

How Accurately and Creatively Can Generative AI Produce Music Across Genres?

A Mixed-Methods Evaluation of ACE-Step 1.5

Flora Wu

Abstract

Generative artificial intelligence music tools such as Suno, Udio, and ACE-Step can now create full songs with vocals, instrumentation, and production quality convincing enough to sound like human-made music. However, musical “quality” is not a single measurable property, and treating “AI music” as one category by itself risks missing details on how performance varies by genre. This study asks: how accurately and creatively can generative AI produce music across different genres? Two hundred tracks were generated using the open-source model ACE-Step 1.5 across ten genres, using five human-written lyric sets to isolate genre as the main variable. Each track was evaluated through two complementary methods: a computational audio analysis using Essentia's Discogs-trained genre classifier, and a human evaluation using a five-criterion rubric assessing musicality. Results suggest that AI-generated music performance is uneven and genre-dependent instead of uniform: genres that relied on standardized, rhythmic structures were classified and rated more consistently than genres that needed more sophisticated harmonies or improvisation. Genre-classification confidence and rubric-based creativity scores did not always align, which may mean the model can recreate surface-level genre traits without capturing the expressions and qualities associated with human musicianship. The study also brings a methodological finding to light: the genre classifier's own training taxonomy assigns subgenre labels unevenly across genres, which can distort genre-accuracy independent of the audio itself. Across 200 generated tracks and 50 human-scored tracks, human genre-match ratings ranged from 41.6% (Classical) to 91.3% (Metal), and Essentia's genre-classification confidence only weakly correlated with creativity scores from the rubric.

Introduction

Over the past few years, generative artificial intelligence has moved from producing synthetic-sounding music to producing full songs with vocals, instrumentation, and production quality that are comparable to human-made music. Tools such as Suno, Udio, and ACE-Step now allow anyone to type in a genre and a set of lyrics and receive a finished track within seconds. For that reason, this study narrows the general question into a more testable one: how accurately and creatively can AI generate music across different genres? This question separates "good" into two dimensions that can be measured separately: accuracy and creativity. Measuring both, genre by genre, gives a more complete picture than a single up or down judgment of "AI vs. human." This research question is explored through three sub-questions: (1) How well can generative AI produce music that matches a given genre and lyrics? (2) How well can generative AI translate or transfer music between genres? (3) Is generative AI equally capable across genres, or does it perform better in some than others?

Methods

This study used a mixed-methods design to assess the accuracy and creativity of AI-generated music across genres, combining quantitative computational audio analysis with qualitative human evaluation. Music samples were generated using ACE-Step 1.5 and evaluated through two independent, complementary processes: a self-coded Essentia pipeline that extracted audio features and genre-classification data, and a human evaluation conducted using a five-criterion scoring rubric. Findings from both methods were combined to provide a more complete picture of AI music quality than either method could offer alone. Ten genres were selected for this study: Pop, Hip-Hop & Rap, Classical, EDM (Electronic Dance Music), Rock, Latin, Country, R&B (Rhythm and Blues), Jazz, and Metal. For each genre, songs were generated using five human-written lyric sets from existing popular songs, with each lyric set paired with all ten genres. These five-lyric-set batches were regenerated across multiple sessions and resulted in a final total of 200 tracks across the ten genres (20 tracks per genre).

Results

Essentia's genre-confidence was computed across all 200 generated tracks, while human rubric scores were obtained for a randomly sampled, genre-stratified subset of 50 tracks. Averaged across all genres, the Essentia summed-metric confidence for all tracks was 29.8% and the maximum-metric confidence was 5.1% for their prompted genre. On the human side, average genre-match ratings across the sampled tracks ranged from 41.6% (Classical) to 91.3% (Metal), with an overall average of 75.2%. Average creativity scores from the rubric ranged from 2.36 (Hip-Hop & Rap) to 2.90 (Metal) out of 4, with an overall mean of 2.61. At the track level (50 tracks with both computational and human data), Essentia's summed confidence correlated moderately with human genre-match ratings (60%), and the maximum metric correlated a little more weakly with the same ratings (54%).

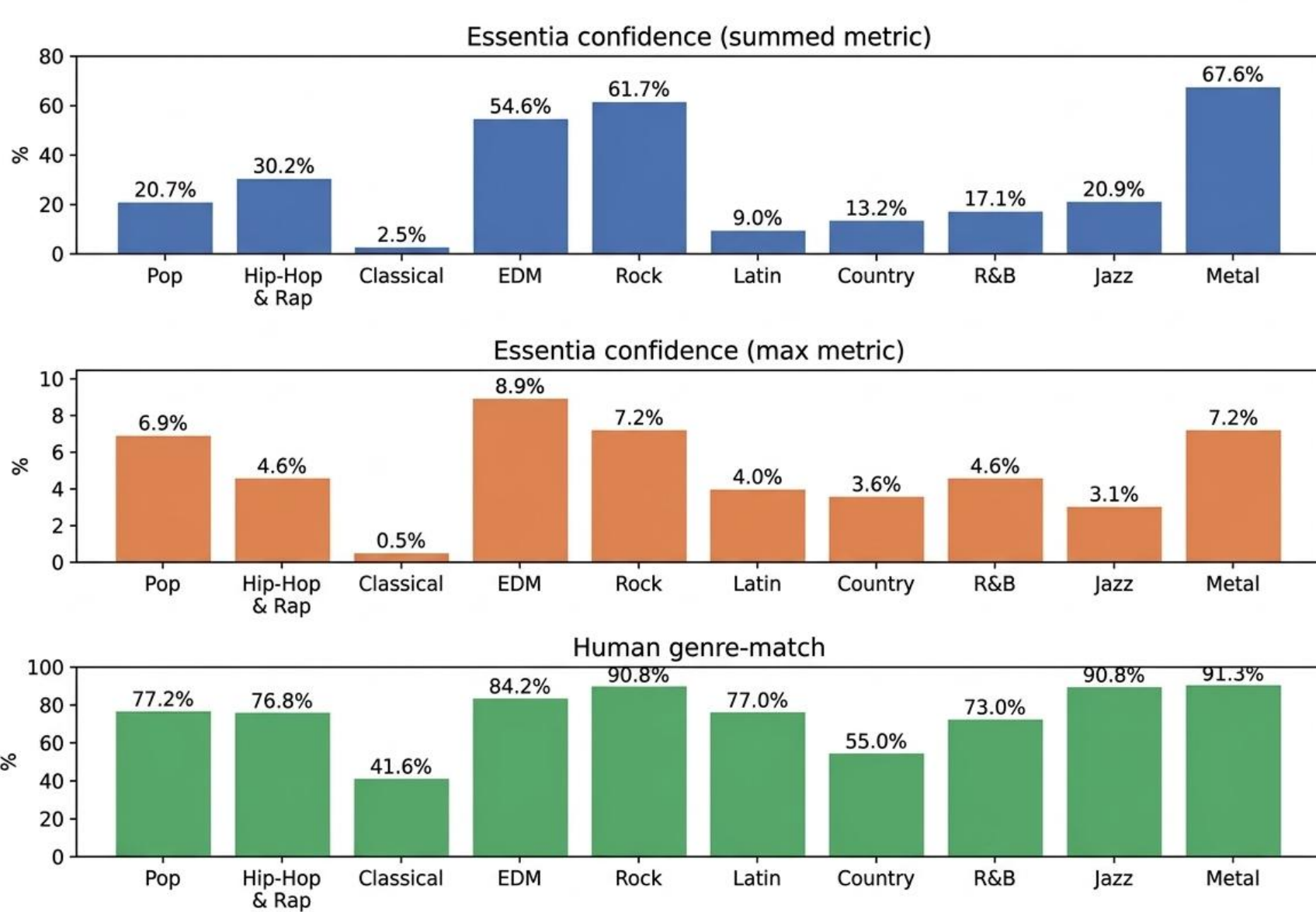


Figure 3. Essentia genre-classification confidence (summed and maximum metrics) and human-judged genre match, by genre.

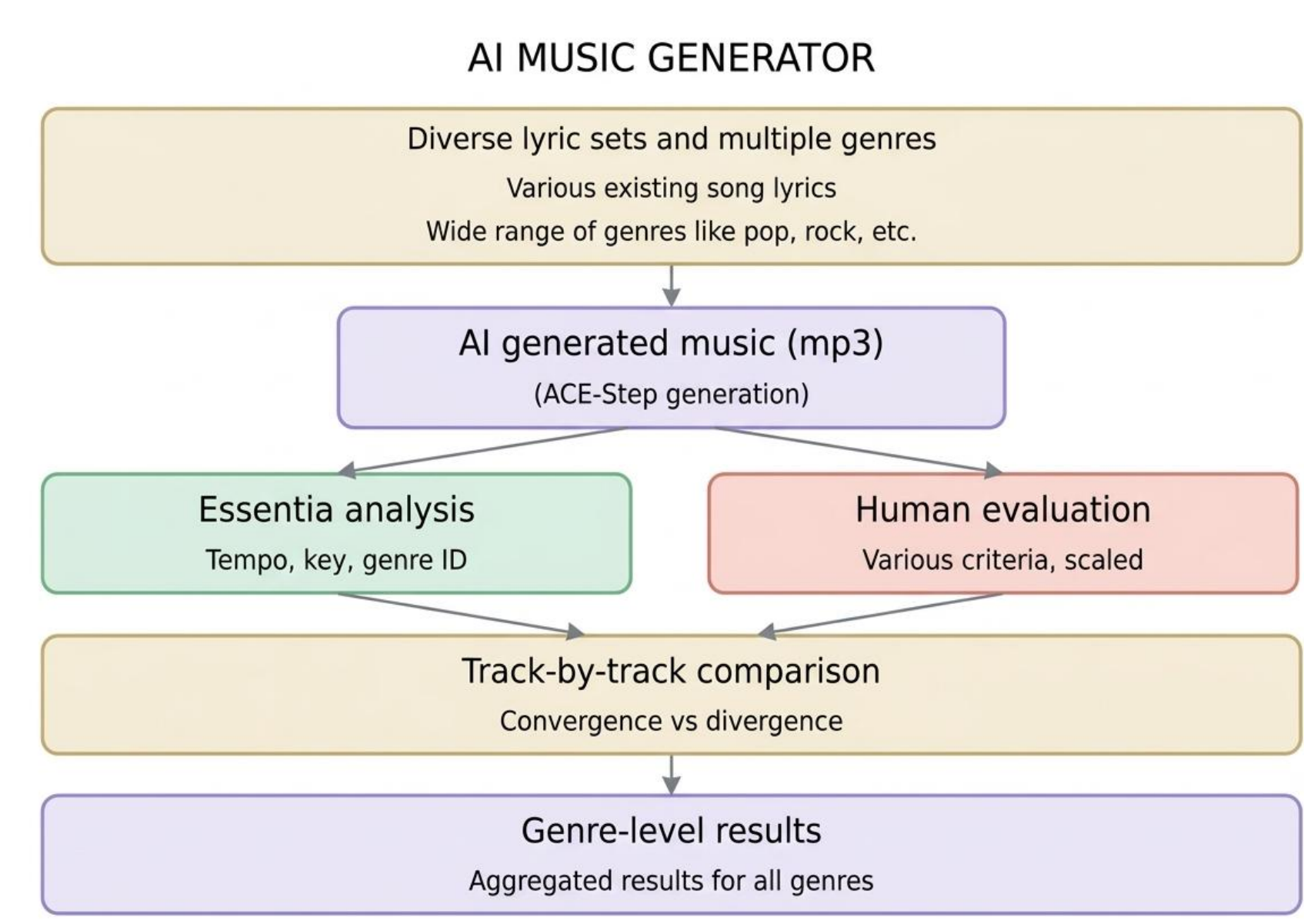


Figure 1. Overview of the study pipeline: generation, computational analysis, and human evaluation.

Discussion

The results show a mixed picture of ACE-Step 1.5's capabilities. For sub-question 1, the model was generally strong at matching lyrics to a target genre with clear, standardized sonic conventions, but weaker at the more expressive qualities, especially in genres that depend on natural harmonies or improvisation. For sub-question 2, the model showed a reasonably flexible ability to recreate with the same lyrical content across genres, but broke down at boundaries that overlap significantly in instrumentation or rhythm. For sub-question 3, performance was uneven across genres, which is evidence against treating "AI music" as a single, uniform category. The genres that scored highest share a common trait: they rely on repetitive, looped structures, clearly defined rhythm chunks, and production techniques that are heavily used in commercially available training data.

Conclusion

This study wanted to answer a specific question: how accurately and creatively can AI generate music across different genres? The answer, based on 200 generated tracks and 50 human-scored tracks spanning ten genres, is that performance is uneven and genre-dependent. The model reliably matched given lyrics to a target genre for standard and rhythmic genres but struggled on genres demanding more sophisticated harmonies or natural sounding improvisation. Its ability to translate the same musical idea across genres was strong at clearly defined genre boundaries but weak where genre styles overlapped. Critically, the model's genre accuracy and its rubric-judged musical quality have a weak correlation: several tracks that were technically recognizable as their intended genre still fell short on the more human qualities of harmony and emotion. As generative AI music tools continue to improve, this genre-by-genre, dimension-by-dimension view might prove more useful than any sweeping verdict for understanding which aspects of these tools are already strong collaborators, and where meaningful gaps between AI and human musicianship remain.

References

Gong et al. ACE-Step. arXiv:2506.00045 (2025).
Gong et al. ACE-Step 1.5. arXiv:2602.00744 (2026).
Bogdanov et al. ESSENTIA. ISMIR (2013). doi.org/177_Paper.pdf
Ipsos/Deezer. AI music study. Deezer Newsroom (2025).
Sun, R. Q. Can you tell AI music from human music? UCSD Guardian (2026).
Sarmiento et al. Between the AI and me... ISMIR (2024). arXiv:2407.21615.
Lecamwasam & Chaudhuri. Perceptions of AI music. arXiv:2506.02856 (2025).
Bogdanov et al. AcousticBrainz genre dataset. ISMIR (2019), 360-367.