Jahrbuch Musikpsychologie
Yearbook of Music Psychology

PsychOpen GOLD

Research Reports

The Creative Musical Achievement of AI Systems Compared to Music Students: A Replication of the Study by Schreiber et al. (2024)

Die kreativen musikalischen Leistungen von KI-Systemen im Vergleich zu Musikstudierenden: Eine Replikation der Studie von Schreiber et al. (2024)

Nicholas Meier[1], Kilian Sander[1], Anton Schreiber[1], Reinhard Kopiez[1]

[1] *Hanover University of Music, Drama and Media, Hannover, Germany.*

Jahrbuch Musikpsychologie, 2025, Vol. 33, Article e221, https://doi.org/10.5964/jbdgm.221

Received: 2025-03-26 • **Accepted:** 2025-06-14 • **Published (VoR):** 2025-07-16

Reviewed by: Wolf-Georg Zaddach; Kathrin Schlemmer.

Corresponding Author: Reinhard Kopiez, Hanover Music Lab, Hanover University of Music, Drama and Media, Neues Haus 1, 30175 Hannover, Germany. E-mail: reinhard.kopiez@hmtm-hannover.de

Supplementary Materials: Code, Data, Materials, Preregistration [see Index of Supplementary Materials]

Abstract

Although the last two years have seen AI systems progress significantly when it comes to generating cultural products like literature, poems, or music, the jury is still out when it comes to determining whether the aesthetic quality of these products increases in tandem with the performance enhancements of underlying large language models (LLMs). We replicated the study by Schreiber et al. (2024) to test whether the creative performance of selected LLMs had improved over the past two years in the musical domain. In an online rating experiment based on a melody continuation paradigm, 75 melodic continuations generated by the AI systems *Qwen 2* (Version 72B Instruct), *Llama 3* (Version 70B Instruct), and *ChatGPT* (Version 4) were compared to 23 solutions composed by humans. The aesthetic quality of the sound examples was then evaluated by $N = 54$ listeners (music students) using four criteria (convincing, logical and meaningful, interesting, and liking). As the first main finding, human-based creative solutions outperformed all three AI systems on all four dependent variables (large effect sizes $1.11 \leq d_z \leq 2.51$), thus confirming the finding by Schreiber et al. (2024). The second main finding revealed a mean (and meaningful) discrimination sensitivity of $d' = 1.09$ for AI- and human-based solutions. We conclude that merely boosting the volume of training of the AI systems does not guarantee correlating improvement in the creative musical output produced under controlled conditions.

Keywords

Artificial Intelligence, AI, generative AI, composition, empirical aesthetics, melody rating, musical creativity, large language models

Zusammenfassung

Obwohl KI-Systeme in den letzten Jahren erhebliche Fortschritte bei der Erzeugung kultureller Produkte wie Literatur, Poesie oder Musik gemacht haben, bleibt die Frage offen, ob die ästhetische Qualität dieser Produkte mit der zunehmenden allgemeinen Leistungsfähigkeit der large language models (LLMs) ebenfalls angewachsen ist. In einer Replikation der Studie von Schreiber et al. (2024), überprüften wir, ob die kreative Leistungsfähigkeit ausgewählter LLMs auf dem Gebiet der Musik zugenommen hat. In einem Online-Rating-Experiment und unter Verwendung eines Melodiefortsetzungsparadigmas wurden 75 Melodiefortsetzungen der KI-Systeme *Qwen 2* (Version 72B Instruct), *Llama 3* (Version 70B Instruct) und *ChatGPT* (Version 4) mit 23 Fortsetzungsvarianten von Musikstudierenden verglichen. Die ästhetische Qualität der Fortsetzungen wurde von $N = 54$ Hörer*innen (Musikstudierende) mittels vier Items (überzeugend, logisch und sinnvoll, interessant, Gefallen) erfasst. Als erstes Hauptergebnis wurden die menschlichen Lösungen auf allen vier Bewertungsmerkmalen besser beurteilt als die KI-Lösungen (große Effektgröße $1.11 \leq d_z \leq 2.51$), was die Ergebnisse von Schreiber et al. (2024) bestätigt. Das zweite Hauptergebnis zeigte eine mittlere Diskriminationssensitivität für die Identifikation des Ursprungs der Melodiefortsetzungen ($d' = 1.09$). Wir schlussfolgern, dass eine bloße Steigerung der Trainingsquantität von KI-Systemen keine Garantie für eine gleichfalls zunehmende ästhetische Qualität des unter kontrollierten Bedingungen erzeugten musikalischen Outputs bedeutet.

Schlüsselwörter

Künstliche Intelligenz, KI, generative KI, Komposition, empirische Ästhetik, Melodiebewertung, musikalische Kreativität, Sprachmodelle

Background

Following the emergence of user-friendly applications like *ChatGPT* (OpenAI, 2022, 2023a, 2024) in recent years, *artificial intelligence* (AI) has already begun to play an increasingly important role in the everyday lives of younger people in particular. Tools like *DALL·E 3* (OpenAI, 2023b) or the generative features implemented in the newer versions of *Photoshop* (Adobe, 2023) further underline the potential of AI to lower the bar to entry still further when it comes to image creation. And it's a similar story in the musical domain, with strong interest in automating composition processes since the emergence of musical dice games during the eighteenth century (Steinbeck, 2016). At the end of the twentieth century, the topic prominently resurfaced in the *Experiments in Musical Intelligence* by the American composer David Cope (1996) and has been top of mind for empirical researchers ever since.

Almost thirty years later, both hardware and software for artificial music creation have become far more powerful and penetrated across the board. Specialized server-

based applications like *AIVA* (Aiva Technologies, 2016) or *Suno AI* (Suno, 2024) are readily available to all with a smartphone and an internet connection. These differ, among other things, in their openness to input and output formats (e.g., MIDI, Python, text-based prompts, sampled sounds). This paves the way for innumerable potential applications involving the deployment of AI systems in co-creativity processes between humans and machines (Gioti, 2021), which may pose a threat to the livelihood of traditional music creators due to, for example, the time and cost efficiency of AI-generated art. This worry is already very much present today, as shown in a recent report conducted on behalf of the German and French collecting societies *GEMA* and *SACEM*, in which 71% of the surveyed members stated that they see their economic foundation threatened by AI (Goldmedia, 2024). Despite an obvious public interest in the topic, research into the creative musical potential of various AI models remains limited. Few studies have conducted controlled blind evaluations and/or compared the compositions of generative AI models with those of human musicians (Schreiber et al., 2024).

Although Oksanen et al. (2023) systematically reviewed a total of 44 empirical studies on AI in the fine arts between 2003 and 2021, only ten fell into the domain of music. Moreover, the content of these studies differed significantly. For example, Frieler and Zaddach (2022) examined participants' ratings of jazz solos, which were either human- or AI-composed. The solos of professional jazz musicians were rated better on average than those of the AI model. Jazz experts were also able to recognize the AI compositions with an accuracy of 64.4%, as opposed to an accuracy of 41.7% for non-experts. Ferreira et al. (2023) also investigated the discrimination skills of participants with different musical expertise in the domain of classical piano music.

As part of his early experiments, David Cope already described a discrimination test to differentiate between human-made and artificially created music. He called it "The Game" (Cope, 2001), and the average success rate of participants always seemed to hover between 40% and 60% (Cope, 1996). Participants with a success rate of over 66% earned the label "high-scorers" (Cope, 2001).

In a recent study by Schreiber et al. (2024), the authors exploratively analyzed the aesthetic ratings of melody continuations composed by the AI systems ChatGPT 3.5 (OpenAI, 2022) and *Magenta Studio* 2 (Google AI, 2023) as well as by a group of music students in a standardized melody continuation task. Participants were presented with ten AI- and ten human-composed melody continuations, respectively. The authors found that the human compositions were rated significantly higher in terms of subjectively perceived quality on all four rating scales used (liking, interesting, logical and meaningful, and convincing) compared to the AI melodies, $F(1, 67) = 91.114$, $p < .001$, Pillai's trace = 0.857, $\eta^2 = 0.576$. Neither the length of the given melodies nor the musical expertise of the participants impacted the ratings significantly.

Research Questions and Study Aim

This study aims to conceptually replicate the findings of Schreiber et al. (2024) using an updated selection of topical (as of June 2024) *large language models* (LLMs): Qwen 2 72B Instruct (Alibaba Cloud, 2024), Llama 3 70B Instruct (Meta, 2024), and GPT-4 (OpenAI, 2023a). Although music-specific AI applications (e.g., UDIO or SUNO) may produce better results, they currently lack open input formats (e.g., based on Python code) required for the application of a standardized melody continuation task. Therefore, these systems are outside the scope of our research. In line with the definition by Wooldridge and Jennings (1995), the selected LLMs are termed "AI systems" as they do not fulfill the criteria of proactive, flexible, or cooperative behavior with other computing systems, which would characterize "AI agent" systems. The AI systems were accessed via the platform *AcademicCloud* (https://academiccloud.de/), which is available to all university members in Lower Saxony.

To extend the stimulus basis of the original study, we used a different melody from the domain of popular music as a starting point for the standardized continuation task. This replication also encompassed the implication of a discrimination task: Based on the *signal detection theory* (SDT), we were interested in the respective discriminative performances of humans and AI systems vis-a-vis creative products.

The following research questions (RQ) were formulated: (RQ_1) Can the significantly poorer subjective assessments of AI systems compared to human compositions in Schreiber et al. (2024) be replicated with another melody task based on a similar highly expert group of participants (music students)? (RQ_2) Do any significant differences emerge in the subjectively perceived quality of the creative products among the three different LLMs used? (RQ_3) Is the overall discrimination performance in the given task (humans vs. AI composer) better than would otherwise happen by chance?

Hypotheses

Based on the results by Schreiber et al. (2024), our first two hypotheses are as follows:

- (H_1) The subjective quality ratings for melody continuations by humans significantly exceed those obtained by three different AI systems.
- (H_2) The subjective ratings of aesthetic quality differ significantly when the results obtained from the three different AI systems are compared.

Based on the results by Cope (1996), our third hypothesis is as follows:

- (H_3) The perceptual discrimination between continuations by humans vs. AI systems does not exceed an extent that would otherwise happen by chance.

Method

Study Design and Pre-Registration

We measured four dependent variables, each involving a single item, to determine subjective ratings for the aesthetic quality of human and AI-based melody continuations. These four scales were adopted from the study conducted by Schreiber et al. (2024): *convincing*, *logical and meaningful*, *interesting*, and *liking*. The condition *composer_specific* served as an independent variable with four gradations: *Qwen 2*, *Llama 3*, *GPT-4*, and *human*. This resulted in a repeated measures design with four independent variables and the composer as within-subject factor. Given our interest in establishing a distinction between AI systems and humans, we aggregated the three AI models in portions of the statistical analysis, resulting in a dichotomous differentiation with two remaining gradations for the independent variable (AI vs. human).

An a priori power analysis was calculated using G*Power (Faul et al., 2009) based on the (large) effect sizes reported by Schreiber et al. (2024). Regarding the relevant Pillai's trace value of 0.857 (the value of the main effect *composer*), the number of participants required to reach the desired power of $1-\beta = .95$ at a standard probability of .05 α-error was a minimum of $N \geq 12$. The study was pre-registered (see Meier et al., 2024).

Musical Stimuli

We used a melody continuation paradigm in line with Schreiber et al. (2024) and set it against the lack of tools for the standardized evaluation of performance in AI and music research, as diagnosed by Mycka and Mańdziuk (2025). As a starting point, we simplified the chorus melody of a German pop ballad (*Durch die schweren Zeiten [Through the Hard Times]* by Udo Lindenberg; see Figure A1). The musical stimuli were created as follows: (a) music students from various study programs (mainly music education) at the Hanover University of Music, Drama and Media composed continuations of the given melody following a standardized instruction (see Appendix 1); (b) The three selected AI systems also continued the melody following the same standardized instruction (see Appendix 2 for the given text prompt). During the next step, which involved converting musical products from all three AI systems into audio files for experimental use, we transcribed the melody output in Python syntax using the module *SCAMP* (Evanstein, 2023) and instructed the systems to return their continuations in the same format.

The resulting melody continuations from the music students and AI systems were then converted into MIDI format using *MuseScore* (MuseScore Ltd, 2024) and SCAMP (Evanstein, 2023) respectively. In the final step, we imported these MIDI files into *Reaper* (Cockos, 2024), before exporting them as MP3 audio files using the *BBC Symphony Orchestra Discover* (Spitfire Audio, 2023) sample library with timbre clarinet. Sound files were normalized to a loudness of -20 LUFS-I. A total of $N = 98$ melodies were generated

(Qwen 2: n = 25, Llama 3: n = 25, GPT-4: n = 25, and human n = 23; see Supplementary Materials section for details, Meier et al., 2025).

Procedure

The study was conducted as an online experiment: 55 participants completed a questionnaire on the *SoSci Survey* platform (https://www.soscisurvey.de/en/index). Participants were informed that they would listen to some melodies, which started in the same way but then continued with either AI- or human-composed portions. After giving their informed consent via a checkbox (see Statement of Ethics), participants had the chance to hear example audio of the given melody, from which the AI systems and music students had established a musical continuation. Participants were presented with a random selection of 20 melodies (five for each condition: Qwen 2, Llama 3, GPT-4, and human) in a randomized, blinded trial, resulting in an incomplete study design. The melodies were rated on a 5-point scale (1 = *not at all* [*gar nicht*] to 5 = *very much* [*sehr*]) using the criteria *convincing*, *logical and meaningful*, *interesting*, and *liking*. There was no trial run. In addition to the rating, participants used an eight-point scale (1 = *clearly human* [*eindeutig Mensch*], 8 = *clearly machine* [*eindeutig Maschine*], see Figure A2) to indicate who they thought had composed the musical continuation, human or AI and how confident they were (with endpoints of the scale indicating higher confidence).

After listening to and rating the melodies, participants specified their gender, age, and musical identity on a single item (Zhang & Schubert, 2019). They were also informed of their number of correct responses in the detection task at the end of the experiment. Given time constraints and the presence of a homogenous group of highly expert participants, we decided against using an additional inventory to control musical sophistication and the use of different melodic probe positions (given melodic lengths), since neither of these variables was found to have any significant impact on the study by Schreiber et al. (2024). The experiment took about 20 minutes to complete, and participants were not remunerated for their involvement.

Sample Description

Of the 55 initial participants, one was excluded due to spurious responses, resulting in a final sample size of N = 54. Participants were recruited via mailing lists at different German-speaking Universities of Music. From the sample, 30 participants (55.6%) were male, 23 (42.6%) female, and one (1.9%) non-binary. The participants were aged between 19 and 61 years (Mdn = 25, IQR = 9). Sporadic outliers from the average age can be explained by the participation of some professors completing the questionnaire in addition to their students. The fact that we actively targeted music university students should have ensured an above-average level of musical expertise for the entire sample, with 45 participants (83.3%) reporting at least 10 years of formal music lessons.

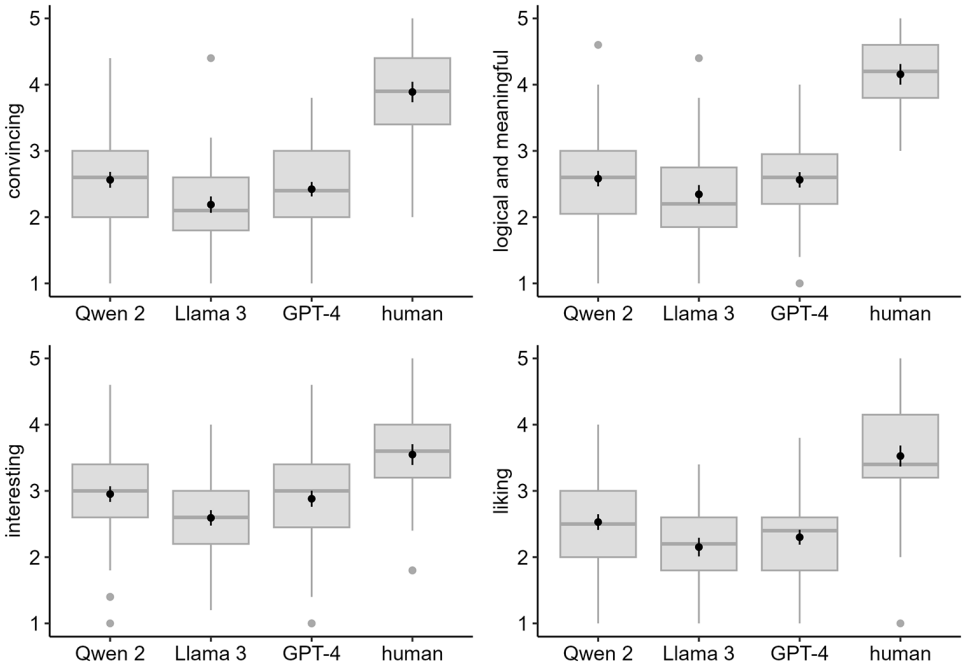
Results

Ratings of the Melodies

As Figure 1 shows, the human-composed melodies were rated higher for all four dependent variables. Comparing the three LLMs, we can see that Qwen 2 always obtained the highest average ratings, followed closely by or tied with GPT-4. Llama 3 always showed the lowest scores.

Figure 1

Box Plots of the Melody Ratings for Each Dependent Variable Grouped by Composer_Specific

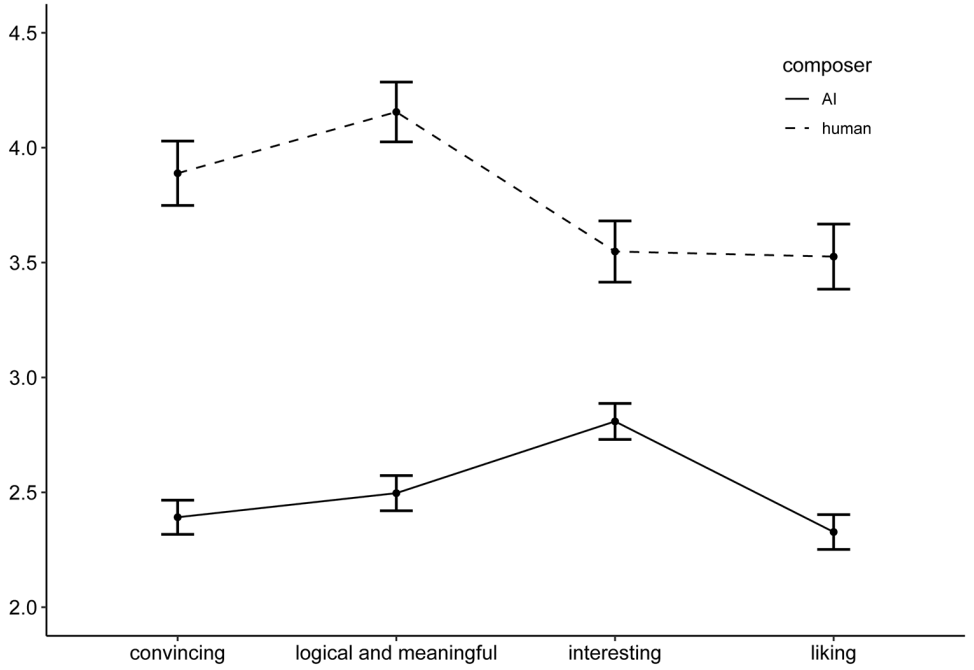


Note. Central tendencies represent the mean with 95% CI, horizontal bars represent the median, and grey dots represent outliers.

Given our key focus on comparing AI systems and humans, we aggregated the melody ratings for the three LLMs into a single independent variable *composer*. Figure 2 thus shows the rating curve across all four dependent variables grouped by AI vs. human.

Figure 2

Error Bar Diagram for Each Dependent Variable in the Dichotomous Comparison of AI vs. Human



Note. The rating scale ranges from 1 (minimum) to 5 (maximum). Error bars represent means with 95% CI.

Accordingly, we conclude that Hypothesis 1 (improved aesthetic verdict on human-based melody continuations compared to three different AI systems) could be confirmed. Results also confirm Hypothesis 2 (significant differences between the three AI systems in the aesthetic quality of the melody continuations).

To test for the main effect of the independent variable *composer*, we calculated a repeated measures MANOVA in R (R Core Team, 2024) using the stats package. This analysis revealed that the *composer* factor impacted significantly on the rating of the four dependent variables, $F(1, 53) = 94.07$, $p < .001$, Pillai's trace = 0.88, $\eta^2 = 0.64$. We also conducted *t*-tests for groupwise comparisons for all four dependent variables from Figure 2. Effect sizes show differences in the vicinity of large effects ($d_Z > 0.8$; for benchmarks see Ellis, 2010) in favor of the human-based compositions (see Table 1). Differences between the four evaluation items and both sources of melodic origin for the original study and its replication are very similar.

Table 1

Effect Sizes for the Comparison of AI vs. Human (t-Tests) Compared to Those Found by Schreiber et al. (2024)

	Present Study			Schreiber et al. (2024)		
		95% CI			95% CI	
DV	Cohen's d_z	LL	UL	Cohen's d_z	LL	UL
convincing	−2.29	−2.80	−1.78	−2.11	−2.53	−1.69
logical and meaningful	−2.51	−3.05	−1.97	−1.93	−2.32	−1.53
interesting	−1.11	−1.45	−0.77	−1.74	−2.11	−1.37
liking	−1.79	−2.22	−1.36	−2.23	−2.66	−1.79

Note. Negative values indicate that AI melody continuations were rated lower than those generated by humans.

To round off this part of the data analysis, we calculated a correlation matrix (Pearson) for all four dependent variables to assess the strength of inter-variable correlations between the different rating scales (see Table 2). The target variables *convincing* and *liking* showed the strongest inter-correlations.

Table 2

Correlation Matrix for the Dependent Variables (Pearson)

Variable	1	2	3	4
1. convincing	—			
2. logical and meaningful	0.89	—		
3. interesting	0.79	0.73	—	
4. liking	0.90	0.79	0.83	—

Note. $p < .001$ for all correlations ($N = 54$).

Discrimination Performance

The second step of data analysis aimed to reveal the underlying discrimination performance of listeners as measured by Signal Detection Theory (SDT). Since the melody continuation falls into two distinct categories (composed either by AI or by humans), with only one at a time presented for evaluation, our discrimination task falls into the family of so-called A-Not A or Yes-No experiments (Bi & Ennis, 2001; Hautus et al., 2021). In our study, a melody continuation composed by an AI system is designated as "A", "Yes", or the presence of the signal, while a human-composed continuation is assigned to the category "Not A", "No", or the absence of the signal—in short, an AI-Not AI design. Following the classification of A-Not A sub-designs provided by Düvel and Kopiez (2022, Table 1), we have a replicated mixed A-Not A (but not paired) design—replicated because

participants were presented with multiple stimuli; mixed because they were presented with stimuli from both categories (A and Not A).

Participants' responses on the eight-point rating scale were dichotomized (AI and human), with responses ranging from one to four being counted as human and five to eight being counted as AI. They were then classified as follows: If an AI-composed melody continuation was correctly identified, the response was coded as a "hit", but if a human-composed melody continuation was misidentified as AI-composed, the response was coded as "false alarm". Conversely, if a human-composed melody continuation was correctly identified as human-composed, the response was coded as "correct rejection", but identifying an AI-composed melody continuation as human-composed triggered a "miss" response. Table 3 shows the respective frequencies of hits, misses, correct rejections, and false alarms. The independence of the stimulus type and the participants' responses can be tested based on this 2 × 2 table. Following Bi (2015) and Brier (1980), the test statistic is a conventional Pearson χ^2 test with a single degree of freedom if the data are transformed with a correction factor. The test results in $\chi^2 = 111.62$, $p < .001$, and with Yates' continuity correction in $\chi^2 = 109.82$, $p < .001$. Accordingly, the null hypothesis that participants' responses are independent of the stimulus type has to be rejected.

Table 3

Frequencies of Signal Detection Theory Response Types

		Melody continuation composed by		
		AI	**Human**	**Row Sums**
Participant responds	AI	Hits	False alarms	
		n = 546	*n* = 62	*n* = 608
		50.6%	5.7%	56.3%
	Human	Misses	Correct Rejections	
		n = 264	*n* = 208	*n* = 472
		24.4%	19.3%	43.7%
Column Sums		*n* = 810	*n* = 270	*n* = 1080
		75.0%	25.0%	100.0%

To quantify a participant's discrimination performance, we calculated the sensitivity d prime (d'), which was based on the participant's hit rate and false-alarm rate (Hautus et al., 2021, p. 7). These rates were then converted to z scores with the inverse of the normal cumulative distribution function (Φ^{-1}, see Equation 1).

$$d' = \Phi^{-1}(\text{hit rate}) - \Phi^{-1}(\text{false-alarm rate}) \quad \text{(Equation 1)}$$

A *d'* value of 0 designates an occurrence that would otherwise happen by chance. Positive values indicate scope for participants to discriminate AI-composed melody continuations from those by humans.

We also calculate the response bias or criterion *c* (see Equation 2).

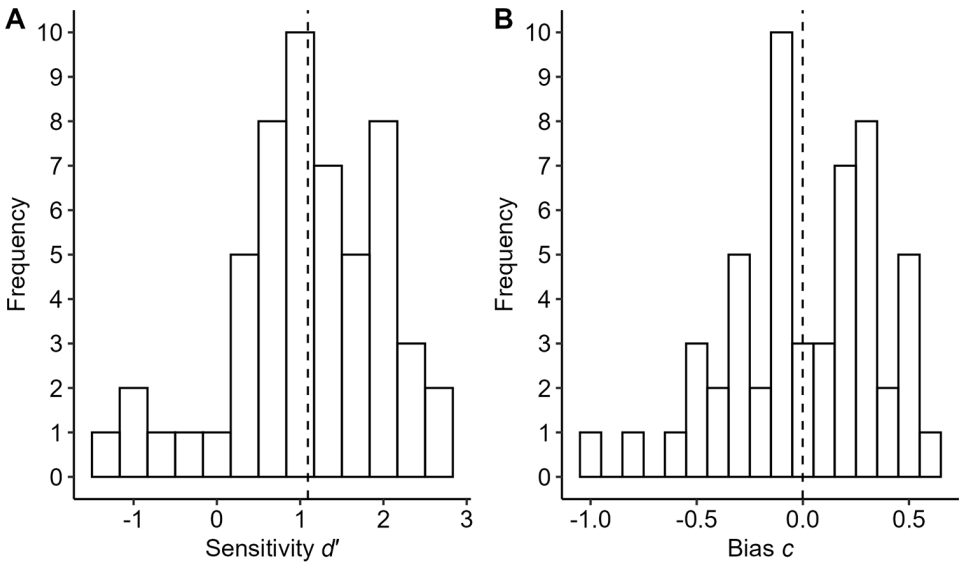
$$c = -\frac{1}{2}\left(\Phi^{-1}(\text{hit rate}) + \Phi^{-1}(\text{false-alarm rate})\right) \qquad \text{(Equation 2)}$$

The response bias reflects the tendency toward one of the two response categories (Hautus et al., 2021). In our case, $c < 0$ indicates a tendency to decide that the melody was continued by an AI system. For both, sensitivity and response bias, hit or false-alarm rates of 0 or 1 had to be corrected (otherwise *d'* or *c* equal $\pm\infty$). Following Hautus et al. (2021, p. 7), a rate of 0 was corrected to $1/(2N)$, where N is the number of trials on which the rate is based; a rate of 1 was corrected to $1-1/(2N)$.

As per Figure 3A, most sensitivity values and their mean of $d' = 1.09$ exceeded 0. A one-tailed one-sample *t*-test also revealed a significant difference from 0, $t(53) = 8.88$, $p < .001$, Cohen's $d = 1.21$, 95% CI [0.61, 1.80]. Based on the benchmarks for *d'* provided by Bi (2015, Table 3.1), a mean sensitivity of $d' = 1.09$ constitutes a meaningful discrimination sensitivity ($0.74 \leq d' \leq 1.81$) between AI- and human-composed melody continuations.

Figure 3

Histograms of Sensitivity and Response Bias



Note. Panel A: Histogram of the sensitivity *d'*. Panel B: Histogram of the response bias *c*. Dashed lines represent the means of *d'* and *c*, respectively.

Considering Figure 3B, the response bias scores seem to be distributed around 0. Indeed, their mean value is almost 0. A two-tailed one-sample *t*-test indicated no difference from 0, $t(53) = -0.02$, $p = .986$, Cohen's $d = -0.00$, 95% CI [−0.55, 0.54].

Table A1 (see Appendix) shows the frequencies of hits and misses for each AI system. The melody continuations by Llama 3 were assigned as AI more often than those by ChatGPT 4 which, in turn, were detected as AI-based more often than those generated by Qwen 2. Accordingly, the frequencies of misses are in the same ranking order as the AI system ratings.

Using the musical identity item (years of formal instrumental lessons), the sample was split into two groups (Group 1: > 10 years, $n_{>\ 10\ \text{years}} = 45$; Group 2: 6–10 years, $n_{6–10\ \text{years}} = 9$). As shown by Figure A3 (see Appendix), the sensitivity values covered a similar range in both groups. However, their means of $d'_{>\ 10\ \text{years}} = 1.18$ and $d'_{6–10\ \text{years}} = 0.68$ were compared using a Welch's *t*-test, leading to a non-significant result of $t(9.86) = -1.24$, $p = .242$.

Discrimination Performance in the Study by Schreiber et al. (2024)

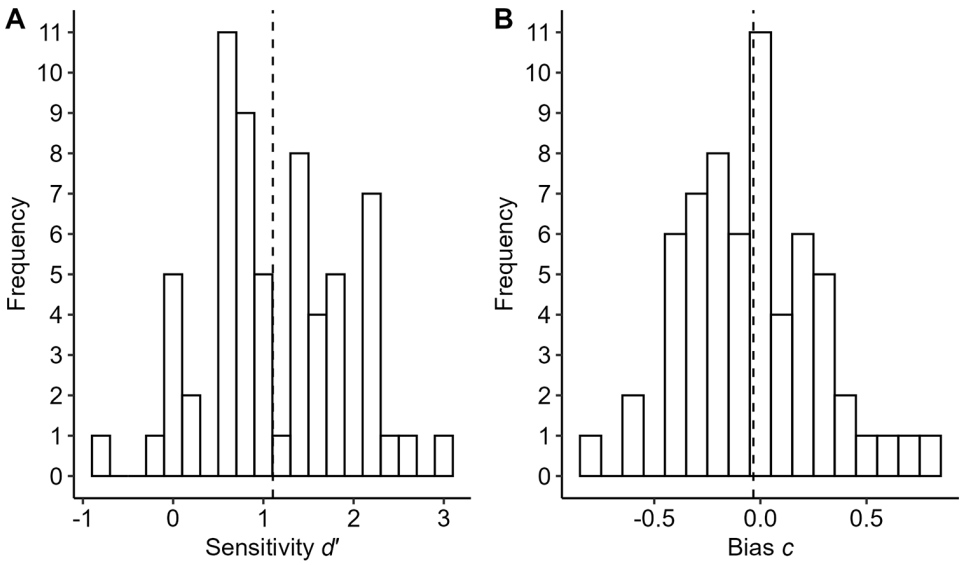
Although Schreiber et al. (2024) did not report any results on discrimination performance, the authors had used the same discrimination paradigm in their study. For this reason, after obtaining the raw data from their study, responses were coded identically to conduct equivalent discrimination analyses. Nine participants were removed from the data set (one due to missing values and spurious responses, and eight because they had been presented with one less stimulus in the detection task) resulting in a valid sample of $N = 62$.

Table A2 (see Appendix) displays the absolute and relative frequencies of the SDT response types. The statistical test on this 2 × 2 table resulted in $\chi^2 = 227.11$, $p < .001$, and with Yates' continuity correction in $\chi^2 = 225.58$, $p < .001$. This means the null hypothesis that participants' responses in the study by Schreiber et al. (2024) are independent of the stimulus type also has to be rejected.

Sensitivity and response bias spawn mean values of $d' = 1.12$ and $c = -0.03$, respectively. As shown by Figure 4, the distributions of sensitivity and response bias broadly resemble those from the replication study (see Figure 3). Based on one-sample *t*-tests, only the sensitivity differs from 0, one-tailed $t(61) = 11.21$, $p < .001$, Cohen's $d = 1.42$, 95% CI [0.86, 1.99], but not the response bias, two-tailed $t(61) = -0.85$, $p = .401$, Cohen's $d = -0.11$, 95% CI [−0.62, 0.40]. As in the data set of our replication study, the mean sensitivity of $d' = 1.12$ in the original study is in the benchmark interval for a meaningful difference between AI- and human-composed melody continuations ($0.74 \leq d' \leq 1.81$; Bi, 2015, Table 3.1).

Figure 4

Histograms of Sensitivity and Response Bias in the Data from Schreiber et al. (2024)



Note. Panel A: Histogram of the sensitivity *d'*. Panel B: Histogram of the response bias *c*. Dashed lines represent the means of *d'* and *c*, respectively.

Similar to the results based on the ratings reported by Schreiber et al. (2024), ChatGPT 3.5 recorded a higher number of misses (i.e., its melody continuations were more often misassigned as of "human" origin) than Magenta (see Table A3). The data by Schreiber et al. (2024) encompassed all three levels of musical identity. As can be seen in Figure A4 (see Appendix), the sensitivity values are similarly distributed in the groups. Neither a one-way analysis of variance, $F(2, 59) = 0.41$, $p = .666$, $\eta_{generalized} = .014$, nor a Kruskal-Wallis test, $\chi^2(2) = 1.58$, $p = .453$, revealed any significant difference in sensitivity values between the three groups.

Discussion

Our results largely correlate those resulting from the original study by Schreiber et al. (2024) and we were able to replicate the authors' main outcome, namely that "the subjectively perceived and empirically confirmed quality of AI compositions is far below human-made compositions." (p. 7) This difference was particularly clear in the case of the two dependent variables *convincing* as well as *logical and meaningful*, where the AI compositions scored over two standard deviations lower than their human-made coun-

terparts. This becomes apparent when listening to some of the compositions generated by the LLMs (see Supplementary Materials section for sound examples, Meier et al., 2025). We agree with Schreiber et al.'s (2024) sentiment, that "the AI melodies sounded illogical and strange to our Western understanding of melodic construction." (p. 7) This becomes especially obvious towards the end of the melodic continuations, since they often stop abruptly and without reaching the correct tonal resolution.

The AI systems used in our study also seemed to lack all concept of tonality. For example, despite having a melody in the key of E minor, the AI systems frequently used an F instead of the correct diatonic scale step of F sharp in their melodic continuations, resulting in an unmistakably off-key result.

Comparing the effect sizes in Table 1 to those reported by Schreiber et al. (2024), we note less of a difference in ratings for the dependent variables *interesting* and *liking*, while the gap for *logical and meaningful* has widened. This suggests that although participants subjectively enjoyed listening to those unconventional melodic continuations, they simultaneously evaluated them as objectively worse in terms of music theory categories. The latter could result from our expert sample recognizing technical flaws more accurately.

The SDT analyses of both, our replication and the original study by Schreiber et al. (2024) show that participants could discriminate AI-composed melody continuations from those created by humans. Furthermore, no significant differences emerged in the detection performance between the two levels of musical identity (> 10 years and 6–10 years of instrumental lessons). In other words, identifying and evaluating AI-generated music does not require high compositional expertise, and the musical sophistication of an average music student seemingly suffices. Based on current research, the question remains open whether listeners without formal music training might also be able to discriminate both sources of musical creation (AI and human) reliably, leveraging only musical capacities acquired through mere everyday exposure to tonal music. This would tally with the "experienced listeners' hypothesis" by Bigand and Poulin-Charronnat (2006), who conclude in their review that even musically untrained listeners could also manage most musical tasks in experimental studies.

Our finding of a better-than-chance level discrimination performance between AI- and human-generated melody continuations contrasts with Cope and his listening test to identify AI-generated musical style imitations known as "The Game", who identified correct responses around at a level commensurate with what would happen by chance. We should also take into account, however, the fact that Cope's musical AI-based examples were composed with many degrees of freedom and very little control in the generation of stimuli by AI because he was interested in the machine's potential to copy a musical style following extensive training on notated score material. Conversely, we focused instead on the creative potential of AI systems to find new and musically valuable solutions within a standardized melody continuation paradigm.

We also conclude that our results are in contrast to the current prevalent concerns expressed by music creators (Goldmedia, 2024) regarding the musical capacities of current AI models. It should be noted here nonetheless that the creative potential of music-specific AI applications remains unclear and may outperform the three LLMs examined. However, recent studies in the field of *music information retrieval* have demonstrated that large language models (LLMs) are being regarded as suitable research instruments for music processing and generation tasks (see https://m-a-p.ai/LLM4Music/). A continuing academic examination of what AI systems can "truly" creatively achieve under controlled conditions is thus highly relevant for the music industry and should underpin the relevant degree programs more firmly going forward (Tillmann & Zaddach, 2024). As AI applications still develop apace, with no signs of stopping, this topic will only become more relevant and the models more advanced. For the moment, however, we conclude that the creative potential of the systems has improved little since the investigation by Schreiber et al. (2024): although conducted in 2023 and based on the previous generation of AI systems (e.g., ChatGPT 3.5), human composers still outperform modern LLMs significantly when comparisons are based on standardized conditions such as a melody continuation task. We conclude that merely boosting the volume of the AI systems' training (for example, boosting the total number of parameters by a factor of 500, as happens between ChatGPT 3 and 4; see Portakal, 2023) does not in itself guarantee a corresponding improvement in the creative musical output produced under controlled conditions.

Accordingly, our evaluation of music-generating AI systems fully tallies with the review by Mycka and Mańdziuk (2025), and within it, the development of objective and standardized production paradigms comes to the fore. Our suggested melody continuation paradigm could be a first step in this direction.

Limitations

To ensure the composition task would be as fair and the results as comparable as possible across all four conditions of the independent variable *composer_specific* (Qwen 2, Llama 3, GPT-4, and human), we opted to use the same standardized melody continuation task as in Schreiber et al. (2024). Similarly, our use of available AI models was considerably restricted by the existence of an interface for the input of a prompt while referencing the original melody.

Currently, AI systems specialized in music production such as Suno AI (Suno, 2024) or AIVA (Aiva Technologies, 2016) do not offer this option of prompt input. Thus, we decided to focus on the comparison of three selected, currently competing LLMs that represent a generic AI approach. We cannot exclude that a more openly phrased prompt or the use of AI models specializing more strongly in the domain of music would yield better results in favor of AI systems. We agree with Schreiber et al. (2024), that

> more studies like this should be conducted to constantly assess the development of the quality of AI compositions. By doing so, music research can make a valuable contribution to musicians and creatives in empirically investigating the progress being made by musical AI. (p. 9)

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: RK is Editor-in-Chief, and KS is Editorial Assistant of the *Jahrbuch Musikpsychologie/Yearbook of Music Psychology*. They were not involved in the editorial process of this manuscript.

Author Contributions: *Nicholas Meier*—Conceptualization | Methodology | Investigation | Data curation | Formal analysis | Visualization | Writing – original draft | Funding acquisition | Project administration. *Kilian Sander*—Conceptualization | Methodology | Investigation | Data curation | Formal analysis | Visualization | Supervision | Validation | Writing – review & editing | Funding acquisition | Project administration. *Anton Schreiber*—Conceptualization | Methodology. *Reinhard Kopiez*—Conceptualization | Methodology | Supervision | Validation | Writing – review & editing | Funding acquisition | Project administration.

Ethics Statement: The present study was conducted in accordance with ethical principles and standards pursuant to the guidelines of the German Society for Psychology (Föderation Deutscher Psychologenvereinigungen, 2022) and with the principles outlined in the Declaration of Helsinki. The study also adhered to the research regulations of the Hanover University of Music, Drama and Media. According to German law, no ethics approval was required. Written informed consent was obtained by reconfirming that all individuals were both willing to participate and had read and understood the instructions and information provided. Participants were informed that participation was voluntary and that they could withdraw from the study at any time. The data were also anonymized and treated confidentially.

Data Availability: For this article, R scripts, data, codebook, and musical stimuli are available (see Meier et al., 2025).

Supplementary Materials

For this article, R scripts, data, codebook, and musical stimuli are available (see Meier et al., 2025). The study was pre-registered (see Meier et al., 2024).

Index of Supplementary Materials

Meier, N., Kopiez, R., & Sander, K. (2024). *Melody continuation with AI: The creative achievement of different language models compared to music students* [Preregistration]. OSF Registries. https://osf.io/2zbkx

Meier, N., Sander, K., Schreiber, A., & Kopiez, R. (2025). *The creative musical achievement of AI systems compared to music students: A replication of the study by Schreiber et al. (2024)* [Data, codebook, code, stimuli]. OSF. https://osf.io/5mcpt/

References

Adobe. (2023). *Photoshop* (Version 25) [Computer software]. https://www.adobe.com/de/products/photoshop.html

Aiva Technologies. (2016). *AIVA* [Computer software]. https://www.aiva.ai/

Alibaba Cloud. (2024). *Qwen* (Version 2) [Computer software]. https://github.com/QwenLM/Qwen2

Bi, J. (2015). *Sensory discrimination tests and measurements: Sensometrics in sensory evaluation* (2nd ed.). Wiley Blackwell.

Bi, J., & Ennis, D. M. (2001). Statistical models for the A-Not A method. *Journal of Sensory Studies, 16*(2), 215–237. https://doi.org/10.1111/j.1745-459X.2001.tb00297.x

Bigand, E., & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. *Cognition, 100*(1), 100–130. https://doi.org/10.1016/j.cognition.2005.11.007

Brier, S. S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika, 67*(3), 591–596. https://doi.org/10.1093/biomet/67.3.591

Cockos. (2024). *Reaper* (Version 7.16) [Computer software]. https://www.reaper.fm/index.php

Cope, D. (1996). *Experiments in musical intelligence.* A-R Editions.

Cope, D. (2001). *Virtual music: Computer synthesis of musical style.* MIT Press.

Düvel, N., & Kopiez, R. (2022). The paired A–Not A design within signal detection theory: Description, differentiation, power analysis and application. *Behavior Research Methods, 54*(5), 2334–2350. https://doi.org/10.3758/s13428-021-01728-w

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results.* Cambridge University Press.

Evanstein, M. (2023). *SCAMP (Suite for Computer-Assisted Music in Python)* (Version 0.9.2) [Computer software]. http://scamp.marcevanstein.com/

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Ferreira, P., Limongi, R., & Fávero, L. P. (2023). Generating music with data: Application of deep learning models for symbolic music composition. *Applied Sciences, 13*(7), Article 543. https://doi.org/10.3390/app13074543

Föderation Deutscher Psychologenvereinigungen. (2022). *Berufsethische Richtlinien [Guidelines for professional ethics].* https://www.dgps.de/fileadmin/user_upload/PDF/Berufsetische_Richtlinien/BER-Foederation-20230426-Web-1.pdf

Frieler, K., & Zaddach, W.-G. (2022). Evaluating an analysis-by-synthesis model for jazz improvisation. *Transactions of the International Society for Music Information Retrieval, 5*(1), 20–34. https://doi.org/10.5334/tismir.87

Gioti, A.-M. (2021). Artificial intelligence for music composition. In E. R. Miranda (Ed.), *Handbook of artificial intelligence for music* (pp. 53–73). Springer International Publishing. https://doi.org/10.1007/978-3-030-72116-9_3

Goldmedia. (2024). *AI and music: Market development of AI in the music sector and impact on music authors and creators in Germany and France.* https://www.goldmedia.com/produkt/study/ki-und-musik/

Google AI. (2023). *Magenta Studio* (Version 2) [Computer software]. https://magenta.tensorflow.org/studio/

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection theory: A user's guide* (3rd ed.). Routledge. https://doi.org/10.4324/9781003203636

Meta. (2024). *Llama* (Version 3) [Computer software]. https://llama.meta.com/llama3/

MuseScore Ltd. (2024). *MuseScore* (Version 4.3.2) [Computer software]. https://musescore.com/about

Mycka, J., & Mańdziuk, J. (2025). Artificial intelligence in music: Recent trends and challenges. *Neural Computing & Applications, 37*(2), 801–839. https://doi.org/10.1007/s00521-024-10555-x

Oksanen, A., Cvetkovic, A., Akin, N., Latikka, R., Bergdahl, J., Chen, Y., & Savela, N. (2023). Artificial intelligence in fine arts: A systematic review of empirical research. *Computers in Human Behavior: Artificial Humans, 1*(2), Article 100004. https://doi.org/10.1016/j.chbah.2023.100004

OpenAI. (2022). *ChatGPT* (Version 3.5) [Computer software]. https://openai.com/chatgpt/

OpenAI. (2023a). *ChatGPT* (Version 4) [Computer software]. https://openai.com/index/gpt-4/

OpenAI. (2023b). *DALL·E* (Version 3) [Computer software]. https://openai.com/index/dall-e-3/

OpenAI. (2024). *ChatGPT* (Version 4o) [Computer software]. https://openai.com/index/hello-gpt-4o/

Portakal, E. (2023, March 20). GPT-3 vs. GPT-4 Vergleich. *TextCortex Blog*. https://textcortex.com/de/post/gpt-3-vs-gpt-4-comparison

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.1) [Computer software]. https://www.r-project.org/

Schreiber, A., Sander, K., Kopiez, R., & Thöne, R. (2024). 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*, Article e195. https://doi.org/10.5964/jbdgm.195

Spitfire Audio. (2023). *BBC Symphony Orchestra Discover* (Version 1.7.0) [Computer software]. https://www.spitfireaudio.com/bbc-symphony-orchestra-discover

Steinbeck, W. (2016). Würfelmusik. In L. Lütteken (Ed.), *MGG Online*. https://www.mgg-online.com/mgg/stable/12552

Suno. (2024). *Suno AI* (Version 3.5) [Computer software]. https://suno.com/

Tillmann, B., & Zaddach, W.-G. (2024). Artificial intelligence in songwriting and composing: Perspectives and challenges in creative practices. In E. Voigts, R. M. Auer, D. Elflein, S. Kunas, J. Röhnert, & C. Seelinger (Eds.), *Artificial intelligence—Intelligent art?: Human-machine interaction and creative practice* (pp. 217–231). transcript. https://doi.org/10.14361/9783839469224

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review, 10*(2), 115–152. https://doi.org/10.1017/S0269888900008122

Zhang, J. D., & Schubert, E. (2019). A single item measure for identifying musician and nonmusician categories based on measures of musical sophistication. *Music Perception, 36*(5), 457–467. https://doi.org/10.1525/mp.2019.36.5.457

Appendix

Appendix 1: Instructions for the Music Students

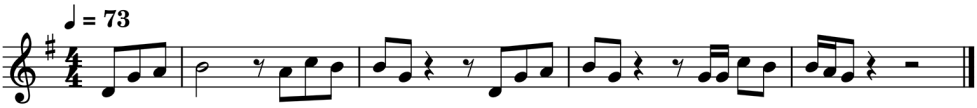
The AIs will complete the beginning of the melody shown below with ten to 20 notes. I therefore ask you to complete the melody shown below (see Figure A1) according to your own ideas. A few rules apply: the continuation should ...

... comprise ten to 20 notes,
... be within the range G3 (g) to G5 (g''),
... contain different note lengths (i.e., not just quarter notes, for example),
... have a clear melodic peak.

We will need three to five continuations per person. You can compose the versions yourself or ask fellow students for versions. You can either write down the continuation on music paper, sing or play it on your instrument of choice and record it, or enter it directly into a notation program (e.g., MuseScore). For later evaluation, notation on the computer is most practical, but not mandatory. Transposing instruments can be notated as they were fingered. For the evaluation, all examples will be transposed to a standard pitch.

Figure A1

Score of the Given Melody for the Continuation Task (Chorus of Durch die schweren Zeiten [Through the Hard Times] by Udo Lindenberg)



Appendix 2: Prompt for the AI Systems (LLMs)

Continue the given melody in the form of a list of (pitch, duration) pairs in Python syntax, where the pitch uses the MIDI pitch standard, and the duration represents the number of quarter notes. Use a pitch of None to represent a rest. Ensure the following:

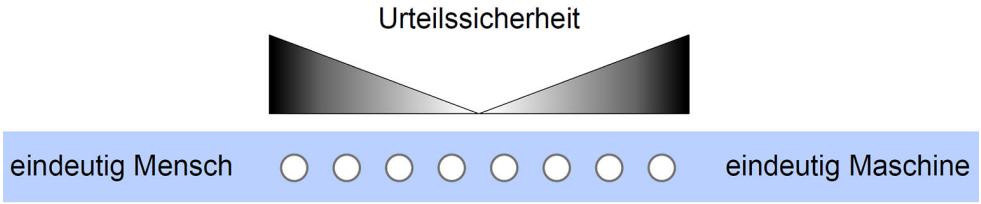
- The continuation stays between MIDI pitch 55 and MIDI pitch 79
- The continuation is between 10 and 20 notes in length
- The melody should be in the style of a pop ballad
- The continuation should use a variety of note lengths
- The continuation should have a clear melodic peak

melody = [(62, 0.5), (67, 0.5), (69, 0.5), (71, 2.0), (None, 0.5), (69, 0.5), (72, 0.5), (71, 0.5), (71, 0.5), (67, 0.5), (None, 1.5), (62, 0.5), (67, 0.5), (69, 0.5), (71, 0.5), (67, 0.5), (None, 1.5), (67, 0.25), (67, 0.25), (72, 0.5), (71, 0.5), (71, 0.25), (69, 0.25), (67, 0.5), (None, 1.5)]

Appendix 3: Discrimination Task and Performance

Figure A2

Response Scale for the Discrimination Task



Note. "eindeutig Mensch" = clearly human; "Urteilssicherheit" = confidence; "eindeutig Maschine" = clearly machine.

Table A1

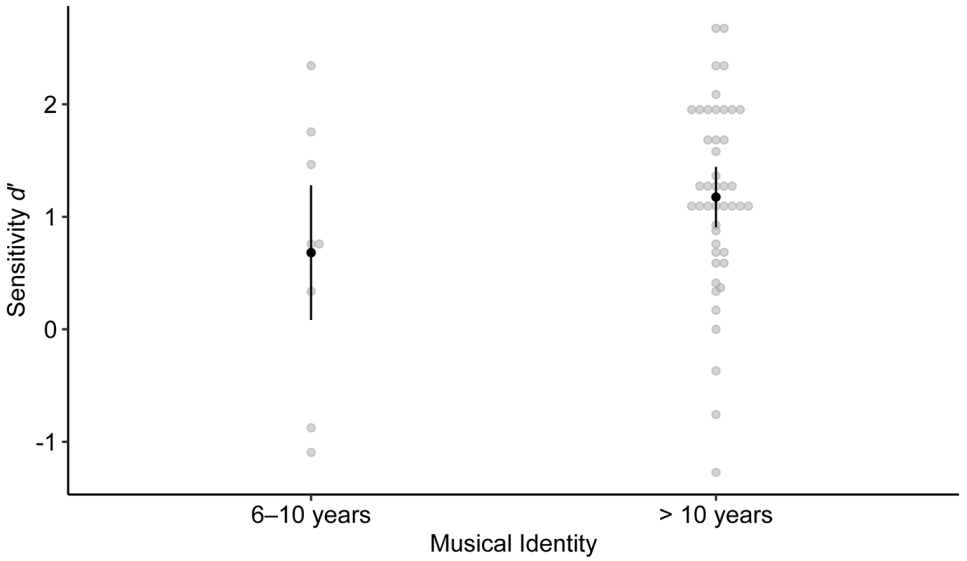
Hits and Misses per AI System

	Hits		Misses		
Composer	***n***	**%**	***N***	**%**	**Total**
Chat GPT-4	181	67.0	89	33.0	270
Llama 3	199	73.7	71	26.3	270
Qwen 2	166	61.5	104	38.5	270

Note. N = 54 participants.

Figure A3

Sensitivity d' by Musical Identity



Note. Grey dots represent raw data. Black dots and error bars represent means and 95% confidence intervals, respectively. $n_{6–10\ \text{years}} = 9$; $n_{>\ 10\ \text{years}} = 45$.

Table A2

Frequencies of Signal Detection Theory Response Types in the Data From Schreiber et al. (2024)

		Melody continuation composed by		
		AI	**Human**	**Row Sums**
Participant responds	AI	Hits	False alarms	
		n = 434	*n* = 197	*n* = 631
		35.0%	15.9%	50.9%
	Human	**Misses**	**Correct Rejections**	
		n = 186	*n* = 423	*n* = 609
		15.0%	34.1%	49.1%
Column Sums		*n* = 620	*n* = 620	*n* = 1240
		50.0%	50.0%	100.0%

Note. $N = 62$ participants.

Table A3

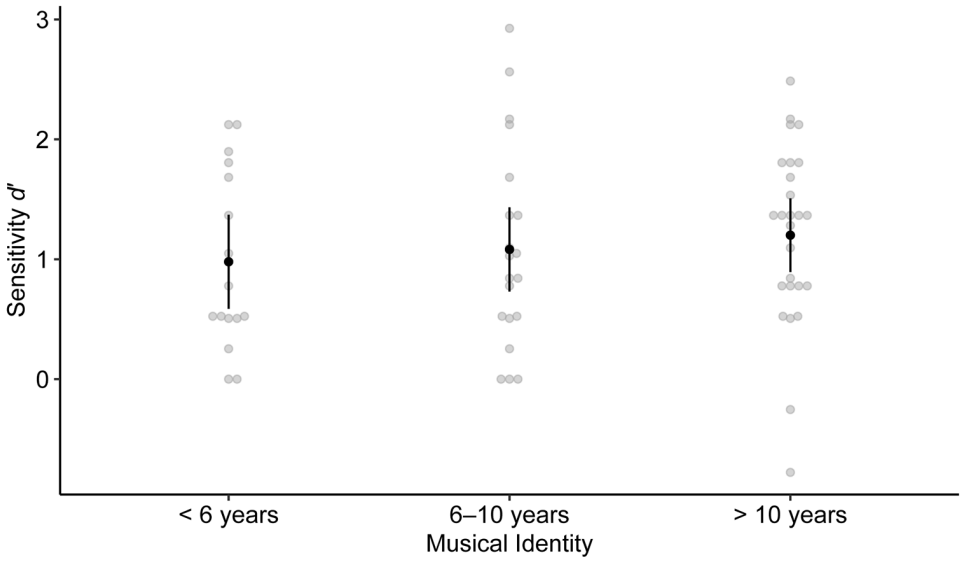
Hits and Misses per AI System in the Data From Schreiber et al. (2024)

AI System	**Hits**		**Misses**		
Composer	***n***	**%**	***n***	**%**	**Total**
Chat GPT 3.5	217	64.2	121	35.8	338
Magenta 2.0	217	77.0	65	23.0	282

Note. $N = 62$ participants.

Figure A4

Sensitivity d' by Musical Identity in the Data From Schreiber et al. (2024)



Note. Grey dots represent raw data. Black dots and error bars represent means and 95% confidence intervals, respectively. $n_{< 6 \text{ years}} = 16$; $n_{6–10 \text{ years}} = 20$; $n_{> 10 \text{ years}} = 26$.