



Audio Engineering Society Conference Paper 27

Presented at the AES International Conference on Machine
Learning and Artificial Intelligence for Audio
2025 September 8–10, London, UK

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (<http://www.aes.org/e-lib>), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Perceiving AI in Music: Human Evaluation of AI-Generated Melodies and AI Detection Sensitivity

Michael Oehler¹, Jasper Oldach¹, and Florian Zwißler²

¹Osnabrück University, Music Technology and Digital Musicology Lab

²Cologne University of Music and Dance

Correspondence should be addressed to Michael Oehler (michael.oehler@uos.de)

ABSTRACT

Recent advancements in generative artificial intelligence (AI) have led to increasingly sophisticated music composition systems, yet their perceived creative quality and detectability remain debated. This study examines human perception of AI-generated music across two paradigms: a melody continuation task, where AI- and human-generated melodies were rated on four aesthetic dimensions, and an AI detection task, where participants distinguished between AI-composed and human-composed songs. Results indicate that AI-generated melodies were consistently rated lower than human compositions, with no significant differences between AI models or the impact of expert feedback. In the detection task, participants identified AI-generated songs at a level above chance, though sensitivity remained low. Notably, musical sophistication, as measured by the Gold-MSI, did not significantly influence ratings or detection accuracy. These findings highlight persistent limitations in AI-generated music while raising questions about the perceptual boundaries between AI and human creativity.

1 Introduction

Recent advancements in generative artificial intelligence (AI) have significantly impacted creative domains, notably musical composition [1]. AI systems, including ChatGPT and Google Magenta, have demonstrated the ability to generate musical content; however, empirical evaluations of their creative capabilities relative to human composers remain sparse. Even recent comprehensive comparative evaluations of AI music-generation systems, such as Yin et al. [2], did not include general-purpose AI models like ChatGPT, focusing instead exclusively on specialized deep-learning-based music generation models. This gap underscores the necessity of systematically assessing

general-purpose AI agents, particularly given their rapid development and broad availability.

Schreiber et al. recently conducted a seminal study assessing AI-generated melodies from ChatGPT and Google Magenta against human-created melodies by music students within a standardized melody continuation paradigm [3]. They found the AI compositions to be significantly inferior in aesthetic quality.

The present research builds on the methodology of [3] with updated AI agents, as the rapid evolution of AI technologies necessitates continuous empirical reassessment. Notably, Google Magenta has seen limited recent developments, whereas newer generative systems like Google Gemini and advanced reasoning

models from ChatGPT (e.g., GPT-4 and later variants) have emerged, offering potentially enhanced creative capabilities relevant for music generation.

1.1 Background and Related Work

A growing body of perceptual research has examined listeners' abilities to detect and evaluate AI-generated music. For instance, Frieler and Zaddach asked jazz experts to rate human-performed solos against algorithmically composed counterparts, finding a clear preference for human performances and identification accuracy around 65 % [4]. Extending this paradigm to a specialized sub-genre, Sarmento et al. showed that progressive-metal guitarists could reliably distinguish AI-generated guitar tablature from human-composed examples, again preferring the latter in subjective ratings [5]. Complementing these discrimination studies, White et al. and Shank et al. revealed a pronounced labeling bias: when identical excerpts are presented as "AI-generated," listeners rate them significantly lower in expressive and aesthetic quality than when labeled "human" [6, 7].

Parallel to these perceptual inquiries, the landscape of generative AI in music has bifurcated into symbolic-domain and audio-domain approaches. Symbolic systems—from early rule-based composition engines [8] to modern neural architectures such as PerformanceRNN [9] and MusicTransformer [10]—produce MIDI sequences that are subsequently rendered via synthesis engines or neural vocoders. By contrast, audio-domain models like Jukebox [11], MusicLM [12], MusicGen [13], MusicLDM [14], Mustango [15], Riffusion [16], and Suno AI [17] operate directly on waveform or spectrogram representations, often leveraging discrete audio tokens or diffusion in a learned latent space. Seed-Music has recently proposed a hybrid pipeline combining autoregressive language modeling with diffusion-based audio refinement, illustrating how high-level symbolic control and fine-grained timbral editing can be unified [18].

Comparative evaluations within the symbolic domain have largely focused on specialist architectures: Yin et al. systematically benchmarked models such as MusicTransformer and CoCoNet in a melody-continuation task [2]. More recently, Schreiber et al. extended this to general-purpose language models—ChatGPT and Google Magenta—finding that both lag behind human composers in aesthetic ratings, with ChatGPT's

continuations rated marginally higher than those from Magenta [3].

Building on these insights, our study bridges the symbolic and audio domains. We first evaluate state-of-the-art autoregressive transformer models (GPT-4, Gemini) in a standardized melody-continuation paradigm. We then assess a prototypical audio-domain diffusion system (Suno AI), which synthesizes rich timbral textures via latent diffusion over spectrogram embeddings, in a signal-detection task to measure listeners' sensitivity to AI generation.

1.2 Aims and Hypotheses

This dual-paradigm evaluation addresses a critical gap by examining both symbolic and audio-based generative systems through the lens of human perception. Unlike prior technical detection approaches [19], our study emphasizes listeners' subjective experiences and qualitative judgments. Based on prior research on AI-generated music perception and human-AI comparisons in musical creativity, the present study formulates the following hypotheses:

H1: AI-generated melodies will receive significantly lower ratings than human-composed melodies across several aesthetic dimensions. This hypothesis aligns with previous studies showing that AI-composed music is often perceived as less expressive and coherent compared to human compositions [3].

H2: AI-generated melodies that incorporate expert feedback (reprompts) will receive higher ratings than those without feedback. Given that AI models are designed to optimize outputs based on iterative refinements, we expect that structured feedback from human experts will improve the perceived quality of AI compositions.

H3: Participants will identify AI-generated songs at a level above chance, indicating some level of sensitivity to distinguishing AI-composed from human-composed music. This expectation follows from previous research showing that listeners can differentiate between AI and human music, albeit with moderate accuracy [4, 7, 5].

H4: Musical sophistication will correlate with AI detection accuracy, such that individuals with higher musical sophistication will demonstrate greater sensitivity in distinguishing AI-generated from human-composed music.

H5: AI-generated songs containing vocals will be more easily identified as AI-generated compared to instrumental AI-generated music. This hypothesis is derived from research suggesting that listeners are particularly sensitive to unnatural renditions of human voice [20, 21, 22].

2 Methods

A listening experiment was conducted in two parts. In the first part, participants qualitatively evaluated stimuli generated using the melody continuation task. The second part utilized a Signal Detection Theory (SDT) paradigm [23, 24], in which listeners decided for each stimulus whether it was AI-generated or human-composed. Additionally, musical sophistication was assessed using the Gold-MSI [25] to examine whether participants' judgments were influenced by their level of musical sophistication.

2.1 Melody continuation task

Following [3], we selected a distinct seven-bar melodic fragment (3/4 meter, 80 bpm) as seed prompt. Eight graduate music students each produced one "human-composed" continuation of this fragment. The same students then prompted two general-purpose AI agents—GPT-4o and Gemini 2.0—to generate continuations under identical instructions, and finally they provided up to three iterative reprompts (expert feedback) to each AI output in order to refine it. Students provided feedback to the AI in the same manner as they would during a live co-composing session, offering guidance on musical shape rather than prescribing exact pitches. For instance, one might say: "Try strengthening the melody in bar 7 and resolve the phrase more harmonically at the end." Note that this feedback deliberately avoided specifying exact notes or sequences, focusing instead on the overall contour and expressive goals.

Unlike specialized music-generation systems (e.g., MuseNet, MusicLM, MusicGen, Seed-Music), which work on raw audio or opaque MIDI representations via hierarchical embeddings and diffusion/transformer architectures, both GPT-4o and Gemini accept symbolic pitch–duration sequences expressed in natural-language prompts and return continuations in the same format. This makes them ideally suited for a standardized, symbolic melody-continuation paradigm.

All continuations—whether human- or AI-generated—adhered to a common set of compositional guidelines: a film-music style, between 10 and 20 notes, varied note durations, a limited pitch range (D3–D5), and a clearly defined melodic climax. By comparing outputs with and without expert feedback, our factorial design yielded 40 stimuli in total:

- 8 human-composed stimuli (HUM)
- 8 GPT-4o stimuli without feedback (GPT_nF)
- 8 GPT-4o stimuli with feedback (GPT_wF)
- 8 Gemini stimuli without feedback (GEM_nF)
- 8 Gemini stimuli with feedback (GEM_wF)

All continuations were encoded as MusicXML, rendered using the oboe timbre from the MuseScore SF3 soundfont, and normalized for loudness to ensure consistency.

In the subsequent listening study, participants rated the resulting melodies on several aesthetic dimensions, specifically focusing on how musically logical and meaningful, interesting, convincing, and personally likable they were, using 5-point Likert scales.

2.2 AI Detection Task

This part of the study follows a more exploratory approach, as it is challenging to balance the use of commercially popular AI music generation systems with the need for strict control over generation parameters. Many widely used AI agents offer limited customization options, making it difficult to systematically manipulate and standardize key musical aspects across different conditions. At the same time, highly controllable research-based AI models often fail to match the naturalness and production quality of commercial systems. Given these constraints, we opted for a commercial AI platform that represents the current state-of-the-art in AI-generated music [26], ensuring the highest level of realism and fidelity while allowing us to study the impact of such technology in real-world usage scenarios.

Eight music students were tasked with using Suno's AI music generation platform [17] to create three musical pieces in a genre/style of their choice, aiming for a particularly natural sound. The specific requirement was to produce one piece featuring male vocals,

one with female vocals, and one instrumental. This approach stems from the understanding that listeners are particularly sensitive to unnatural renditions of human voice, as indicated by several studies [20, 21, 22]. Suno was selected for this task due to its recent Version 4 release, which has been noted for producing more natural-sounding music, especially in vocal performances.

To maximize the perceived naturalness of the AI-generated pieces, music students were allowed to refine their compositions iteratively, adjusting generation parameters until they closely matched their artistic expectations. Additionally, participants were asked to select lesser-known, non-AI-composed songs that closely matched their AI-generated pieces in terms of genre/style, instrumentation etc. This selection aimed to facilitate a comparative analysis between AI-generated and human-composed music.

From the total of 48 musical pieces, comparable 30-second excerpts were extracted, containing characteristic elements such as verse and chorus sections. All excerpts were loudness-normalized to ensure consistent listening conditions.

2.3 Participants

An a priori sample size calculation was conducted using G*Power [27] to ensure sufficient statistical power for the planned repeated-measures ANOVA. While Schreiber et al. [3] found a large effect size in their study, we opted for a more conservative approach due to advancements in AI agents. Specifically, we based our calculation on a small effect size ($f = 0.2$) to account for potential differences in perceptual sensitivity. With a power of 0.95 and a significance level of 0.05, this resulted in a required sample size of 48 participants. To compensate for potential exclusions due to incomplete data, we recruited 60 participants. After data screening, two participants had to be excluded due to missing or incomplete responses, leaving a final sample of 58 participants for analysis.

The final sample consisted of 37 female, 20 male, and 1 non-binary participant, with a mean age of 28.93 years ($SD = 14.68$). To assess participants' overall level of musical sophistication, the general factor of the Gold-MSI was measured, yielding an average score of 77.5, with values ranging from 33 to 112.

Written informed consent was obtained from all participants prior to the experiment. Participants were not paid an honorarium but could withdraw at any time without penalty. All procedures were conducted in accordance with the Declaration of Helsinki and the guidelines of the German Psychological Society.

2.4 Procedure

Due to the relatively large number of stimuli used in both parts of the study, an incomplete within-subjects design was implemented. Each participant was randomly assigned stimuli produced by one of the eight music students, evaluating five melodies and six musical pieces via an online study conducted using SoSciSurvey [28]. This approach ensured direct comparability between AI-generated and human-composed stimuli while maintaining a manageable workload for participants. Consequently, stimuli created by each music student were evaluated by approximately seven to eight participants.

Each participant first evaluated the melodies, presented in random order, by rating them on four dimensions—musically logical and meaningful, interesting, convincing, and personal liking—using a 5-point Likert scale. Subsequently, the six musical pieces were also presented in randomized order. After listening to each piece, participants indicated whether they believed it was composed by AI or a human composer. Additionally, they reported whether they recognized the song; trials involving recognized songs were excluded from further analysis.

At the end of the experiment, participants provided demographic information, and the general factor of the Gold-MSI was assessed to measure their overall level of musical sophistication.

3 Results

3.1 Melody continuation task

3.1.1 Analysis of Variance (ANOVA)

Before each repeated-measures ANOVA, we assessed the assumption of sphericity using Mauchly's test. Results were as follows: for interesting, $W(9) = 0.754$, $p = 0.082$; for logical and meaningful, $W(9) = 0.850$, $p = 0.454$; for convincing, $W(9) = 0.948$, $p = 0.968$; and for liking, $W(9) = 0.934$, $p = 0.930$. In all cases,

$p \geq 0.05$, indicating that the sphericity assumption held and that uncorrected degrees of freedom could be used.

A series of repeated-measures ANOVAs were then conducted to examine whether the ratings of musical stimuli differed across the five conditions. The results revealed significant main effects of category for all four dependent variables (interesting, logical and meaningful, convincing, and liking). The effect sizes (Partial Eta-Squared) ranged from 0.098 to 0.148, indicating small to moderate effects (see Table 1).

For interesting, a significant effect of category was found, $F(4, 224) = 8.85, p < 0.001, \eta_p^2 = 0.1365$, suggesting that the perceived interestingness of the melodies differed significantly between groups. Similarly, significant main effects were observed for logical and meaningful ($F(4, 224) = 6.09, p < 0.001, \eta_p^2 = 0.0981$), convincing ($F(4, 224) = 6.89, p < 0.001, \eta_p^2 = 0.1096$), and liking ($F(4, 224) = 9.76, p < 0.001, \eta_p^2 = 0.1485$), indicating systematic differences in how the musical continuations were evaluated across conditions.

Table 1: Repeated Measures ANOVA Results

DV	$F(4, 224)$	p	η_p^2
<i>Interesting</i>	8.85	< 0.001	0.1365
<i>Logical</i>	6.09	< 0.001	0.0981
<i>Convincing</i>	6.89	< 0.001	0.1096
<i>Liking</i>	9.76	< 0.001	0.1485

3.1.2 Multiple Regression with Musical Sophistication

To examine whether musical sophistication (Gold-MSI) influenced ratings, multiple regression analyses were conducted with category, Gold-MSI, and their interaction as predictors. The models revealed relatively low explanatory power (R^2 ranging from 0.087 to 0.124), suggesting that only a small portion of the variance in ratings could be attributed to these factors. Importantly, Gold-MSI was not a significant predictor in any model ($p > 0.1$), indicating that individual musical sophistication did not systematically influence the evaluations of the melodic continuations. Moreover, no significant interactions between Gold-MSI and category were observed, suggesting that the effect of musical sophistication did not vary across AI- and human-generated melodies.

3.1.3 Post-hoc Comparisons

Bonferroni-corrected post-hoc paired t-tests were conducted to further explore differences between the conditions. Across all four dependent variables, a consistent pattern emerged: significant differences were found between AI-generated melodies (GPT and GEM) and human-generated melodies (HUM), while AI models did not significantly differ from each other. These results are illustrated in Figure 1 and detailed in Table 2, which also report Cohen's d as a measure of effect size.

Table 2: Post-hoc Comparisons Across Dimensions

	t	p	d
<i>Interesting</i>			
GPT_nF vs. HUM	-5.04	< .001	-0.67
GPT_wF vs. HUM	-2.93	.049	-0.39
GEM_nF vs. HUM	-4.22	< .001	-0.56
GEM_wF vs. HUM	-4.93	< .001	-0.65
<i>Logical and Meaningful</i>			
GPT_nF vs. HUM	-4.16	.001	-0.55
GPT_wF vs. HUM	-3.72	.004	-0.49
GEM_wF vs. HUM	-4.67	< .001	-0.62
<i>Convincing</i>			
GPT_nF vs. HUM	-3.10	.030	-0.41
GPT_wF vs. HUM	-3.48	.009	-0.46
GEM_nF vs. HUM	-3.20	.022	-0.42
GEM_wF vs. HUM	-4.87	< .001	-0.65
<i>Liking</i>			
GPT_nF vs. HUM	-5.10	< .001	-0.68
GPT_wF vs. HUM	-4.14	.001	-0.55
GEM_nF vs. HUM	-4.97	< .001	-0.66
GEM_wF vs. HUM	-5.01	< .001	-0.66

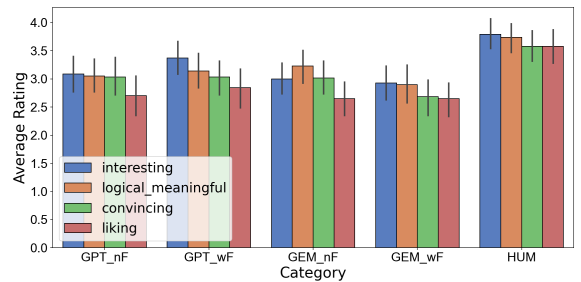


Fig. 1: Comparison of ratings across categories and dependent variables. The rating scale ranged from 1 (not at all) to 5 (very much). Error bars represent 95% confidence intervals.

3.2 AI Detection Task

For each participant, a d' value was calculated based on the Signal Detection Theory (SDT) paradigm [23, 24] to measure their sensitivity in distinguishing AI-generated from human-composed musical pieces ($d' = Z(\text{hit rate}) - Z(\text{false alarm rate})$). The average d' value across all participants was $M = 0.711$ ($SD = 1.212$), indicating that participants were able to identify Suno-generated songs at a level above chance. However, the sensitivity was low, suggesting that differentiating AI-generated from human-composed music remained challenging.

To investigate whether musical sophistication, as measured by the Gold-MSI, is associated with the ability to distinguish AI-generated from human-composed music, several analyses were conducted.

A median split on the Gold-MSI scores divided participants into high and low musical sophistication groups. A two-sample t -test comparing their sensitivity (d') values revealed no significant difference between groups, $t(56) = 0.471$, $p = .640$, Cohen's $d = 0.124$. The average d' value for the high musical sophistication group was $M = 0.787$ ($SD = 1.354$), while the low musical sophistication group had an average d' of $M = 0.636$ ($SD = 1.069$).

Figure 2 shows a Kernel Density Estimation (KDE) plot illustrating the distribution of d' values for both high and low musical sophistication groups. The left-skewed distribution observed in the low Gold-MSI group suggests that while most participants performed similarly to the high Gold-MSI group, a subset exhibited notably lower sensitivity (d'), indicating greater individual variability in distinguishing between AI-generated and human-composed music.

Additionally, a Pearson correlation analysis between Gold-MSI scores and d' values yielded $r = 0.119$, $p = .375$, indicating no significant linear relationship. This further supports the notion that higher musical sophistication does not necessarily correspond to greater sensitivity in identifying AI-generated music.

To examine whether sensitivity (d') in distinguishing AI-generated from human-composed music varied across the three song variants (male vocals, female vocals, and instrumental), a series of chi-square tests were conducted. The analysis compared the detection rates of AI-generated (Suno) and human-composed (HUM)

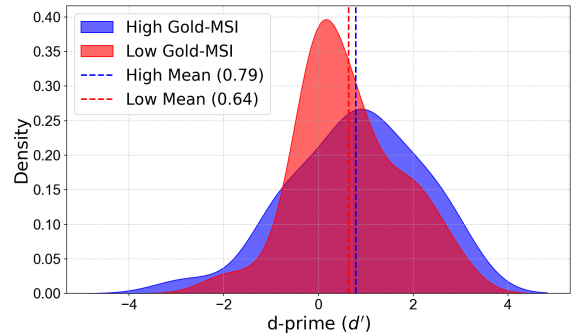


Fig. 2: Kernel Density Estimation (KDE) plot for d' by Gold-MSI group.

songs across these categories, with Cramér's V calculated as a measure of effect size.

The results indicated that none of the comparisons reached statistical significance (all p -values $> .05$), suggesting participants were not systematically more or less accurate in identifying AI-generated music based on the presence or type of vocals. Across all conditions, effect sizes remained small (Cramér's $V < 0.3$), further reinforcing the conclusion that the presence of male vocals, female vocals, or the absence of vocals did not substantially impact participants' ability to distinguish AI-generated from human-composed songs.

4 Discussion

The present study aimed to assess the perceived quality and detectability of AI-generated music compared to human-composed music across two experimental paradigms: a melody continuation task and an AI detection task.

4.1 Evaluation of AI-Generated Melodies

The results of the melody continuation task indicate that AI-generated melodies were consistently rated lower than human-composed melodies across all four aesthetic dimensions (interesting, logical and meaningful, convincing, and liking). This pattern replicates previous findings by Schreiber et al. [3], reinforcing the notion that, despite recent advancements, AI-generated melodies still exhibit qualitative shortcomings when compared to human compositions.

Importantly, the presence of iterative expert feedback during AI generation (GPT_wF, GEM_wF) did not

lead to significantly higher ratings compared to their respective no-feedback conditions (GPT_nF, GEM_nF). This suggests that while AI systems can integrate feedback in text-based tasks, their ability to refine creative outputs in music generation may be more limited. Unlike human composers, who can abstract and creatively reinterpret feedback, AI systems may struggle to make meaningful structural and stylistic improvements beyond surface-level modifications.

Furthermore, post-hoc comparisons revealed that while significant differences existed between AI-generated and human-composed melodies, no significant differences were found between the different AI models or between their with- and without-feedback conditions. This suggests that the type of AI system (GPT vs. GEM) had little impact on perceived quality, and that feedback did not enhance compositional effectiveness in the present study.

Musical sophistication, as measured by the Gold-MSI, did not significantly influence ratings of AI- or human-generated melodies, nor did it moderate the effect of category. Consistent with the findings of [3], this suggests that both musically trained and untrained listeners evaluate AI-generated music similarly, reflecting a shared perception of its aesthetic limitations.

4.2 Detection of AI-Generated Music

The results from the AI detection task revealed that participants were able to distinguish AI-generated from human-composed songs at a level above chance ($d' = 0.711$), yet their sensitivity remained relatively low. This aligns with prior studies showing that while AI-generated music can be recognized with some accuracy, it often approximates human-like compositional characteristics closely enough to make identification difficult [4, 7, 5].

No significant differences were found between the ability to detect AI-generated music based on whether the song contained male vocals, female vocals, or no vocals. This finding contradicts the expectation that participants would be particularly sensitive to unnatural vocal rendering, as suggested by previous research on voice perception [20, 21, 22]. While it remains possible that AI-generated vocals introduce subtle unnatural characteristics, these were not strong enough to consistently impact detection accuracy.

Furthermore, musical sophistication did not significantly correlate with sensitivity in AI detection, and no differences were found between participants with high and low Gold-MSI scores. This suggests that experience and expertise in music do not necessarily enhance the ability to identify AI-generated music, potentially because the distinctions between AI and human compositions are not solely based on musical complexity but also on production quality and stylistic coherence.

5 Conclusions

While this study provides insights into the perception and detection of AI-generated music, several limitations should be acknowledged.

First, the study focused on a relatively small set of AI-generated and human-composed melodies and songs, limiting the generalizability of the findings. Although the stimuli were carefully selected and controlled, future studies should expand the dataset to include a broader range of musical styles, complexity levels, and AI generation techniques to assess whether the observed effects hold across different musical contexts.

Second, the AI models used in the melody continuation task (GPT-4o and Gemini 2.0) were not originally designed for music composition, but rather adapted for this purpose. While these models have shown promise in generating musical structures, their outputs may not fully reflect the capabilities of dedicated music AI systems such as ByteDance’s Seed-Music, OpenAI’s MuseNet, Google’s MusicLM, or Suno’s commercial platform. A direct comparison between general-purpose AI models and music-specific AI systems could provide a more comprehensive understanding of their respective strengths and weaknesses.

Third, the effectiveness of expert feedback on AI-generated melodies may have been constrained by the structure of the feedback process itself. Music students provided iterative feedback to AI models, but the way in which AI systems processed and incorporated this feedback remains unclear. It is possible that the current generation of AI models lacks the capacity to meaningfully refine compositions in response to nuanced musical critique. A detailed log of reprompt messages, coupled with an automated evaluation of melodic features (e.g., comparing pitch histograms or harmonic transitions before vs. after feedback), would help quantify the actual impact of expert guidance. Future studies could also explore alternative interaction paradigms,

such as real-time co-creative systems where AI and human composers iteratively develop music together, rather than AI simply responding to written prompts.

Fourth, in the AI detection task, only one commercial AI system (Suno) was tested. While Suno represents a state-of-the-art model in AI-generated music [26], the extent to which the findings generalize to other AI music generators remains uncertain. Different AI systems may have distinct sonic characteristics that influence detectability, and the use of a single system limits the scope of conclusions about AI music detection in general.

Finally, the study employed an incomplete within-subjects design, meaning that each participant evaluated only a subset of the available stimuli. While this approach was necessary to manage cognitive load and ensure comparability, it may have introduced variability in participants' exposure to different compositional styles. Future research could explore fully within-subjects designs or counterbalanced approaches to enhance the robustness of the findings.

Addressing these limitations in future research will be essential for developing a more comprehensive understanding of AI-generated music, its creative potential, and how human listeners perceive and interact with it.

References

- [1] Mitra, R. and Zualkernan, I., "Music Generation Using Deep Learning and Generative AI: A Systematic Review," *IEEE Access*, 2025.
- [2] Yin, Z., Reuben, F., Stepney, S., and Collins, T., "Deep learning's shallow gains: a comparative evaluation of algorithms for automatic music generation," *Machine Learning*, 112(5), pp. 1785–1822, 2023.
- [3] Schreiber, A., Sander, K., Kopiez, R., and Thöne, R., "The creative performance of the AI agents ChatGPT and Google Magenta compared to human-based solutions in a standardized melody continuation task," *Jahrbuch Musikpsychologie*, 32, p. e195, 2024, doi:10.5964/jbdgm.195.
- [4] Frieler, K. and Zaddach, W.-G., "Evaluating an analysis-by-synthesis model for Jazz improvisation," *Transactions of the International Society for Music Information Retrieval*, 5(1), pp. 20–34, 2022.
- [5] Sarmiento, P., Loth, J., and Barthet, M., "Between the AI and Me: Analysing Listeners' Perspectives on AI-and Human-Composed Progressive Metal Music," *arXiv preprint arXiv:2407.21615*, 2024.
- [6] White, C. W., Kapoor, K., Cosme-Clifford, N., Symons, J., and von Mutius, L., "Humans Perceive Ai-Generated Music as Less Expressive than Comparable Human-Made Content," *Available at SSRN 5087035*, 2025.
- [7] Shank, D. B., Stefanik, C., Stuhlsatz, C., Kacirek, K., and Belfi, A. M., "AI composer bias: Listeners like music less when they think it was composed by an AI," *Journal of Experimental Psychology: Applied*, 29(3), pp. 676–692, 2023, doi:10.1037/xap0000447.
- [8] Cope, D., "Experiments in musical intelligence (EMI): Non-linear linguistic-based composition," *Journal of New Music Research*, 18(1-2), pp. 117–139, 1989.
- [9] Simon, I. and Oore, S., "Performance RNN: Generating Music with Expressive Timing and Dynamics," <https://magenta.tensorflow.org/performance-rnn>, 2017.
- [10] Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A. M., Hoffman, M. D., Dinculescu, M., and Eck, D., "Music Transformer: Generating Music with Long-Term Structure," in *7th International Conference on Learning Representations (ICLR)*, 2019.
- [11] Dhariwal, P., Jun, H., Payne, C., et al., "Jukebox: A Generative Model for Music," 2020, arXiv preprint.
- [12] Agostinelli, A. et al., "MusicLM: Generating Music From Text," 2023, arXiv preprint.
- [13] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A., "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, 36, pp. 47704–47720, 2023.
- [14] Chen, K., Wu, Y., Liu, H., Nezhurina, M., Berg-Kirkpatrick, T., and Dubnov, S., "Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies," in *ICASSP 2024*, pp. 1206–1210, IEEE, 2024.

- [15] Melechovsky, J., Guo, Z., Ghosal, D., Majumder, N., Herremans, D., and Poria, S., “Mustango: Toward controllable text-to-music generation,” *arXiv preprint arXiv:2311.08355*, 2023.
- [16] Riffusion, “Riffusion AI,” 2023, accessed: 2025-03-13.
- [17] Suno AI, “Suno AI Audio Generation,” 2023, accessed: 2025-03-13.
- [18] Bai, Y., Chen, H., Chen, J., Chen, Z., Deng, Y., Dong, X., Hantrakul, L., Hao, W., Huang, Q., Huang, Z., et al., “Seed-music: A unified framework for high quality and controlled music generation,” *arXiv preprint arXiv:2409.09214*, 2024.
- [19] Afchar, D., Meseguer-Brocal, G., and Hennequin, R., “AI-Generated Music Detection and its Challenges,” *arXiv preprint arXiv:2501.10111*, 2025.
- [20] Nussbaum, C., Frühholz, S., and Schweinberger, S. R., “Understanding voice naturalness,” *Trends in Cognitive Sciences*, 2025.
- [21] Pandey, A., Edlund, J., Le Maguer, S., and Harte, N., “Listener sensitivity to deviating obstruents in WaveNet,” *Proc. Interspeech. Dublin, Ireland*, pp. 1080–1084, 2023.
- [22] Kühne, K., Fischer, M. H., and Zhou, Y., “The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study,” *Frontiers in neurorobotics*, 14, p. 593732, 2020.
- [23] Thurstone, L. L., “A law of comparative judgment,” *Psychological review*, 34(4), pp. 27–286, 1927.
- [24] Green, D. M., Swets, J. A., et al., *Signal detection theory and psychophysics*, volume 1, Wiley New York, 1966.
- [25] Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L., “The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population,” *PLoS ONE*, 9(2), p. e89642, 2014.
- [26] Grötschla, F., Solak, A., Lanzendörfer, L. A., and Wattenhofer, R., “Benchmarking Music Generation Models and Metrics via Human Preference Studies,” in *ICASSP 2025*, pp. 1–5, IEEE, 2025.
- [27] Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A., “G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences,” *Behavior Research Methods*, 39(2), pp. 175–191, 2007, doi:10.3758/BF03193146.
- [28] Leiner, D. J., “SoSci Survey: The Solution for Online Surveys,” <https://www.soscisurvey.de>, 2025, accessed: 2025-03-13.