and ornamentation-driven melodic movement.

South Asian classical music is among the most well-documented non-Western traditions, making it an ideal testbed for this study. Its theoretical framework is organized around raga-based constraints (Pudaruth, 2016), cyclic tala structures (Rowell, 1986), and ornamentation-driven melodic transitions (Basu et al., 2023), which impose long-range dependencies distinct from West-

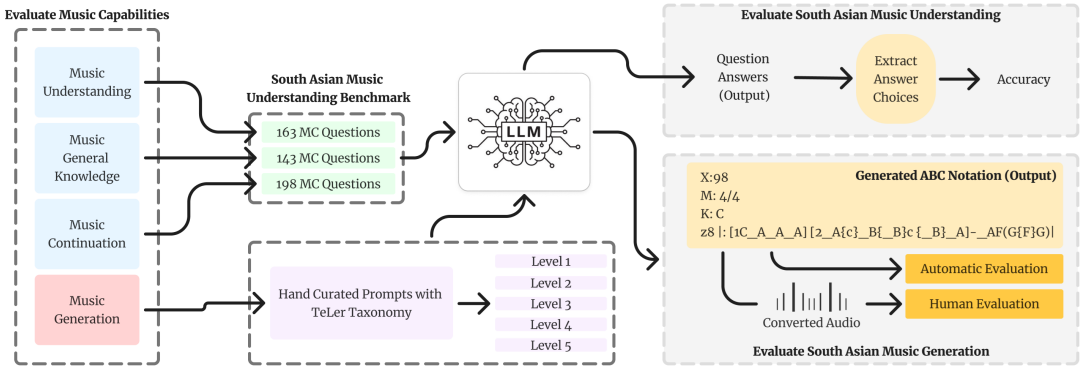

Figure 1: LLMs' South Asian Musical Capabilities Evaluation Pipeline

ern harmony-driven composition (Agarwal et al., 2013). Within this tradition, Bengali classical music forms such as *Rabindra Sangeet* (Tagore, 2014) and *Nazrul Sangeet* (Das, 1996) combine distinct compositional structures with raga- and tala-derived melodic rules, and lyrically driven, stylistically constrained ornamentation.

In this work, we introduce a two-part, comprehensive evaluation framework to assess LLM competence in South Asian classical music systematically (Figure 1). For music understanding capabilities, we construct a music understanding benchmark grounded in Bengali and Hindustani classical traditions and conduct a large-scale evaluation across **33** open-source and frontier LLMs. To evaluate the model's music generation capability, we design a **5-level** controlled prompting framework that incorporates specific Bengali classical stylistic constraints, focusing on *Rabindra* and *Nazrul* Sangeet, and evaluates model outputs in ABC notation using both automatic structural metrics and human judgment. Our contributions are:

- We introduce a **504-question benchmark** spanning three subtasks: 1) Music Theory Understanding (Hindustani classical theory as practiced in Bengali traditions including *Raga Grammar*, *Tala Systems*, *Thaat classification*, and *Ornamentation*), 2) Music General Knowledge (composers and historical figures, instrument knowledge, and regional stylistic traditions), and 3) Music Continuation (compositional reasoning in ABC notation) – serving as the first foundational diagnostic of LLM competence in South Asian classical music in a symbolic setting.
- We curate 100 manually reference scores (50 Rabindra Sangeet and 50 Nazrul Sangeet) in ABC notation derived from official Swaralipi sources, serving as both continuation seeds for the benchmark and references for generation evaluation. We will publicly release the benchmark and curated reference scores.
- We propose a 5-level prompting framework that progressively introduces musical constraints (e.g., scale, rhythm, genre, and stylistic cues) to analyze controllable symbolic music generation.
- We evaluate 33 LLMs on the benchmark and the top 9 models on generation, revealing a large capability gap (e.g., Gemini 2.5 Pro up to 90.8% vs. most open-source models at 23-40% on the benchmark). Despite achieving reasonable structural validity, even the strongest model produces stylistically faithful outputs only 40% of the time, with surface characteristics failing to capture style-specific constraints. We further show that existing automatic metrics fail to capture culturally grounded stylistic properties, highlighting the need for culturally aware evaluation of symbolic music generation.

2 Background & Related Work

2.1 South Asian Classical Music and Bengali Classical Traditions

South Asian classical music follows principles that are fundamentally different from those of Western tonal systems (Agarwal et al., 2013). Rather than harmonic chord progressions, it is structured around raga, a melodic framework that specifies permitted pitches, expressive mood (Pudaruth, 2016), and tala. This cyclic rhythmic system organizes time through repeating beat cycles (Rowell, 1986). Melodic motion is further shaped by gamakas (raga-specific ornamentations) (Basu et al., 2023). These principles appear across multiple Bengali musical traditions including khayal, thumri, baul, and kirtan. Among them, Rabindra Sangeet and Nazrul Sangeet are two of the most influential and well-documented forms. Rabindra Sangeet comprises 2,232 songs by Rabindranath Tagore (Tagore, 1960) and is charac-

terized by close coupling between lyrical prosody and melodic contour, drawing from Hindustani classical, Bengali folk, and Western influences (Gupta, 2016). Nazrul Sangeet includes nearly 4,000 compositions by Kazi Nazrul Islam (Das, 1996) and adheres closely to classical performance conventions (Langley, 2009). Their well-defined compositional rules and reliance on raga-tala structure make them a challenging testbed for evaluating culturally grounded music reasoning in LLMs.

2.2 Music representation

Music for computational processing can be represented either as transformed audio waveforms (Logan et al., 2000; Takuya, 1999) or as symbolic notations that encode musical structure (Rothstein, 1995; Good, 2021; Zhao et al., 2025a; Nienhuys and Nieuwenhuizen, 2003). Among symbolic formats, **ABC Notation** (Walshaw, 2021) encodes pitch, rhythm, meter, and key signature in compact plain text. ABC notation produces significantly shorter sequences than MIDI or MusicXML (Yuan et al., 2024), and LLMs perform better on ABC-encoded tasks (Qu et al., 2024), making it a practical choice without the complexity of audio encoding.

2.3 Music Understanding Benchmarks

Music understanding benchmarks have evolved from narrow Music Information Retrieval (MIR) tasks such as genre classification, beat tracking, and chord estimation (Downie et al., 2014) to broader multi-task evaluations covering acoustic, semantic, and structural reasoning. Benchmarks such as MARBLE (Yuan et al., 2023) and CMI-Bench (Yinghao et al., 2025) evaluate acoustic and semantic capabilities, while AIR-Bench (Yang et al., 2024) and MuChoMusic (Weck et al., 2024) extend evaluation to multimodal audio-language reasoning. For symbolic music, MusicTheoryBench (Yuan et al., 2024), ABC-Eval (Zhao et al., 2025a), ZIQI-Eval (Li et al., 2024), and WildScore (Mundada et al., 2025) benchmark LLM competence primarily in Western tonal traditions. However, existing benchmarks exhibit these limitations: non-Western traditions, including South Asian classical music, are largely absent; symbolic reasoning under raga-based compositional constraints is not evaluated; and culturally grounded musical structures such as raga grammar and tala organization are not represented. Although ZIQI-Eval includes limited South Asian content, it focuses only on surface-level factual questions. These gaps motivate the benchmark introduced in this work.

2.4 Music Generation Frameworks

Based on representations, we observe music generation and evaluation frameworks in both the audio and symbolic domains.

Audio Models Transformer-based (Agostinelli et al., 2023; Kreuk et al., 2022) and diffusion-based audio music generation architectures (Yuan et al., 2025; Liu et al., 2025; Ning et al., 2025) have advanced rapidly but remain constrained by distributional bias toward Western-centric training corpora, with performance degrading across underrepresented cultural traditions (Marták et al., 2026; Ujgare et al., 2025; Sawaengsawangarom et al., 2025). Models further struggle with fine-grained controllability, and consistent text–audio alignment, while diffusion-based approaches remain computationally expensive (Zhao et al., 2025b; Kong, 2025; Zhang et al., 2024). Objective metrics further correlate poorly with human musical judgment, and standardized cross-cultural benchmarks remain limited (Lemercier et al., 2025; Dong, 2025; Chiu, 2025), motivating complementary investigation in symbolic domains.

Symbolic Models In symbolic music generation, models producing scores in formats such as MIDI or ABC notation have progressed from GAN-based systems (MuseGAN (Dong et al., 2018)) to transformer (Museformer (Yu et al., 2022)) and diffusion-based approaches (Zhang et al., 2023), enabling multi-instrument composition. However, symbolic datasets remain small and culturally narrow (Wu and Sun, 2022), exhibit weak harmonic control (Li, 2025), and lack unified objective and perceptual evaluation framework (Liang et al., 2025).

LLMs for Symbolic Music Generation and Non-Western Domains Text-based LLMs have extended symbolic generation capabilities, with ChatMusician (Yuan et al., 2024) treating ABC notation as a second language, SongComposer (Ding et al., 2025) enabling lyrics and melody generation, and ComposerX (Deng et al., 2024) and CoComposer (Xing et al., 2025) extending this through multi-agent prompting frameworks. Cross-lingual transfer studies further show that English-centric pretraining underperforms on low-resource languages (Mehta et al., 2025), a pattern that extends to the music domain, motivating us to conduct a systematic evaluation of LLM competence in South Asian symbolic musical capabilities grounded in Hindustani and Bengali classical traditions.

3 Methodology

3.1 South Asian Music Understanding Benchmark Curation

We curated a 504-question multiple-choice benchmark to evaluate LLMs' capabilities in understanding South Asian music, spanning Hindustani classical music theory and Bengali classical traditions. The benchmark includes three subtasks reflecting distinct but complementary knowledge domains:

- **Music Theory Understanding (163 Questions)** evaluates LLMs' comprehension of the theoretical grammar underlying Hindustani classical music, as practiced in Bengali traditions such as Raga grammar, tala systems, thaat classification, ornamentation, and notation.
- **Music General Knowledge (143 questions)** evaluate factual and cultural knowledge grounded in Bengali and broader Hindustani musical heritage covering Composers, historical figures, instruments, and regional stylistic traditions.
- Similar to (Li et al., 2024), **Music Continuation (198 questions)** task evaluates whether LLMs can apply structural musical knowledge in practice. Each question presents the beginning of a Rabindra or Nazrul Sangeet in ABC notation and asks the model to select the most structurally coherent continuation from four options.

All Examples for each subtask from the benchmark are provided in Appendix 4. All questions were reviewed by an expert in Bengali classical music.

3.1.1 Language and Extraction

Theory and knowledge questions were curated exclusively in Bengali to preserve terminology. Models were prompted in Bengali and allowed to respond freely. Responses in Bengali or English were processed using language-agnostic option-letter extraction (A-D). Outputs were first scanned for explicit answer declarations; otherwise, standalone option letters were counted. Ties or missing predictions were marked incorrect. Full extraction rules are provided in Appendix C.

3.1.2 Distractor Construction

A corpus of 100 manually transcribed ABC scores (50 Rabindra, 50 Nazrul) from official Swaralipi (Hindustani music notation system) served both as continuation seeds and generation references. Continuation distractors were constructed by sampling melodic segments from other songs within the annotated set, deliberately introducing structural mismatches through cross-genre substitution and rhythmic inconsistencies, ensuring that distractors are plausible surface-level continuations while being structurally incorrect with respect to raga consistency or tala alignment.

3.2 Music Generation Prompt Design

We designed five prompt levels that progressively increase contextual information, following the TELeR taxonomy (Santu and Feng, 2023). Prompts incrementally introduce scale, time signature, genre, lyrical theme, and stylistic constraints to examine how conditioning affects generation quality.

For 50 Rabindra and 50 Nazrul lyrics, five prompt variants were used per model, yielding 500 prompts in total. Metadata from the original compositions was incorporated in natural-language form to enable controlled comparison with the references. Full prompts are provided in Appendix B.

4 Experimental Design

4.1 Models

We evaluated 30 open-source LLMs across multiple families and parameter scales, along with 3 proprietary systems, on the music understanding task. All models were evaluated at temperature = 1.0, top-p = 1.0, top-k = -1, a fixed seed of 492, and a maximum generation length of 4096 tokens. For the music generation task, the top 9 models (scoring above 40% on average in 3 benchmark subtasks) were selected for evaluation. QwQ-32B-Preview and Phi-4 were excluded despite meeting the threshold, due to high ABC syntax error rates. Table 1 summarizes all models.

4.2 Evaluation metrics

4.2.1 Music Understanding Evaluation

We extracted the answer choices and calculated the accuracy scores between the ground truth and the generated answer.

4.2.2 Music Generation Evaluation

Automatic Evaluation For the automatic evaluation of generated scores, we structure the evaluation framework into two sections: 1) reference-based evaluation and 2) reference-free evaluation.

In *Reference-Based Evaluation*, we compare the generated lyrics with the original ABC notation that the lyrics belong to, using the widely used KL Divergence (Kullback, 1951).

In *Reference Free Evaluation*, we evaluate the generations on whether the generated score adhered

Family	Models
Open-source	
Qwen 2.5 (Yang et al., 2025)	0.5B–14B It
QwQ (Team, 2024)	32B Preview
Falcon 3 (Almazrouei et al., 2023)	3B, 7B, 10B It
DeepSeek R1 (Guo et al., 2025)	Distill-Qwen (1.5B, 7B, 14B, 32B), Distill-Llama-8B
Gemma 2 (Team et al., 2024)	2B It
Gemma 3 (Team et al., 2025)	1B It
OLMo 2 (OLMo et al., 2024)	7B, 13B It
OLMo (Groeneveld et al., 2024)	7B It
Llama 3.x (Grattafiori et al., 2024)	3.2 (1B, 3B), 3.1 (8B) It
Phi 3.5 (Abdin et al., 2024a)	Mini
Phi 4 (Abdin et al., 2024b)	Phi-4, Phi-4 Mini
Mistral (Jiang et al., 2023)	7B It-v0.3
Cogito-v1-preview (Deep Cogito, 2025)	Llama-3B,8B Preview
ChatMusician (Yuan et al., 2024)	fine-tuned Llama2
Proprietary	
GPT-3.5 (OpenAI, 2022)	—
GPT-4o (Hurst et al., 2024)	—
Gemini 2.5 Pro (Comanici et al., 2025)	—

Table 1: Models evaluated for South Asian music understanding task. "It" stands for Instruct models.

to the given prompt or not, with metrics like Adherence to scale, ABC syntax accuracy (ABC Parse Success Rate), etc. Other aspects of the prompt (e.g., genre, lyrics) were left for human evaluation due to the abstract nature of the task. We further computed feature quality metrics including Repetition rate (Yuan et al., 2024) and Pitch Histogram Entropy (Wu and Yang, 2020). Appendix D provides details on each metric and its implementation.

Human Evaluation Automatic metrics cannot fully capture perceptual quality or stylistic authenticity. We therefore conducted a human evaluation with three annotators who had formal training in music theory and familiarity with both Rabindra and Nazrul Sangeet. A total of 180 samples were evaluated, comprising 20 randomly selected generations (with balanced representation of Rabindra and Nazrul Sangeet) per model from the top nine performing models, all generated using Level 3 prompts, ensuring enough parseable audio samples were available for each model. Level 3 was chosen because it provides explicit structural constraints without extensive stylistic scaffolding; a quantitative justification is provided in Section 5.2.1. Annotators followed standardized annotation guidelines and rated each sample on six criteria: Structureness, Genre analysis, Style analysis, Emotion, Adherence to the instruction, and Harmoniousness. The annotation guideline is provided in Appendix E.

Ordinal criteria (e.g., Structureness) were mapped to numeric scales (1-5) for calculation. For Genre and Style Analysis, we compared annotator-assigned labels against the prompt specification and reported accuracy as the percentage of matches. Labels were determined by majority voting across three annotators; samples with no majority agreement were counted as mismatches.

5 Results

In this section, we focus on answering several research questions: **RQ-1:** Do current LLMs understand grammatical concepts of Indian and Bengali classical music? **RQ-2:** How well do current LLMs exhibit factual and cultural knowledge of Bengali and Hindustani musical heritage? **RQ-3:** Can current LLMs generate stylistically faithful symbolic music scores adhering to the compositional conventions of Bengali classical genres? **RQ-4:** Are the current symbolic music generation evaluation metrics reliable for capturing the structural and stylistic properties of South Asian symbolic music?

5.1 South Asian Music Understanding

Figure 2 presents accuracy scores across all three benchmark subtasks for the 33 evaluated models. Addressing **RQ-1** and **RQ-2**, results indicate that most open-source models show limited understanding of both the theoretical grammar and cultural knowledge underlying Hindustani and Bengali classical music, with Music Theory Understanding scores remaining in the 23-40% range. Poor performance on Music Continuation further extends this deficit beyond theory to the practical application of South Asian musical grammar. Gemini 2.5 Pro substantially outperforms all others across all three subtasks (90.8%, 89.4%, 85.2%), followed by GPT-4o and GPT-3.5, indicating that while frontier models exhibit emergent comprehension of South Asian musical knowledge, most open-source LLMs remain ill-equipped for this culturally grounded domain. Several noteworthy observations:

- *Scaling yields diminishing returns within families.* Within the Qwen2.5 family, the 32B, 14B, and 7B variants score nearly identically on Music Theory Understanding (36.2%, 36.8%, 35.6%), and DeepSeek-R1-Distill Qwen 14B and 32B are identical (39.3%), suggesting a performance ceiling regardless of scale.
- *Music Continuation is anomalously high for some models.* Several mid-tier models score surprisingly high on Music Continuation despite having a poor Music Theory Understanding score (Qwen2.5-32B: 69.3%, Qwen2.5-14B: 60.8%),

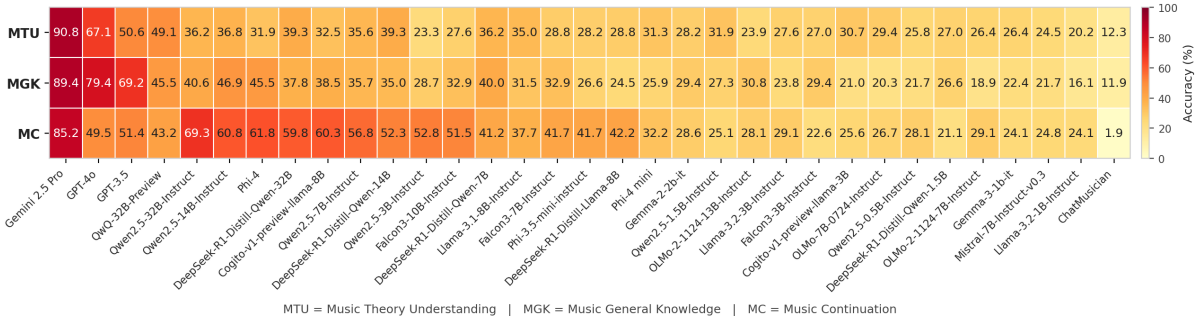

Figure 2: Heatmap of 33 model accuracy scores across three South Asian music understanding subtasks. Rows correspond to tasks and columns correspond to models. Darker colors indicate higher accuracy. Models are sorted by theory understanding accuracy.

suggesting the task may be partially solvable through surface-level pattern matching rather than genuine musical reasoning.

- *Reasoning ability does not substitute for musical knowledge.* QwQ-32B-Preview, a reasoning-focused model, underperforms GPT-3.5 on Theory Understanding (49.1% vs. 50.6%), indicating that general chain-of-thought reasoning does not compensate for the absence of culturally grounded musical knowledge.
- *ChatMusician completely fails on South Asian content.* The only music-specific fine-tuned model scores lowest across all subtasks (12.3%, 11.9%, 1.9%). Output inspection reveals question repetition in Music Continuation and hallucinated Bengali non-answers, confirming that fine-tuning on Western ABC notation does not transfer to South Asian traditions.

5.2 South Asian Music Generation

5.2.1 Automatic Evaluation

Figure 3 shows automatic evaluation scores for the top 9 models across prompt levels L2-L5 (Level 1 excluded due to noisy, unconstrained generations; see Appendix B). Gemini 2.5 Pro achieves the best scores across all metrics, and its pitch distributions most closely match the reference entropy of **2.86**. Among open-source models, Cogito-v1-preview-llama-8B and DeepSeek-R1-Distill-Qwen-14B show relatively competitive KL Divergence scores, though both are considerably lagging behind Gemini 2.5 Pro.

Levels 2 and 3 consistently achieve competitive scores across all metrics (Table 6 shows mean scores across levels), while Levels 4 and 5 yield no appreciable improvement despite additional contextual detail, suggesting that beyond a moderate level of prompt specificity, additional information does not meaningfully enhance generation quality. Level 3 was therefore selected as the representative prompt level for human evaluation.

5.2.2 Human Evaluation

Table 2 presents human evaluation results. Gemini 2.5 Pro achieves the highest scores across all ordinal criteria (3.25-3.65), followed by Cogito-v1-preview-llama-8B. Most open-source models score below 2.5, indicating that structurally parseable outputs do not necessarily translate to perceptually coherent or emotionally aligned melodies.

For Genre and Style Accuracy, Gemini 2.5 Pro maintains strong adherence to the broader South Asian classical genre (95%) but correctly generates the Rabindra or Nazrul Sangeet style in only 40% of cases, revealing the central finding for **RQ-3**. Figure 5 shows that for most models, the majority of generated outputs fall into no recognizable category, suggesting that current LLMs can approximate the surface characteristics of South Asian classical music but lack the compositional specificity needed for faithful stylistic control. It is worth noting that ABC notation cannot approximate some raga-specific gamakas such as meend and microtonal inflections, which may partly contribute to the gap between structural validity and stylistic faithfulness beyond model limitations.

We computed Inter-Annotator Agreement (IAA) across 3 annotators using Krippendorff's α (Krippendorff, 2011), mean weighted Cohen's κ (Cohen, 1960) for ordinal criteria (Structureness, Emotion, Adherence, Harmoniousness, rated on a 1-5 scale), and Fleiss' κ (Fleiss, 1971) for nominal criteria (Genre Analysis and Style Analysis, treated as unordered categorical labels). Majority agreement is reported as the percentage of items where at least 2 of 3 annotators assigned the same label. For ordinal criteria, both α and κ indicates moderate agreement for Structureness ($\alpha = 0.437$, $\kappa = 0.441$) and Harmoniousness ($\alpha = 0.430$, $\kappa = 0.431$), suggesting

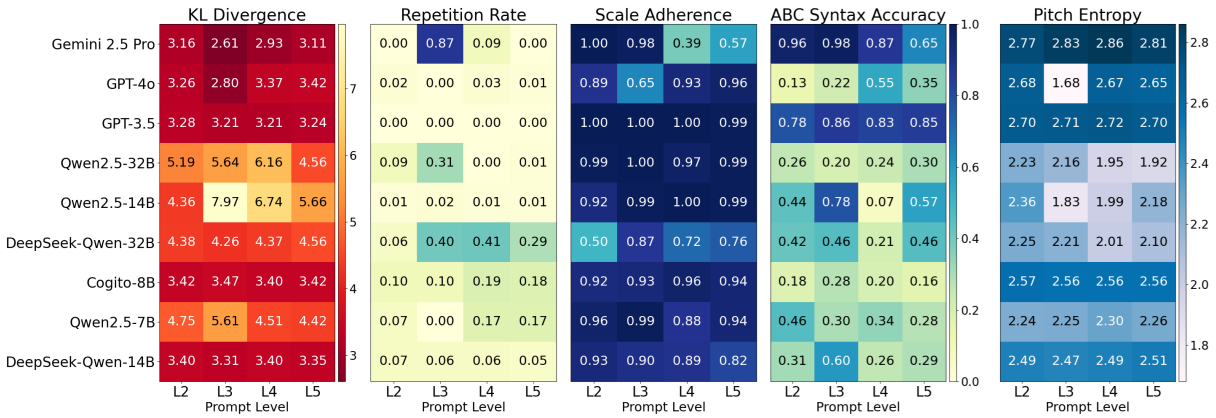

Figure 3: Heatmaps of automatic evaluation metrics across models and prompt levels. Rows correspond to models and columns correspond to prompt levels (L2-L5). Darker colors indicate stronger performance for each metric.

Model	Structureness (1-5) ↑	Emotion (1-5) ↑	Adherence to Instruction (1-5) ↑	Harmoniousness (1-5) ↑	Genre Accuracy ↑	Style Accuracy ↑
Gemini 2.5 Pro	**3.65**	**3.42**	**3.25**	**3.53**	**0.95**	**0.40**
GPT-4o	2.96	2.18	2.35	2.77	0.75	0.40
GPT-3.5	2.43	2.12	2.22	2.97	0.50	0.25
Qwen2.5-32B-Instruct	1.50	1.33	1.55	1.62	0.05	0.05
Qwen2.5-14B-Instruct	1.95	1.35	1.60	2.05	0.00	0.00
DeepSeek-R1-Distill-Qwen-32B	2.07	2.18	2.17	2.25	0.55	0.25
Cogito-v1-preview-llama-8B	3.30	3.00	3.08	3.35	0.65	0.30
Qwen2.5-7B-Instruct	2.33	2.53	2.25	2.32	0.55	0.40
DeepSeek-R1-Distill-Qwen-14B	2.07	2.33	2.05	2.30	0.60	0.25

Table 2: Human evaluation results on 6 criteria. Columns 2-5 show average annotator scores on ordinal criteria (1-5), columns 6-7 show majority-vote accuracy for Genre and Style Analysis.

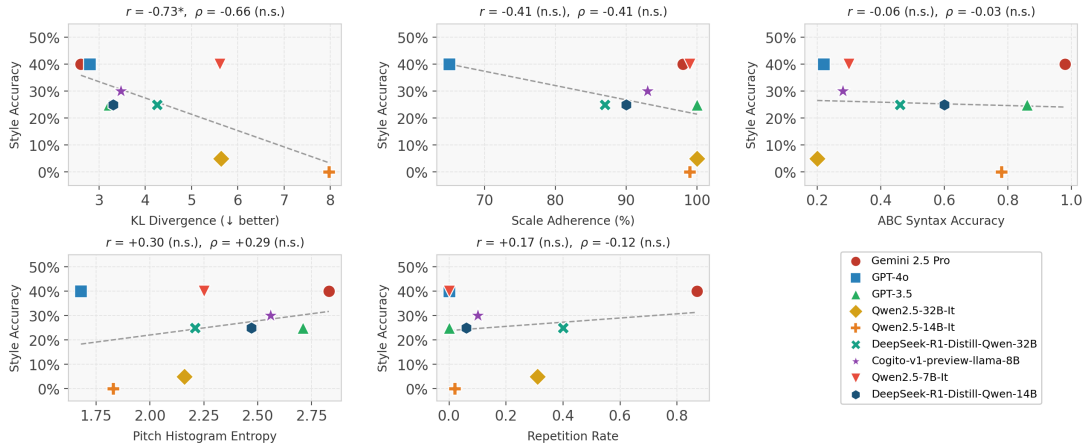

Figure 4: Pearson and Spearman correlations between each automatic evaluation metric and human Style Accuracy across nine models (Level 3 generations). Each point represents one model; the dashed line shows the linear trend. * indicates $p < 0.05$ and n.s. indicates not significant.

Criteria	α	κ	Majority agreement
Structureness	0.437	0.441	83.2%
Emotion	0.325	0.347	78.8%
Adherence	0.268	0.282	77.1%
Harmoniousness	0.430	0.431	81.6%
Genre Analysis	0.152	0.151	100%
Style Analysis	0.206	0.204	92.2%

Table 3: IAA across 6 criteria. α: Krippendorff's alpha, κ: mean weighted Cohen's Kappa (ordinal) / Fleiss' Kappa (nominal).

annotators share reasonable common understanding of structural and harmonic quality, with majority agreement of 83.2% and 81.6% respectively. Adherence to instruction and Emotion show weaker agreement (α = 0.268-0.325), reflecting the inherently subjective nature of these criteria. For nominal criteria, Genre Analysis and Style Analysis show low α and κ ($\alpha = 0.152$, $\kappa = 0.151$; $\alpha = 0.206$, $\kappa = 0.204$), Yet the majority agreement remains high at 100% and 92.2% respectively. This divergence indicates that the low α is driven by the highly skewed label distribution of the generated outputs rather than by genuine annotator disagreement, since annotators consistently converged on the same label. Still, that label was predominantly

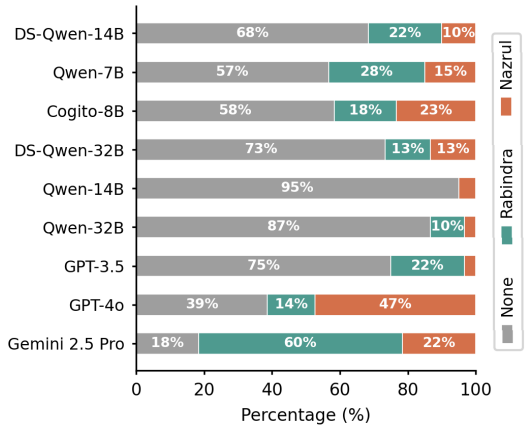

Figure 5: Percentage of generated Styles (Rabindra, Nazrul, None) for each. Here, DS-Qwen = DeepSeek-R1-Distill-Qwen, Qwen = Qwen 2.5 instruct models

one category, confirming that models genuinely failed to generate stylistically distinctive outputs rather than annotators being unable to classify them. Thus answering **RQ-3**: current LLMs demonstrate partial competence in structural generation but fall substantially short of faithful stylistic control in Bengali classical genres, posing a critical open challenge for culturally grounded music modeling.

We further examined whether improved musical understanding translates into improved generation quality in the models. Figure 6 shows Music Theory Understanding and Music General Knowledge have moderate positive correlations with ordinal human evaluation criteria (Structureness, Emotion, Adherence to instruction and Harmoniousness) ($r = 0.35 - 0.67$), suggesting stronger theoretical knowledge tends to produce structurally sounder and emotionally richer outputs. However, correlations with Genre and Style Accuracy are weaker, and Music Continuation is essentially uncorrelated with Style Accuracy ($r = -0.03$), indicating that structural recognition ability does not predict generative stylistic competence.

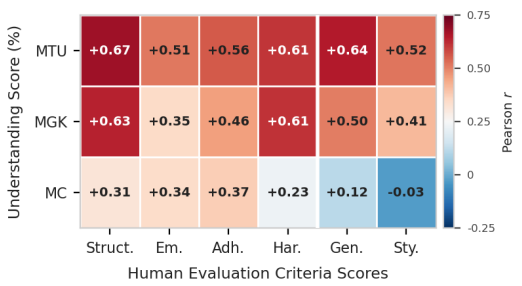

Figure 6: Pearson correlation between Music understanding tasks vs Human evaluation criteria scores on the generation task for the top 9 models. MTU=Music Theory Understanding, MGK=Music General Knowledge, MC=Music Continuation; the x-axis presents structure, emotion, adherence to instruction, harmoniousness, genre, and style accuracy, respectively.

5.2.3 Limitations in Automatic Evaluation

To assess whether automatic metrics reliably capture stylistic faithfulness, we computed Pearson correlations between each automatic metric and human Style Accuracy across all evaluated models (Figure 4). Among five metrics, only KL Divergence shows a statistically significant correlation ($r = -0.73$, $p < 0.05$), yet even the strongest model (Gemini 2.5 Pro) reaches only 40% style accuracy. Scale Adherence and ABC Syntax Accuracy show near-zero correlations ($r = -0.41$ and $r = -0.06$, both non-significant), models achieving 99-100% Scale Adherence score as low as 0% on Style Accuracy, confirming that syntactic correctness is entirely uncorrelated with stylistic faithfulness. Answering **RQ-4**, existing metrics operate at the syntactic and pitch-distribution level and can be satisfied trivially while producing outputs that annotators unanimously classify as stylistically unrecognizable. Developing metrics sensitive to raga-specific ornamentation and the compositional conventions of Bengali classical traditions remains an open problem.

6 Conclusion

We present a systematic evaluation pipeline of 33 open-source and frontier LLMs on South Asian symbolic music understanding and generation, through a Bengali classical music benchmark spanning music theory, cultural knowledge, symbolic continuation, and ABC notation generation. Our results reveal a pronounced capability gap: Gemini 2.5 Pro achieves **90.8%**, **89.4%**, and **85.2%** across the three understanding subtasks, while most open-source models remain in the **23-40%** range. However, we observed a critical gap between structural generation competence and fine-grained stylistic faithfulness: despite reaching **95%** South Asian genre adherence, produces recognizable Rabindra/Nazrul-style melodies only **40%** of the time. Furthermore, existing automatic metrics operating at the syntactic and pitch-distribution level are insufficient for culturally grounded evaluation, highlighting the need for reliable evaluation of non-Western symbolic music as an open problem. These findings suggest that while frontier LLMs offer a promising foundation for low-resource South Asian music modeling, achieving faithful stylistic control and developing culturally grounded evaluation frameworks for music generation remain critical challenges for the field.

Limitations

In this work, the generation evaluation focuses exclusively on two Bengali classical traditions, Rabindra and Nazrul Sangeet, which, while culturally significant and computationally underrepresented, do not fully represent the breadth of South Asian or even Bengali musical heritage. The scope of the Music Theory Understanding subtask is similarly constrained to Hindustani classical theory as practiced in the Bengali tradition. Thus, results on this benchmark should not be generalized to South Asian music competence broadly. Our choice of ABC notation, while motivated by both modeling and data constraints, imposes a ceiling on stylistic faithfulness that is independent of model capability. This information loss occurs at two levels: Swaralipi itself is a skeletal scaffold that hints at ornamentation through hyphens and slur marks but does not prescriptively encode *gamakas*, *meend*, or microtonal inflections, as these are transmitted through oral performance traditions. ABC notation further approximates these hints through grace notes and ties, but continuous pitch glides and microtonal inflections remain inexpressible. Richer formats such as MIDI or MusicXML could better capture such ornamentations. Unfortunately, direct conversion from Swaralipi into MIDI or MusicXML is infeasible because its solfège-based representation lacks the absolute pitch and timing encodings required by these formats. Consequently, reported style accuracy scores reflect faithfulness within ABC's expressive bounds, and the true gap between generated outputs and authentic raga performance practice may be larger than our evaluation captures.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. Preprint, arXiv:2404.14219.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024b. Phi-4 technical report. arXiv preprint arXiv:2412.08905.

Parul Agarwal, Harish Karnick, and Bhiksha Raj. 2013. A comparative study of indian and western music forms. In ISMIR, pages 29–34.

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 arXiv:2301.11325.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, and 1 others. 2023. The falcon series of open language models. arXiv preprint arXiv:2311.16867.

Medha Basu, Archi Banerjee, Shankha Sanyal, Kumardeb Banerjee, and Dipak Ghosh. 2023. Styles and rhythms of musical transitions in indian ragas: An acoustical exploration. In International Symposium on Frontiers of Research in Speech and Music, pages 251–265. Springer.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

Tiffany Chiu. 2025. Artificial intelligence in music generation: Models, evaluation, applications, and future prospects. Theoretical and Natural Science, 151:51–59.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1):37–46.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261.

Sudhin Das. 1996. Nazrul Sangeet Swarlipi (Series). Nazrul Institute, Dhaka. Authoritative swarlipi editions of Nazrul songs prepared by Sudhin Das.

Deep Cogito. 2025. Cogito v1 preview: Introducing IDA as a path to general superintelligence.

Qixin Deng, Qikai Yang, Ruibin Yuan, Yipeng Huang, Yi Wang, Xubo Liu, Zeyue Tian, Jiahao Pan, Ge Zhang, Hanfeng Lin, Yizhi Li, Yinghao Ma, Jie Fu, Chenghua Lin, Emmanouil Benetos, Wenwu Wang, Guangyu Xia, Wei Xue, and Yike Guo. 2024. Composerx: Multi-agent symbolic music composition with llms. Preprint, arXiv:2404.18081.

Shuangrui Ding, Zihan Liu, Xiaoyi Dong, Pan Zhang, Rui Qian, Junhao Huang, Conghui He, Dahua Lin, and Jiaqi Wang. 2025. Songcomposer: A

large language model for lyric and melody generation in song composition. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7108–7127.

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.

Jiaxing Dong. 2025. A survey of text-to-music generation with deep learning. Applied and Computational Engineering, 119:73–80.

J Stephen Downie, Xiao Hu, Jin Ha Lee, Kahyun Choi, Sally Jo Cunningham, and Yun Hao. 2014. Ten years of mirex: reflections, challenges and opportunities. ISMIR.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5):378.

M Good. 2021. Musicxml: An internet-friendly format for sheet music (2001).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, and 1 others. 2024. Olmo: Accelerating the science of language models. In Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers), pages 15789–15809.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Kalyan Sen Gupta. 2016. The Philosophy of Rabindranath Tagore. Routledge.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. Preprint, arXiv:2310.06825.

Xiangchen Kong. 2025. Deep learning in music generation: a comprehensive investigation of models, challenges and future directions. In ITM Web of Conferences, volume 70, page 04027. EDP Sciences.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. arXiv preprint arXiv:2209.15352.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Solomon Kullback. 1951. Kullback-leibler divergence.

Winston E. Langley. 2009. Kazi Nazrul Islam: The Voice of Poetry and the Struggle for Human Wholeness. Nazrul Institute, Dhaka, Bangladesh.

Jean-Marie Lemercier, Julius Richter, Simon Welker, Eloi Moliner, Vesa Välimäki, and Timo Gerkmann. 2025. Diffusion models for audio restoration: A review. IEEE Signal Processing Magazine, 41(6):72–84.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. Advances in neural information processing systems, 35:3843–3857.

Jiajia Li, Lu Yang, Mingni Tang, Cong Chen, Zuchao Li, Ping Wang, and Hai Zhao. 2024. The music maestro or the musically challenged, a massive music evaluation benchmark for large language models. Preprint, arXiv:2406.15885.

Jiawei Li, Yang Gao, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, and 1 others. 2025. Fundamental capabilities and applications of large language models: A survey. ACM Computing Surveys, 58(2):1–42.

Yang Li. 2025. Erld-hc: Entropy-regularized latent diffusion for harmony-constrained symbolic music generation. Entropy, 27(9):901.

Qian Liang, Menghaoran Tang, and Yi Zeng. 2025. Muspike: A benchmark and evaluation framework for symbolic music generation with spiking neural networks. Preprint, arXiv:2508.19251.

Zihan Liu, Shuangrui Ding, Zhixiong Zhang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. 2025. Songgen: A single stage auto-regressive transformer for text-to-song generation. arXiv preprint arXiv:2502.13128.

Beth Logan and 1 others. 2000. Mel frequency cepstral coefficients for music modeling. In Ismir, volume 270, page 11. Plymouth, MA.

Lukáš Samuel Marták, Patricia Hu, and Gerhard Widmer. 2026. Sound and music biases in deep music transcription models: a systematic analysis. EURASIP Journal on Audio, Speech, and Music Processing, 2026(1):5.

Atharva Mehta, Shivam Chauhan, Amirbek Djanibekov, Atharva Kulkarni, Gus Xia, and Monojit Choudhury. 2025. 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.

Gagan Mundada, Yash Vishe, Amit Namburi, Xin Xu, Zachary Novack, Julian McAuley, and Junda Wu. 2025. Wildscore: Benchmarking mllms in-the-wild symbolic music reasoning. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16858–16874.

Han-Wen Nienhuys and Jan Nieuwenhuizen. 2003. Lilypond, a system for automated music engraving. In Proceedings of the xiv colloquium on musical informatics (xiv cim 2003), volume 1, pages 167–171. Citeseer.

Ziqian Ning, Huakang Chen, Yuepeng Jiang, Chunbo Hao, Guobin Ma, Shuai Wang, Jixun Yao, and Lei Xie. 2025. Diffrhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. arXiv preprint arXiv:2503.01183.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, and 1 others. 2024. 2 olmo 2 furious. arXiv preprint arXiv:2501.00656.

OpenAI. 2022. ChatGPT: Optimizing language models for dialogue.

Santosh Kumar Pudaruth. 2016. A reflection on the aesthetics of indian music, with special reference to hindustani raga-sangita. Sage Open, 6(4):2158244016674512.

Xingwei Qu, Yuelin Bai, Yinghao Ma, Ziya Zhou, Ka Man Lo, Jiaheng Liu, Ruibin Yuan, Lejun Min, Xueling Liu, Tianyu Zhang, and 1 others. 2024. Mupt: A generative symbolic music pretrained transformer. arXiv preprint arXiv:2404.06393.

Joseph Rothstein. 1995. MIDI: A comprehensive introduction, volume 7. AR Editions, Inc.

Lewis Rowell. 1986. The ancient tala system: A comparative approach. The Journal of the Music Academy, 57:83–99.

Shubhra Kanti Karmaker Santu and Dongji Feng. 2023. Teler: A general taxonomy of llm prompts for benchmarking complex tasks. Preprint, arXiv:2305.11430.

Pasin Sawaengsawangarom, Suparoek Phongoen, and Papis Wongchaisuwat. 2025. Deep learning for music genre classification: A case study of thai music. In Proceedings of the 2025 9th International Conference on Control Engineering and Artificial Intelligence, pages 40–44.

Rabindranath Tagore. 1960. Gitabitan. Visva-Bharati. Collection of the 2,232 songs of Rabindranath Tagore.

Rabindranath Tagore. 2014. Gitabitan. Visva-Bharati, Santiniketan, India. Original compositions written between 1880–1941.

Fujishima Takuya. 1999. Realtime chord recognition of musical sound: Asystem using common lisp music. In Proceedings of the International Computer Music Conference 1999, Beijing.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. arXiv preprint arXiv:2503.19786.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118.

Qwen Team. 2024. Qwq: Reflect deeply on the boundaries of the unknown.

Nitin S. Ujgare, Roshni Bhamare, Apurwa Patil, Yogeshwari Rajput, Pranjal Patil, and Pravin Pachorkar. 2025. A comprehensive survey on machine and deep learning techniques for genre prediction in folk songs. In 2025 International Conference on Future Technologies (ICFT), pages 1–6.

Chris Walshaw. 2021. The abc music standard 2.1 (dec 2011).

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. arXiv preprint arXiv:2408.01337.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Shangda Wu and Maosong Sun. 2022. Exploring the efficacy of pre-trained checkpoints in text-to-music generation task. arXiv preprint arXiv:2211.11216.

Shih-Lun Wu and Yi-Hsuan Yang. 2020. The jazz transformer on the front line: Exploring the shortcomings of ai-composed music through quantitative measures. arXiv preprint arXiv:2008.01307.

Peiwen Xing, Aske Plaat, and Niki van Stein. 2025. Co-composer: Llm multi-agent collaborative music composition. Preprint, arXiv:2509.00132.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, and 1 others. 2025. Qwen2. 5-1m technical report. arXiv preprint arXiv:2501.15383.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and 1 others. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1979–1998.

MA Yinghao, Siyou Li, Juntao Yu, Emmanouil Benetos, and Akira Maezawa. 2025. Cmi-bench: A comprehensive benchmark for evaluating music instruction following. In Ismir 2025 Hybrid Conference.

Botao Yu, Peiling Lu, Rui Wang, Wei Hu, Xu Tan, Wei Ye, Shikun Zhang, Tao Qin, and Tie-Yan Liu. 2022. Museformer: Transformer with fine- and coarse-grained attention for music generation. Advances in neural information processing systems, 35:1376–1388.

Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, and 1 others. 2025. Yue: Scaling open foundation models for long-form music generation. arXiv preprint arXiv:2503.08638.

Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, and 1 others. 2024. Chatmusician: Understanding and generating music intrinsically with llm. arXiv preprint arXiv:2402.16153.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, and 1 others. 2023. Marble: Music audio representation benchmark for universal evaluation. Advances in Neural Information Processing Systems, 36:39626–39647.

Chen Zhang, Yi Ren, Kejun Zhang, and Shuicheng Yan. 2023. Sdmuse: Stochastic differential music editing and generation via hybrid representation. IEEE Transactions on Multimedia.

Ran Zhang and Steffen Eger. 2024. Llm-based multi-agent poetry generation in non-cooperative environments. arXiv preprint arXiv:2409.03659.

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.

Jiahao Zhao, Yunjia Li, Wei Li, and Kazuyoshi Yoshii. 2025a. Abc-eval: Benchmarking large language models on symbolic music understanding and instruction following. arXiv preprint arXiv:2509.23350.

Yujia Zhao, Mingzhi Yang, Yujia Lin, Xiaohong Zhang, Feifei Shi, Zongjie Wang, Jianguo Ding, and Huansheng Ning. 2025b. Ai-enabled text-to-music generation: A comprehensive review of methods, frameworks, and future directions. Electronics, 14(6):1197.

Zoie Zhao, Sophie Song, Bridget Duah, Jamie Macbeth, Scott Carter, Monica P Van, Nayeli Suseth Bravo, Matthew Klenk, Kate Sick, and Alexandre LS Filipowicz. 2023. More human than human: Llm-generated narratives outperform human-llm interleaved narratives. In Proceedings of the 15th Conference on Creativity and Cognition, pages 368–370.

A Benchmark Examples

Table 4 shows one example per sub-task from the South Asian Music Understanding Benchmark.

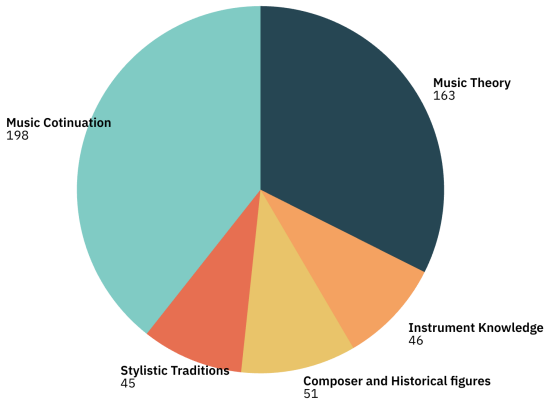

Figure 7: Number of questions per sub-theme

B Prompts for Music Generation

Table 5 shows the exact curated 5 level of prompts with increasing amount of details following the TELeR taxonomy (Santu and Feng, 2023) for music generation in ABC notation. The meta-data that was provided were lyrics, theme, time signature from the original songs. Since the official Bengali Swaralipi notation system represents pitches using solfège syllables (সা, রে, গা, মা, পা, ধা, নি) relative to a tonic without specifying an absolute pitch, a mapping convention was required during the manual transcription of Swaralipi scores into ABC notation. As ABC notation requires an explicit key signature declaration (`K:`), we adopted the convention of mapping সা (Sa, the tonic) to C, yielding the

Benchmark Sub-Task	Example
Music General Knowledge	ঠুংরী গানের সর্বপ্রথম প্রচলন হয় কোথায়? A. বানারসী B. আগ্রা C. দিল্লী D. লক্ষ্ণৌ What is the correct answer? Strictly output only the answer as option A or B or C or D in this following format: [[Answer:]] **Translation:** Where was Thumri music first popularized? A. Benares B. Agra C. Delhi D. Lucknow
Music Theory Understanding	কোনটি সত্য নয়? A. ঠাট কখনো সাত স্বরের কম বা বেশি দিয়ে রচিত হয়না। B. ঠাট কেবল অবরোহী হয়। C. ঠাটের সংখ্যা মাত্র ১০ টি। D. ১০ টি ঠাট থেকেই সকল রাগের সৃষ্টি। What is the correct answer? Strictly output only the answer as option A or B or C or D in this following format: [[Answer:]] **Translation:** Which of the following is not true? A. A Thaat is never composed with fewer or more than seven notes. B. A Thaat is only descending in nature. C. There are only 10 Thaats. D. All ragas originate from these 10 Thaats.
Music Continuation	Please select the most matching melody continuation segment based on the entered melody: X: 1 T: হলুদ গাঁদার ফুল M: 3/4 K: C |: (D2 F-F) D2 | (F2 G) | A3 | | (D F-F) D2 | (F2 G) | A3 | | A2 (A_B | A) (G F) | (F G) (GA |G) (F E)| | (E F E) | (DC _BA,2) | (A, C2-C) (C D) | A. | B cB AB ABA | (GA G3) | GA G FE D | EG3 GG | | A B c ced | c3 | C C D EF | (FGF E3) | | F G {G}AG F | ({F}GF E2) DC | D EF {D}E DED | C4 | B. | (E F) (E | D) D2 | {D}C3-C3 :| | G2 G-G G2 | A3 | G3 | | E (E G) | G G2 | A3 | G3 | | (G A) (A| c) (c d) | c_B3 | A AG2 | C. | (GA G) F | E F2-F2 D_E | _E (_E D) | | z2 CD | D (C B,) |: C ({B}C D_E) | _E D2 | | D (D _E) | F G2 | F (F _E) | D (D_E F_E) | D. | _A2 _A2 | B2 c c | _d3 c | (cB -c) (_B -_A) :| | _A _e _e _e-_e _e (_e _d)| _d _d c2 | _B _B c2 | | (_d3 -C) | c_B c_B (_B _A) |

Table 4: Examples from each of the subtasks of the South Asian Music Understanding Benchmark used for evaluating Music General Knowledge, Music Theory Understanding, and Music continuation. Figure 7 the theme-wise number distribution of questions of the benchmark.

correspondence: সা=C, রে=D, গা=E, মা=F, পা=G, ধা=A, নি=B. All 100 ground truth ABC notations were transcribed under this fixed mapping, resulting in all scores being rooted in Scale C. Consequently, all music generation prompts were constrained to Scale C to ensure a consistent and fair basis for comparison between the generated outputs and the ground truth scores.

C Answer Extractor

To extract the one letter answer (A,B,C or D) for the multiple choice question, first, the model response is optionally stripped of the original prompt to avoid matches from the question or answer choices. The system then attempts to extract a definitive answer using a set of regular-expression patterns that capture explicit answer declarations (e.g., "[[Answer: C]]", "The correct answer is B", or "Option D is correct"). If a match is found, the corresponding option (A-D) is returned and the prediction is deemed unambiguous. When no explicit pattern is detected, a fallback strategy is applied in which all standalone mentions of answer letters (A–D) are identified and counted. If a single option occurs most frequently, it is selected as the model's prediction; if multiple options are tied for the highest frequency, the

Level	Prompt
Level 1 – Minimum Details	Generate a melody in ABC notation for the full lyrics [lyrics].
Level 2 – Moderate Details	Generate a melody in ABC notation for the full lyrics [lyrics], in Scale C with 3/4 time signature, without any comments and empty lines. The song will have a indian classical music theme and cover the theme of the lyrics which is about [theme] and feature instruments used in indian classical music. The song should be about 2–3 minutes length.
Level 3 – Moderate Listwise Details	Generate a melody in ABC notation for the given lyrics by performing the following tasks: 1. Ensure the melody adheres to a Indian classical music style. 2. It should sound like a [Rabindra/Nazrul] Song. 3. Use Scale C and [time signature] time signature. 4. Strictly generate just notations, avoid any comments and extra line. 5. Cover the lyrical theme: [theme], ensuring the output spans ∼2–3 minutes in play length. 6. The music should feature instruments used in indian classical music. [Lyrics]
Level 4 – Significant Details + User Expectation	Level 3 prompt + A good output should: • Be musically coherent and playable. • Clearly reflect the theme [theme]. • Maintain stylistic authenticity of Indian classical music. • Be structured in valid ABC notation syntax. Lyrics: [lyrics]
Level 5 – Maximum Details	Level 4 prompt + Additional background: Indian classical music typically involves raga-based melodies which works with a specific set of notes, ornamentations (gamakas), microtonal variations and an emotional or spiritual mood that matches the lyrics well and rhythmic cycles (taals). Unlike Western classical music, Indian classical tradition emphasizes monophonic texture -centered on melodic development rather than harmonic progression. Lyrics: [lyrics]

Table 5: Prompt levels for ABC melody generation from lyrics.

Metric	L2	L3	L4	L5
KL Divergence (↓)	**3.911**	4.320	4.232	3.971
Repetition Rate (%) (↑)	0.047	**0.196**	0.107	0.080
Scale Adherence (%) (↑)	0.901	**0.923**	0.860	0.884
Pitch Histogram Entropy	2.477	2.300	2.394	2.411
ABC Syntax Accuracy (%) (↑)	0.438	**0.520**	0.397	0.434

Table 6: Mean automatic evaluation scores across all 9 models per prompt level (L2–L5).

prediction is marked as ambiguous. Finally, if no valid answer option can be extracted from the output, the response is classified as ambiguous with no predicted answer. The ambiguous matches were checked manually afterwards for definitive answers.

D Automatic Evaluation Metrics

D.1 KL Divergence

KL Divergence measures the statistical distance between two probability distributions, quantifying how much one distribution differs from a reference distribution, where a lower value indicates greater similarity between the two distributions. We computed the KL Divergence between the pitch distributions of the original and generated ABC notation scores as a reference-based evaluation metric. For each song, pitches were extracted from the ABC strings using a regular expression pattern that captures standard note names (A-G), their octave modifiers, and accidentals, followed by normalization to remove octave indicators and convert to uppercase. A pitch histogram P was then constructed for each score by computing the relative frequency of each pitch class, formally defined as $P(x) = \frac{c(x)}{\sum_{x'} c(x')}$, where $c(x)$ denotes the count of pitch x. To measure how much the generated pitch distribution Q diverges from the original distribution P, we computed the KL Divergence as:

$$D_{KL}(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \quad (1)$$

To handle zero-probability pitch classes, a small smoothing term $\epsilon = 10^{-10}$ was added to both distributions before normalization, ensuring numerical stability. The final reported score is the average KL Divergence across all song pairs within each model and prompt level, where lower values indicate that the generated pitch distribution more closely resembles the original.

D.2 Pitch Histogram Entropy

Pitch Histogram Entropy measures the diversity and uniformity of pitch usage in a melody, where a higher value indicates a more varied and evenly distributed use of pitches across the generated score.

We computed the Pitch Histogram Entropy of the generated ABC notation scores as a reference-free evaluation metric. For each generated score, pitches were extracted using the same regular expression pattern described above, normalized to remove octave indicators and converted to uppercase. A pitch probability distribution was then constructed by computing the relative frequency of each pitch class x as $P(x) = \frac{c(x)}{\sum_{x'} c(x')}$, where $c(x)$ denotes the count of pitch x. The entropy of the pitch histogram was then computed as:

$$H = -\sum_{x} P(x) \log_2 P(x) \tag{2}$$

The final reported score is the average entropy across all generated scores per model and prompt level. A higher entropy value reflects greater melodic diversity, suggesting the model utilizes a wider and more balanced range of pitches, whereas a lower entropy indicates a tendency toward repetitive or pitch-restricted melodic output.

D.3 Repetition Rate

Repetition Rate measures the proportion of generated scores that contain explicit repeat signs, where a higher value indicates greater use of structured musical repetition, which is a characteristic feature of well-formed compositions.

We computed the Repetition Rate of the generated ABC notation scores as a reference-free evaluation metric reflecting structural organization. For each generated score, we checked for the presence of the ABC notation repeat sign `|:`, which denotes a repeated section in the score. The Repetition Rate was then computed as:

$$\text{Repetition Rate} = \frac{\text{\# of scores containing } |: }{\text{Total scores}} \tag{3}$$

The final reported score is the proportion of scores per model and prompt level that contained at least one repeat sign. A higher repetition rate suggests that the model generates more structurally coherent melodies with intentional repeated sections, which aligns with the compositional conventions of both Rabindra and Nazrul Sangeet where returning melodic phrases such as the sthayi are a defining structural characteristic.

D.4 ABC Syntax Accuracy

ABC Parse Success Rate measures the proportion of generated scores that are syntactically valid and parseable as ABC notation, where a higher value indicates that the model produces more well-formed and structurally correct musical scores.

We computed the ABC Parse Success Rate of the generated scores as a reference-free evaluation metric reflecting syntactic correctness. For each generated ABC notation file, we attempted to parse and convert it to MIDI using the **music21** library[1]. If the parse succeeded and a valid MIDI file was produced, the score was counted as a successful parse; if an exception was raised during conversion, the file was counted as a parse failure. The ABC Parse Success Rate was then computed as:

$$\text{ABC Parse Success Rate} = \frac{\text{\#of successfully parsed scores}}{\text{Total scores}} \tag{4}$$

Successfully parsed files were subsequently converted to WAV audio using **FluidSynth**[2] with a SoundFont file, enabling both syntactic validation and downstream human evaluation of the generated audio. The final reported score is the proportion of scores per model and prompt level that were successfully parsed, where a higher rate indicates the model generates more syntactically valid ABC notation that conforms to the formal grammar expected by standard ABC parsers.

D.5 Scale Adherence

Scale Adherence measures the proportion of generated scores that strictly conform to the specified target scale, where a higher value indicates that the model generates melodies that remain within the prescribed set of allowable pitches.

We computed the Scale Adherence of the generated ABC notation scores as a reference-free evaluation metric reflecting pitch-level instruction following. Since all prompts specified Scale C, we defined the allowed pitch set as the C major scale {C, D, E, F, G, A, B} along with a set of permitted flat pitches {♭E, ♭A, ♭B} that are characteristic of certain Indian classical ragas within that scale context. For each generated score, pitches were extracted and normalized using the same procedure described above. A score was considered scale-adherent if and only if every extracted pitch belonged to the allowed pitch set, with any occurrence of a sharp

[1] https://music21.org/music21docs/

[2] https://github.com/FluidSynth/fluidsynth

(ˆ) or natural sign (=) prefix immediately marking the score as non-adherent. The Scale Adherence rate was then computed as:

$$\text{Scale Adherence} = \frac{\text{\# of scores fully adhering to Scale C}}{\text{Total scores}} \tag{5}$$

The final reported score is the proportion of scores per model and prompt level that passed the scale adherence check. A higher score indicates the model more reliably respects the scale constraint provided in the prompt, which is a necessary condition for stylistic and theoretical alignment with the target musical tradition.

E Annotation Guideline

Annotators were first briefed about the project and how the collected data was going to be utilized. Consent was taken on data usage and no personal information were collected during the evaluation. The annotators were given 9 anonymous folders including 20 samples each, not disclosing the models' names. The instruction given to the annotators is written verbatim below:

Each annotator receives three items: A folder named "Audio files", An Excel file (annotator#.xlsx) and this instructions document.

1. Folder structure: Inside the "Audio files" folder, there are 9 subfolders. Each subfolder corresponds to a different large language model (LLM) and contains 20 audio files generated by that model. The names of these 9 subfolders exactly match the sheet names in the Excel file.

2. Matching folders and sheets: In annotator#.xlsx, there are 9 sheets, each named after one of the subfolders. When you work on a specific folder, make sure you are also filling out the sheet with the same name. *Example:* If you are annotating audios from the folder "Folder 1", fill in the sheet named "Folder 1".

3. Understanding the Excel file: Each row in a sheet represents one audio file. The "song no." column specifies the exact file name you need to work on. The full file path is shown as: *folder_name/song_name.wav*. *Example:* "Folder 1/5.wav" means the file is inside "Folder 1".

4. Annotation procedure: Find the audio file mentioned in the "song no." column. Read the prompt in the same row of the Excel sheet, this describes what the model was asked to generate. Listen carefully to the corresponding audio file. Fill in the remaining columns in that row based on the instructions given below.

Note: Even if the lyrics are well-known Najrul/Rabindra songs, as LLMs are asked to generate new melodies, the new melody might be totally different from the original song that we are used to listening to. Annotators should assess based on how the new melody aligns with the given theme and lyrics.

Each of the cells includes a drop-down menu to select your answer from. The drop-down button will show up on the top-right corner of the cell when you select it.

Annotation Guideline: Given a prompt including lyrics, theme, scale and time signature and their corresponding generated melody, listen to the melody and score it according to the criteria below.

Structureness Is the melody structured nicely and has distinct sections like:

- **sthayi** (স্থায়ী): Functions like the "base" section to which the singer returns after improvisations.
- **antara** (অন্তরা): Expands the melody, usually in a higher octave; provides lyrical variation before returning to sthayi and can be paired with the sthayi (the performance alternates between them).

Score	Description
Unstructured	The melody is linear and repetitive, with no clear return point or section contrast. No identifiable sthayi or antara.
Slightly Structured	There are minor variations or pauses suggesting different ideas, but sections are not clearly distinguished or cyclic.
Moderately Structured	The melody shows partial organization – a recognizable main idea (sthayi) with one contrasting section, though transitions may be unclear.
Well Structured	The melody clearly alternates between sthayi and antara, with evident octave or lyrical variation before returning to the base theme.
Highly Structured	The melody exhibits strong, distinct sections – sthayi and antara are clearly defined, cyclic, balanced, and thematically coherent throughout.

Genre Analysis Can the melody be classified as Indian classical music, including characteristics such as microtones, ornamentations/gamakas, and intricate tunes? Answer *Yes* or *No*.

Style Analysis Can the melody be classified as having a feel like Rabindra Sangeet (রবীন্দ্র সঙ্গীত) or Nazrul Sangeet (নজরুল সঙ্গীত)? Answer *Rabindra*, *Nazrul*, or *None*.

Emotion Does the emotion of the melody align with the given lyrics?

Score	Description
Not Aligned at All	The melody conveys a completely different or conflicting emotion compared to the lyrics (e.g., cheerful tune for sorrowful lyrics).
Weakly Aligned	A few emotional cues match, but overall the melody feels disconnected or mismatched with the lyrical sentiment.
Moderately Aligned	The melody captures the general mood but lacks depth or consistency in expressing the intended emotion.
Well Aligned	The melody successfully supports and enhances the lyrical emotion, showing consistency in tone and phrasing.
Strongly Aligned	The melody deeply embodies the lyrical emotion – dynamics, phrasing, and tonal quality fully reinforce the sentiment of the lyrics.

Adherence to Instruction (REL) How much does the melody adhere to the given instructions?

Score	Description
Not at All	The melody ignores the provided instructions entirely – wrong scale, rhythm, emotion, or structure; unrelated to the given prompt.
Slightly Adhering	Some aspects of the instructions are present (e.g., correct tempo or partial scale), but major elements like theme or mood are missing or inconsistent.
Moderately Adhering	The melody reflects several key aspects of the instructions (scale, emotion, or timing) but lacks precision or full consistency.
Well Adhering	The melody follows most of the given instructions accurately – correct scale, fitting time signature, and thematic relevance with minor deviations.
Fully Adhering	The melody strictly and consistently follows all given instructions (scale, time signature, emotion, theme, structure) with high fidelity and musical coherence.

Harmoniousness Is the melody properly in tune, harmonious, and interesting?.

Score	Description
Not Harmonious	The melody sounds off-key, dissonant, or unpleasant. Pitches clash frequently, creating instability or discomfort.
Weakly Harmonious	Some notes or intervals fit the scale, but tuning inconsistencies or awkward jumps make it sound rough or unbalanced.
Moderately Harmonious	The melody generally stays in tune and follows harmonic expectations, but lacks richness or engaging transitions.
Well Harmonious	The melody is tuneful and balanced; intervals and phrasing flow naturally with few discordant moments. Pleasant but not deeply striking.
Highly Harmonious & Engaging	The melody is perfectly in tune, musically rich, and emotionally engaging. It maintains consonance, variation, and aesthetic appeal throughout.