IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Technical Evaluations in the Field of AI Music Composition

[1]Prithaviraj Ashok Londhe, [2]Prof. M. S. Dighe, [3]Prof. L. S Bhadake, [4]Dr. H.B. Jadhav
[1]Student, Department of Computer Engineering
[2,3]Assistant Prof, Department of Computer Engineering
[3]HOD, Department of Computer Engineering
Adsul's Technical Campus, Chas, Ahilyanagar, India

Abstract: *AI-based music composition has progressed from symbolic melody generation to systems that produce audio from natural-language prompts, melodic references, or combinations of both. However, the perceived usefulness of a generated track cannot be captured by audio realism alone. A system may sound convincing yet ignore the requested mood, lose musical structure over time, offer little editing control, or provide insufficient information about the provenance of its training data. This paper develops a practical, multidimensional framework for evaluating AI music composition systems. The framework organizes evaluation into five dimensions: audio and musical quality, prompt and structure alignment, controllability and editability, workflow usefulness, and responsible-use transparency. It proposes a staged evaluation protocol combining listening tests, task-based assessment, technical checks, and documentation review. The framework is a conceptual synthesis informed by published work on text-to-music generation, controllable music models, synthetic-music detection, and the supplied systematic review of AI music composition. No new experimental dataset or performance results are claimed. The paper's contribution is an evaluation blueprint that researchers, developers, educators, and music practitioners can adapt before conducting empirical validation.*

Keywords: AI music composition, text-to-music, generative AI, music evaluation, controllability, responsible AI, provenance

1. INTRODUCTION

Artificial intelligence has become an active area of research in music composition, production, and sound design. Earlier computational approaches often worked with symbolic representations such as notes, durations, and musical rules. More recent generative systems can produce waveforms or compressed audio tokens and accept natural-language descriptions as input. This shift lowers the technical barrier to experimentation, but it also changes what should be evaluated: a user is no longer judging only whether a melody is plausible, but whether an entire audio output satisfies an intention.

Recent text-to-music research illustrates this change. MusicLM frames generation as hierarchical sequence-to-sequence modelling and explores music generation conditioned on text and melody . MusicGen uses a single-stage Transformer language model over compressed discrete audio representations and supports text or melody conditioning . These systems demonstrate different design choices for connecting human instructions to audio. A broad review by Yang et al. maps the development, research hotspots, and technical evaluations of AI music composition across literature published from 2000 to 2023 .

Despite progress, a single quality score can hide important differences between systems. A generated track may be acoustically clean but musically repetitive; it may match a genre label while missing the requested instrumentation; or it may satisfy a prompt but be difficult to revise. In addition, the provenance of training material and the identification of synthetic audio are increasingly relevant to responsible deployment .

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40525

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

This paper proposes a practical framework for evaluating these dimensions together. It does not claim to present a new trained model, a user study, or measured benchmark results. Instead, it synthesizes relevant research into a structured evaluation blueprint and identifies how the blueprint can be tested in future work.

II. BACKGROUND AND RELATED WORK

A. From Symbolic Composition to Audio Generation

Computational music generation spans several representational levels. Symbolic systems generate note events or scores, which can be inspected and edited directly. Audio-generation systems operate closer to the final listening experience and must model timbre, dynamics, texture, and temporal continuity as well as musical organization. Audio tokenization can compress a waveform into a sequence of discrete representations, making sequence models applicable to music generation. The representation selected by a system affects its computational demands, the detail it can preserve, and the types of controls that can be offered.

B. Text and Melody Conditioning

Text conditioning lets a user describe attributes such as genre, mood, tempo, instrumentation, or scene. MusicLM demonstrates text-conditioned generation and melody-guided transformation , while MusicGen presents a controllable approach that accepts textual descriptions or melodic features . Such controls are useful, but a prompt is inherently ambiguous: terms such as 'cinematic', 'warm', or 'energetic' can mean different things to different listeners. Therefore, evaluation should check both technical output quality and alignment with a clearly defined user intention.

C. Evaluation and Responsible Use

The supplied systematic review surveys AI music composition development and technical evaluation, providing a broad map of the field Separate work has begun to address whether synthetic music can be detected and attributed: Comanducci, Bestagini, and Tubaro introduce FakeMusicCaps as a dataset for evaluating detection and attribution of music generated by text-to-music models . Research on AI and copyright in music also highlights dataset transparency, creator participation, and the ability to opt out as important policy considerations . These strands suggest that evaluation should include provenance and governance rather than treating them as unrelated administrative concerns.

III. PROBLEM STATEMENT AND OBJECTIVES

The central problem is the lack of a compact, practical evaluation structure that considers the complete user experience of an AI music composition system while keeping musical quality, instruction following, editability, workflow value, and responsible-use information distinct. If these dimensions are collapsed into one score, weaknesses can be obscured and comparisons between tools may become misleading.

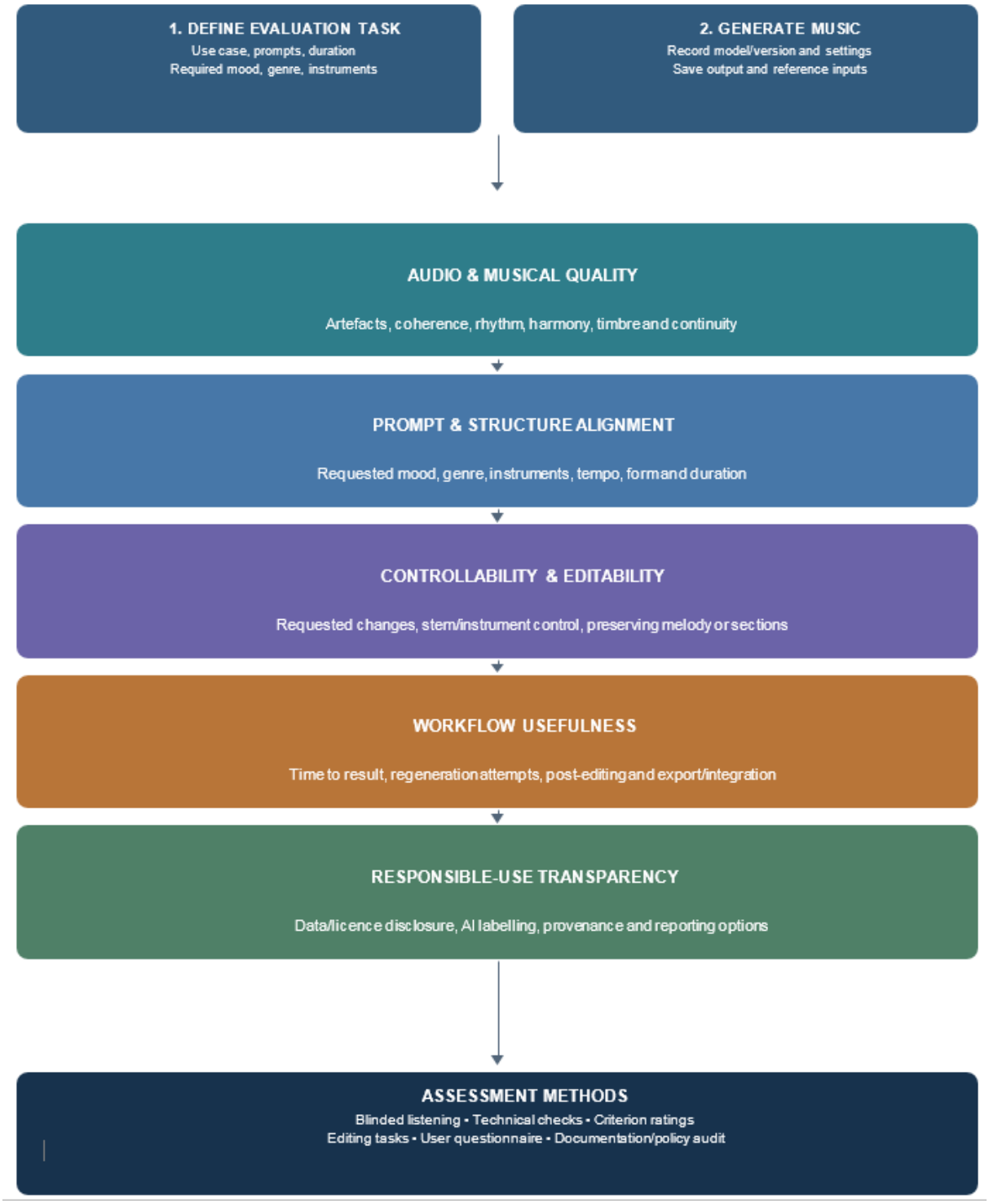
The objectives of this paper are:

- to define a set of complementary dimensions for evaluating AI-generated music;
- to suggest measurable indicators and appropriate assessment methods for each dimension;
- to outline a reproducible evaluation protocol that can be adapted to different tools and use cases; and
- to identify research and deployment risks that should be addressed before broad adoption.

IV. PROPOSED EVALUATION FRAMEWORK

The proposed framework contains five dimensions. Each dimension should be reported separately before any aggregate score is considered. The dimensions are intended as a starting point, not as a validated universal benchmark.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2



Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40525

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

A. Dimension 1: Audio and Musical Quality

This dimension assesses whether the output is listenable and musically coherent. Indicators may include audible artefacts, clipping, abrupt transitions, rhythmic stability, melodic coherence, harmonic consistency, timbral balance, and continuity over the requested duration. Some indicators can be checked computationally, but musical coherence and listening comfort require human judgement. A clean waveform should not automatically be treated as a successful composition.

B. Dimension 2: Prompt and Structure Alignment

Prompt alignment evaluates whether the generated music reflects the requested attributes. A prompt should be decomposed into explicit criteria—for example, target mood, prominent instruments, approximate tempo, genre, vocal or instrumental status, and intended use. Assessors can then rate each criterion as absent, partially present, or clearly present. Structural alignment should separately consider whether the output develops as requested, such as building toward a climax or maintaining a consistent background texture. This reduces the risk of relying only on a broad overall impression.

C. Dimension 3: Controllability and Editability

A practical system should allow users to steer results and revise them without restarting the entire process. Relevant indicators include support for melody conditioning, duration control, section-level regeneration, variation generation, instrument or stem control, and preservation of user-provided musical material. A simple task-based test can ask participants to make a specified change—such as reducing percussion while retaining the melody—and record whether the result meets the instruction, how many attempts were required, and how much manual editing followed.

D. Dimension 4: Workflow Usefulness

The usefulness of generated music depends on the context. A creator seeking a short background cue may value speed and predictable duration, while a composer may prioritize editable structure and control over individual parts. Evaluation should therefore use defined tasks and user groups rather than assuming one output style is best for every application. Suggested indicators include time to obtain an acceptable result, number of regeneration attempts, amount of post-editing, ease of exporting or integrating the audio, and user-reported confidence in the final selection.

E. Dimension 5: Responsible-Use Transparency

This dimension examines the information and safeguards surrounding the model. It can include disclosure of known training-data sources and licences, clear terms for commercial use, mechanisms for reporting suspected imitation, provenance metadata, and transparent labelling of AI-generated outputs. These indicators do not determine legal status on their own. They help users understand uncertainty and make informed decisions. Detection systems can support analysis, but detection should not be presented as infallible; the FakeMusicCaps work provides a research resource for studying detection and attribution .

TABLE I. PROPOSED DIMENSIONS AND INDICATORS

Dimension	Example indicators	Suggested method
Audio and musical quality	Artefacts, coherence, transitions, rhythmic and harmonic stability	Listening rubric plus technical audio checks
Prompt and structure alignment	Presence of requested mood, instruments, tempo, form and duration	Criterion-by-criterion human ratings
Controllability and editability	Success of requested changes; preservation of selected elements	Standardized editing tasks and attempt counts
Workflow usefulness	Time, regeneration count, post-editing effort, export/integration	Task observation and user questionnaire
Responsible-use transparency	Dataset/licence disclosure, output labelling, provenance and reporting options	Documentation and policy audit

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

V. RECOMMENDED EVALUATION PROTOCOL

A reproducible evaluation should begin with a clearly specified use case and a fixed set of prompts. The following protocol is proposed for future empirical testing.

• Define the task: specify whether the target is a short background cue, a full composition, a melody variation, or a production asset. Record duration, intended use, and required musical attributes.

• Prepare a prompt set: create prompts that vary in genre, mood, instrumentation, structural requirements, and specificity. Include repeated prompts to examine output variability.

• Standardize generation conditions: record the system and version, date, settings, prompt text, seed where available, duration, and any reference audio or melody used.

• Use blinded listening where feasible: randomize file order and hide system names from assessors. Include multiple listeners and a clear rating rubric to reduce individual preference effects.

• Measure task outcomes: record alignment ratings, technical defects, editing success, time to acceptable output, number of attempts, and post-processing effort.

• Audit documentation: record what the provider discloses about data, licensing, commercial use, output labelling, and provenance. Mark unknown information as unknown rather than assuming compliance.

• Report results by dimension: publish the prompts, protocol, evaluation criteria, limitations, and aggregate results. Do not combine dimensions into a single ranking unless the weighting method is justified.

For a formal experiment, the protocol should include an appropriate sample size, participant recruitment and consent procedures, inter-rater agreement, and a pre-defined analysis plan. If the study uses copyrighted recordings or human-subject data, the research team should review the applicable permissions, institutional requirements, and local law before collecting or publishing material.

VI. DISCUSSION

A. Why Separate Scores Matter

A multidimensional report makes trade-offs visible. A system might achieve strong prompt alignment but provide limited editability; another may produce less polished audio but offer more control over musical structure. These systems serve different purposes. Reporting each dimension separately is more informative than presenting one unexplained 'best model' score.

B. Human Judgement and Automated Metrics

Automated metrics can help identify technical defects or compare measurable properties, but music quality is contextual. A deliberately dissonant or noisy output may be appropriate for an experimental composition, while the same characteristics may be unacceptable in a calm meditation cue. Human listening should therefore be linked to task-specific criteria, and automated metrics should be described as proxies rather than complete measures of creativity or usefulness.

C. Data, Attribution, and Creator Interests

Training-data provenance and licensing are difficult to infer from the audio output alone. Research on artist perspectives has emphasized dataset transparency, opt-out possibilities, and communication between stakeholders . In practice, a model's documentation may be incomplete or change over time. The framework consequently treats transparency as a reportable dimension and encourages explicit statements of uncertainty instead of unsupported legal conclusions.

D. Limitations

1. The framework has not yet been validated through experiments or user studies.
2. Musical quality assessment may vary between listeners.
3. Evaluation results may differ across genres, cultures, and user groups.
4. Some AI music tools may not provide all the controls required for evaluation.
5. Training-data sources and licensing information may be incomplete.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

6. The framework cannot independently establish copyright compliance or infringement..

VII. FUTURE SCOPE

Future research can operationalize the framework as an open benchmark containing prompts, scoring rubrics, test tracks, and documentation-audit criteria. Studies should compare human ratings with automated measures and investigate whether ratings remain stable across listener backgrounds and musical traditions. Another direction is to evaluate provenance methods and synthetic-audio detection under realistic conditions, including modified or mixed human–AI recordings. Finally, longitudinal studies could examine whether AI music tools improve creative exploration or simply shift effort toward prompt iteration and post-editing.

VIII. CONCLUSION

AI music composition systems should be evaluated not only by how realistic their outputs sound, but also by how accurately they follow instructions, how well users can revise them, how effectively they fit real workflows, and how transparently they communicate responsible-use information. This paper proposes a five-dimension evaluation framework and a reproducible protocol intended for future empirical validation. The framework draws on published work in text-to-music generation, the supplied systematic review, and research on synthetic-music attribution and creator interests. Its central recommendation is to report the dimensions separately, document uncertainty, and avoid unsupported claims of universal superiority. Empirical studies using diverse users and tasks are required before the framework can be considered validated.

REFERENCES

[1] W. Yang, L. Shen, C.-F. Huang, J. Lee, and X. Zhao, “Development Status, Frontier Hotspots, and Technical Evaluations in the Field of AI Music Composition Since the 21st Century: A Systematic Review,” IEEE Access, vol. 12, pp. 89452–89476, 2024, doi: 10.1109/ACCESS.2024.3419050.

[2] A. Agostinelli et al., “MusicLM: Generating Music From Text,” arXiv:2301.11325, 2023. [Online]. Available: https://arxiv.org/abs/2301.11325.

[3] J. Copet et al., “Simple and Controllable Music Generation,” Advances in Neural Information Processing Systems, vol. 36, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/94b472a1842cd7c56dcb125fb2765fbd-Abstract-Conference.html.

[4] L. Comanducci, P. Bestagini, and S. Tubaro, “FakeMusicCaps: A Dataset for Detection and Attribution of Synthetic Music Generated via Text-to-Music Models,” Journal of Imaging, vol. 11, no. 7, art. 242, 2025, doi: 10.3390/jimaging11070242.

[5] L. Shroff, “AI & Copyright: A Case Study of the Music Industry,” GRACE: Global Review of AI Community Ethics, vol. 2, no. 1, 2024, doi: 10.60690/7pqxkr18.

[6] J. D. Fernandez and F. Vico, “AI methods in algorithmic composition: A comprehensive survey,” Journal of Artificial Intelligence Research, vol. 48, pp. 513–582, 2013, doi: 10.1613/jair.3908.

[7] C.-H. Liu and C.-K. Ting, “Computational intelligence in music composition: A survey,” IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 1, no. 1, pp. 2–15, 2017, doi: 10.1109/TETCI.2016.2642200.

[8] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, “MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018, doi: 10.1609/aaai.v32i1.11312.