

# A Multi-Genre Study of Identification and Style Bias of AI-Generated Music

Joseph Longardner

Sichuan Conservatory of Music

[longardnerjj@gmail.com](mailto:longardnerjj@gmail.com)

## Abstract

This study examines the perceptual boundaries between human-produced and AI-generated music through a controlled listening experiment involving 120 participants from the Sichuan Conservatory of Music. Participants evaluated 720 excerpts across six genres (Western Classical, Jazz, Rock, Pop, R&B and Chinese Traditional), with a proportional number of AI and human tracks. The AI music was generated using *Suno AI* and *Udio AI*, while the human tracks were sourced from commercial recordings. Identification accuracy was strongly genre- and model-dependent: Chinese Traditional and Western Classical were most often classified correctly, Jazz and R&B showed the lowest AI detectability, and human-produced Rock was frequently misidentified as AI-generated. Prior experience with AI music creation was associated with improved AI detection, indicating a familiarity effect. These results show that authorship judgements are shaped by sonic features, listener background and evolving AI capabilities, and that the direction of misattribution is an informative outcome. The study recommends genre-specific benchmarks for model evaluation, routine reporting of confusion matrices and longitudinal tracking so that AI development progress can be assessed against real listening behaviour. The study also supports institutional policy that pairs hands-on generation training with feature-level listening. Findings contribute to music cognition, media psychology and debates on creativity in an era of algorithmic production.

**Keywords:** AI-generated music, perceptual bias of music, human-AI authorship expectancy, *Suno AI*, *Udio AI*

## 1 Introduction

As systems using artificial intelligence (AI) become increasingly embedded in musical production (Deruty et al., 2022), questions of evaluation (Xiong et al., 2023), authorship (Pujari & Wilson, 2023) and pedagogy (Bell, 2025) have become important areas of research. In this context, the stakes in AI-human comparison extend beyond identification accuracy to the legitimacy of creative labour (Owen, 2025) and the standards by which listeners, educators and institutions evaluate musical worth (Hassink, 2024). As algorithmically produced art enters the mainstream (Deruty et al., 2022), audience conceptions of authenticity and cultural capital are affected (Hsu, 2025). Curricula and assessment practices in music education must adapt to new forms of production (Cheng, 2024), while cultural policy (DeVereaux et al., 2025) and copyright law grapple with originality and ownership (Surbhi & Roy, 2024). For researchers, shifting perceptual expectations (Hong et al., 2021) necessitate investigative studies designed to distinguish between familiarity effects and genuine gains in generative capability (Ma & Yu, 2025). As hybrid workflows evolve (White, 2025), integrating digital audio workstation (DAW) practices with generative systems, the boundary between composition and curation blurs (Vengathattil, 2025), reshaping how agency, intention and expertise are recognised across musical communities. Because stylistic perception is cultivated through cultural learning (Stevens, 2012), learned aesthetic preferences can bias judgements toward local genre norms, altering how music is evaluated (Susino & Schubert, 2019), whether AI-generated or human-produced.

These considerations motivate the present inquiry into how listeners hear and judge AI relative to human-produced music.

Within this broader context, AI is transforming how music is produced and evaluated (Wittschen, 2025). Music occupies a distinctive position in the arts because it unites computational structure with emotional expression (Aljanaki, 1987). Recent research suggests that AI-generated compositions can evoke emotional and aesthetic responses comparable to those elicited by human music (van Schaik, 2024), yet significant limitations persist. Although there is an “arms race” in computational detection of AI-generated music (Vila et al., 2025), perceptual judgement remains essential. Understanding how listeners perceive and distinguish AI-generated from human-produced works offers valuable insight into how audiences experience creativity in the age of intelligent systems (Lecamwasam & Chaudhuri, 2025).

Because generative systems still struggle to achieve coherence in formal structure, phrasing and expressive nuance (Hernandez-Olivan & Beltran, 2021), documenting these developments is important. Listener evaluation of these systems is key but can be influenced by demographic and experiential factors, such as formal musical training and prior exposure to and experience with AI technologies (Olayeni, 2023). A central concept in this process is perceptual authenticity, which concerns the extent to which listeners interpret musical gestures as the result of human intention and creative agency (van Schaik, 2024).

As Stevens (2012) shows, listener judgements are shaped by cultural learning. Through exposure and pedagogy, people develop genre norms for melody, rhythm and timbre that guide aesthetic evaluation (Guo et al., 2024). These expectations make familiar music easier to process, while habituation reduces sensitivity to cues (salient acoustic features such as timing, timbre, texture, ornamentation, and production artefacts) that are rare or expressed differently outside one’s home tradition. As a result, the same sonic feature may sound like an expressive nuance to one culture but like an artefact to another. In behavioural identification tasks, learned expectations can shift response criteria, and stimulus generalisation can lead to asymmetric misattributions when genre cues match a listener’s learned norms or when systems reproduce surface regularities without deeper constraints, consistent with habituation and fluency accounts of perception (Huron, 2006). Because the present participant pool comes from a single conservatory in China, cultural learning is a key consideration when interpreting accuracy and misattribution across genres.

On this basis, this study situates listener perception within a wider societal context when differences between AI and human authorship are evaluated. AI-human comparisons have become increasingly significant because they touch several interconnected topics: commercial trust (how audiences value authenticity in AI-assisted productions) (Hazzard et al., 2024; Drott, 2021); pedagogical design (how music education adapts to generative technologies) (Chen & Sun, 2024); creative attribution (how authorship and originality are understood) (Wang, 2025); and cultural policy (how societies define and regulate creativity in machine-made art) (DeVereaux et al., 2025). As Shank et al. (2023) demonstrate, listener perception is not only an aesthetic matter but also a social one, because individuals tend to evaluate music less favourably when they believe it to be AI-generated, even when the composition itself is identical to a human-produced recording. These considerations emphasise the need to examine perceptual authenticity empirically at a time when distinctions between human and machine creativity are increasingly blurred.

Research on listener perception of AI-generated music reveals a nuanced landscape in which outcomes vary by model, genre and methodology. A recurring pattern is that AI outputs are often judged as lacking emotional depth or expressive nuance. For example, White et al. (2025) report lower expressiveness ratings for AI than for human works, highlighting a gap between technical output and expectations of musical artistry. The same study shows that judgements of expressiveness and enjoyment depend not only on the sounds themselves but also on who listeners believe created the music, indicating a framing-driven bias that can shape reception irrespective of intrinsic musical qualities. At the same time, findings on bias are inconsistent: some studies report a tendency to evaluate AI-generated pieces less favourably (van Schaik, 2024), whereas others find minimal differences relative to human-produced work (Zenieris, 2023). Taken together, these results suggest that listener bias is contingent on contextual variables such

as prior exposure to AI technologies, the stylistic competence of the AI output, and beliefs about authorship.

Against this backdrop, bias can persist even when participants cannot reliably distinguish AI from human-produced content. Listeners often associate AI outputs with slower tempos, greater predictability and a “digital” or quantised quality (Shank et al., 2023). One plausible reason for this “digital” label is tokenisation. Tokenisation is defined as the segmentation of language into machine-readable units such as characters or words, which enables AI models to encode and generate sequences efficiently (Youvan, 2025). During training and generation, musical parameters including pitch, timing, tempo and control events are segmented into fixed tokens, a representation that promotes grid-aligned timing and short-term prediction (Kumar & Sarmento, 2023).

As AI systems have become increasingly sophisticated, distinguishing between AI- and human-produced music has grown more challenging, underscoring the need for methodological rigour in studies of perceptual identification and aesthetic evaluation (White et al., 2025). At the same time, longitudinal research is needed to track how these perceptions evolve: continued exposure to AI-generated music may attenuate initial biases, leading to greater acceptance and more balanced aesthetic judgements (White et al., 2025). As models improve and audiences become more accustomed to AI in creative domains, sustained research will be essential to clarify the long-term implications for music composition and reception.

This study is situated at a critical point of creative reflection, marking a transitional phase between traditional DAW production, fully prompt-driven AI composition, and emerging hybrid workflows that integrate both approaches (Ribeiro & Marins, 2024; Schwartz, 2025). Rather than emphasising industrial or commercial implications, the research focuses on the aesthetic and perceptual dimensions of this technological shift. Specifically, it investigates how trained musicians, ranging from undergraduate students to full professors, interpret, evaluate and attribute authorship to musical excerpts generated by the contemporary AI systems *Suno AI* and *Udio AI* (henceforth referred to simply as *Suno* and *Udio*) in comparison with human-produced compositions across six genres: Western Classical, Jazz, Rock, Pop, R&B and Chinese Traditional.

A total of 720 musical excerpts were curated and organised into 120 audio groups, each containing six tracks, one per genre. Participants listened to each audio group and made authorship judgements, identifying whether each track was AI-generated or human-produced without receiving any prior guidance or training.

At its core, this research is diagnostic, assessing perceptual thresholds in AI-human authorship identification, but it also carries an interpretive dimension, examining how aesthetic judgements evolve as creative technologies mature. By focusing on perceptual universality, the study seeks to document how people negotiate authenticity, agency and creativity during a period of profound technological transition.

## 1.1 Research Questions

Building on the framework of perceptual authenticity and aesthetic evaluation, this study addresses the following research questions:

1. Authorship Recognition: To what extent can participants with musical training accurately identify whether musical excerpts are AI-generated or human-produced, and what does this reveal about current perceptual thresholds of authenticity?
2. Genre Influences on Listener Perception: How does musical genre affect listeners’ ability to recognise authorship, evaluate expressive detail and perceive stylistic authenticity?
3. Experience and Technological Literacy: How does prior experience with AI music generation or computer-based composition affect the perceptual accuracy of AI-generated music?

4. Attribution Bias and Trust: Are listeners more likely to misattribute human-produced music as AI-generated or vice versa, and what do these attribution patterns suggest about expectation and aesthetic bias in AI-authored music?

## 1.2 Hypotheses

Drawing on existing research in computational creativity (Carnovalini & Rodà, 2020) and music cognition (Scherer & Zentner, 2008), the study advances the following hypotheses concerning listener accuracy, stylistic bias and experience:

- **H1:** Accuracy rates are expected to be higher in genres that exhibit a clear “hierarchy of structural significance” (Krumhansl, 1991) (e.g., Western Classical and Chinese Traditional), where formal organisation and tonal predictability aid in recognition. In contrast, improvisation-centred styles (e.g., Jazz, R&B) may yield lower accuracy because contemporary systems can convincingly emulate call-and-response and other surface conventions and have been reported to have passed the “Turing test for Jazz” improvisation, reducing human *versus* AI discriminability (Miller, 2020).
- **H2:** Misattributions of AI-generated music as human-produced are likely to occur more frequently in Pop and Rock, where AI models have demonstrated stronger emotional response with listeners (Chia et al., 2025) and are often perceived as producing compositions that feel “novel” or creatively distinctive (Carnovalini & Rodà, 2020).
- **H3:** Participants with computer music or AI music generation experience will show higher accuracy, consistent with evidence that emotional and evaluative responses are shaped by “exposure” and higher “perceptual fluency” (Scherer & Zentner, 2008).
- **H4:** Chinese Traditional music will show high, if not the highest, identification accuracy among genres because this group of participants possesses greater stylistic familiarity with this repertoire, and because music from the Global South is underrepresented in common AI training data (Mehta et al., 2025), increasing AI detectability for this genre.

Given that all participants were affiliated with a single Chinese conservatory (demographic details are provided in Section 2.1), cultural homogeneity may amplify genre-contingent effects. The author therefore interprets accuracy and misattribution in relation to cultural proximity and training histories, linking RQ1 and RQ2 to culturally learned listening habits regarding genre cues and RQ3 to experiential familiarity with AI-generated sounds. This framing clarifies why Chinese Traditional and Western Classical might yield higher discriminability in this cohort, and why Jazz and R&B may show lower AI detectability despite potential comparable surface fidelity.

## 1.3 Evaluative Framework

This study aligns with the Standardised Procedure for Evaluating Creative Systems (SPECS) (Jordanous, 2012), which provides a conceptual model for assessing creative performance in generative systems. Although the framework is not applied in its entirety, its three-part structure – which defines creativity, identifies evaluative criteria and establishes systematic methods of analysis – forms the foundation of the study’s methodological approach. In this context, creativity is understood as the perceived ability of a piece to be deemed original, emotionally expressive and stylistically comparable to human-produced music. Perceptual authenticity, therefore, functions both as a measure of creativity and as an indicator of AI’s ability plausibly to simulate emotionally and stylistically accurate music within the given genera. Participants were not informed of the author’s use of the SPECS framework; it informed *post-hoc* interpretation and the brief follow-up interviews only.

The study evaluates five key criteria adapted from creative systems assessment:

- **Aesthetic Appeal:** The extent to which participants find the music convincing and artistically cohesive.
- **Technical Proficiency:** The structural soundness of musical elements, including harmony, rhythm and timbre.
- **Stylistic Competence:** Adherence to genre-specific conventions and practices.
- **Emotional Expressivity:** The capacity to convey nuanced emotional content.
- **Novelty:** Inferred from misattributions, suggesting a level of synthesis beyond imitation.

Through this evaluative framework, the study aims to document how participants perceive authenticity, expressivity and creativity at a technological moment where the distinction between human and artificial music is becoming increasingly difficult to define.

## 2 Methodology

### 2.1 Participant Demographics

A total of 120 musicians from the Sichuan Conservatory of Music participated, including undergraduates, postgraduates and faculty aged 18 to 55. All participants had passed the Chinese Ministry of Education’s *Yikao* (Arts College Entrance Examination), which certifies “robust musical competency” (Lin & Weatherly, 2024). Participants were recruited through regular class sessions for students and through direct invitations to faculty members and institutional colleagues. Inclusion criteria were current enrolment or an academic appointment at the Conservatory; no additional exclusion criteria were applied. Demographic and background variables included age, gender, primary musical background, absolute pitch, and prior experience with AI-based music creation. Age distribution was 17–22 years, 62.5% (n = 75), 23–28 years, 16.67% (n = 20), 29–40 years, 10% (n = 12), and 41–55 years, 10.83% (n = 13). Gender was 55% male (n = 66) and 45% female (n = 54). Primary musical backgrounds were Jazz, 53.3% (n = 64), Pop, 30.8% (n = 37), Western Classical, 10.8% (n = 13) and Computer Music, 5% (n = 6). Prior experience creating AI music was 26.7% yes (n = 32) and 73.3% no (n = 88).

### 2.2 Materials

#### 2.2.1 Selection of AI Generation Software and Musical Genres

Audio samples were generated using two AI platforms, *Suno* and *Udio*, selected because they are among the most widely discussed and frequently analysed AI music generation systems in recent academic literature (Bown, 2025; Casini et al., 2025; Fiorino & Riera, 2025; Nayar, 2025; Rahman et al., 2024; Riedl, 2025; Vila et al., 2025; White et al., 2025). Human-produced music was sourced from *Spotify* and served as a control for the direct comparison of listener perception and identification accuracy. The dataset included 720 samples, with 240 from each source and 40 excerpts per genre, separated into 120 audio groups. The six genres studied were Western Classical, Jazz, Rock, Pop, R&B and Chinese Traditional.

These genres were selected for both stylistic breadth and cultural relevance. Western Classical was included for its pervasive presence in education and media, providing a familiar reference point. Jazz, together with R&B, was chosen for its improvisational character and its roots in African-American traditions, enabling comparison across culturally expressive styles outside the cohort’s home tradition. Rock and Pop were included as representative popular genres where AI generation has shown strong performance (Chia et al., 2025). Finally, Chinese Traditional anchors the set in the participants’ local cultural context, allowing examination of how listeners perceive AI interpretations of their own repertoire. Mehta et al. (2025) also note the near absence

of Global South music in AI training datasets, making its inclusion essential for evaluating cross-cultural limitations in generative models.

### 2.2.2 Standardisation of AI Output and Controlled Timeframe for Audio Collection

To maintain consistency in the audio examples, all AI-generated music was collected from 11–15 February 2025. This approach reduced the risk of variation caused by updates to the AI systems, such as changes in model architecture, introduction of new training data, or adjustments to model weights and parameters that could influence the nature of the generated music.

### 2.2.3 Controlled Recording Conditions

All audio examples were recorded directly from their respective source using *Audacity*'s loop-back feature to ensure stable sound quality across all sources (Berger & Neitsch, 2023), minimising potential variability from platform-specific download differences. Each track was exported as a stereo MP3 at 44,100 Hz and 320 kbps, with no changes made to volume, panning or dynamic range. Furthermore, no compression, noise reduction, or normalisation was applied, preserving the integrity of the original audio. All tracks were processed similarly to create a uniform and controlled listening environment for analysis.

## 2.3 AI Model Configurations

### 2.3.1 *Suno* AI Model: Configurations and Limitations

For all *Suno*-generated samples, the V4 model was used. Due to the 200-character prompt limit, prompts were carefully phrased to maximise stylistic accuracy. The instrumental toggle was enabled for Western Classical, Jazz and Chinese Traditional to avoid unintended vocals. The Classic Lyrics Model was applied to maintain coherence in instrumental output. For vocal genres including Rock, Pop and R&B, the Beta-phase ReMi Lyrics Model was employed. Although described as *Suno*'s newest and most creative model, ReMi could produce content that some listeners might find offensive because of its experimental status during the study.

### 2.3.2 *Udio* AI Model: Configurations and Limitations

For AI-generated samples using the *Udio* platform, version 1.5 was utilised with the udio-130 setting to create tracks two minutes and ten seconds in length. To maintain consistency in instrumental genres, including Western Classical, Jazz and Chinese Traditional, the Instrumental setting was selected to exclude vocals. For Rock, Pop and R&B, the Auto-Generate (Lyrics) option was enabled to incorporate vocals. The default settings were not adjusted to balance prompt accuracy and the naturalness of the output. The settings used are detailed below:

- Prompt Strength was set to 50%
- Lyrics Strength was set at 50%
- Clarity was set to 25%
- Generation Quality was set to “High”

### 2.3.3 Human-Produced Music Selection

Human-produced music was sourced from *Spotify Premium* at 320 kbps AAC to maintain audio quality comparable to the AI-generated tracks. Playback and recording conditions were standardised to avoid compression artefacts and preserve evaluation integrity. Because human selections required manual curation, the author applied a familiarity screening process to minimise awareness bias among Chinese conservatory students and faculty. Tracks showing clear public familiarity, such as prior chart prominence, enduring cultural presence, or high media exposure through film, television, or advertising, were excluded. Given the global circulation of Western media, these criteria were applied even when a Western track might be less familiar in China.

For the Chinese Traditional category, a colleague at the Sichuan Conservatory reviewed the song list to flag repertoire of high cultural significance or commonly recognised works and any flagged items were replaced with less prominent alternatives. Final inclusion decisions were based on the author’s professional experience and understanding of Chinese listening contexts.

Because *Spotify*’s recommendation algorithms can sometimes include AI-generated music in curated playlists (Kratochvil, 2025), additional verification steps were taken to confirm the human origin of every selected track. For each inclusion, the author documented the song title, artist and album, along with the associated *Spotify* playlist, as recorded in Appendix A. The playlists themselves, nine in total, were chosen for their stylistic alignment with the six genres used in this study and for their relative obscurity to participants.

Within some of the six selected genres, certain subgenres, such as Bubble-gum Pop, were incorporated to align more closely with the AI-generated outputs and to ensure stylistic comparability. While all tracks were manually reviewed, the author used *Spotify*’s internal tagging system to locate appropriate playlists. As Johnson (2020) notes, *Spotify*’s genre taxonomy is uneven and mainstream styles such as pop and hip hop are broadly categorised, while rock is generally subdivided into finer stylistic distinctions. These algorithmically mediated classifications influence visibility, discoverability and cultural framing and may also affect the likelihood that participants recognise particular tracks. The author, therefore, acknowledges that genre and sub-genre identifiers used in this study may not always correspond precisely to their conventional stylistic definitions.

To reduce subjectivity, “stylistic similarity” between AI outputs and human controls was assessed using a predefined rubric based on music-information-retrieval and music-theoretical accounts of stylistic features. The rubric, as outlined by Ching-Hua Chuan (2013), considered (1) instrumentation and texture; (2) timbre at the mix level; (3) harmonic vocabulary and progression; (4) rhythmic profile, metre and tempo range; and (5) melodic contour and register. In applying this rubric, the AI excerpts were analysed holistically and human-produced control tracks were selected to reduce subjectivity and to align the selection procedure with established song-level style-identification practice.

Together, these curation procedures ensured that all human-produced examples were authentic in origin, appropriately matched in style to the AI-generated material, and culturally balanced for valid cross-cultural perceptual comparison.

## **2.4 Prompt Engineering**

### **2.4.1 Standardised Prompts**

The study employed standardised prompt language to control variables across *Suno* and *Udio* while allowing each model’s internal stylistic algorithms to operate freely. Prompts were formulated abstractly to align with how generative AI systems categorise musical information by style, texture and emotional intent rather than by prescriptive structural instruction (Rickard, 2022). In each case, the author allowed the AI models to determine the natural duration of the generated piece and then trimmed outputs to the target length for consistency. All prompts were written in English and up to three rounds of pilot testing were conducted to refine the language before the final prompt selection.

Across both platforms, the same prompt text was used verbatim for each genre, with no modifications to syntax or descriptive terms. Although genre-specific input noticeably differed in style, their prompt structures and phrasing remained constant to isolate model behaviour rather than prompt variation. No stylistic balancing was imposed, because the intention was to have genre distinctiveness.

To verify consistency, the author conducted listening tests to ensure that outputs from *Suno* and *Udio* were comparable in timbre, tempo range and general stylistic fit. It should be noted that the Chinese Traditional genre required additional prompt refinement, supporting the argument that there is limited representation of non-Western music in current AI training data (Mehta et al., 2025).

### 2.4.2 Prompt Examples

Western Classical: Create a western classical piece in the style of the 1700s–1800s with period instruments such as harpsichord, fortepiano, violin, cello, or full orchestra.

Jazz: Create a mid-1960s jazz piece featuring saxophone, trumpet, trombone, piano, bass, and drums. Capture rich harmonies, swing or modal grooves, expressive solos, a hard bop sound, and improvisation.

Rock: Create a 1970s–1990s rock song with electric guitar, bass, drums, and either male or female vocals. Incorporate powerful riffs, energetic rhythms, and strong vocals, ranging from classic rock and hard rock to grunge and alternative rock.

Pop: Create a modern pop song featuring male or female vocals, catchy melodies, upbeat rhythms, and polished production. Use synthesisers, guitars, bass, and drums to craft a radio-friendly sound with strong hooks.

R&B: Create an R&B track featuring male or female vocals, smooth melodies, and soulful rhythms. Include background harmonies, lush chords, and a groovy rhythm section, drawing inspiration from both classic and modern R&B styles.

Chinese Traditional: Create a Chinese traditional piece incorporating guzheng, erhu, pipa, or dizi. Use strong melodies, ornamentation, pentatonic scales, traditional rhythms, and Chinese percussion to achieve cultural authenticity.

## 2.5 Output Standardisation and Excerpt Length

Although the AI platforms could generate full-length compositions exceeding four minutes, this study used 30–45-second excerpts that were manually edited by the author in *Audacity*, a decision informed by Strauss et al. (2024), whose multi-genre listening protocol employed excerpts averaging 40.8 seconds. The selected duration was chosen to balance statistical reliability and perceptual accuracy and was deemed to possess sufficient musical information for valid comparison while minimising fatigue and maintaining experimental brevity. Each excerpt began at a randomised point within the track; however, when a randomly selected segment did not include an internal change in texture, the start time was adjusted to a different point on the track that did include a texture change. This ensured that all excerpts contained two distinct textural sections while avoiding systematic selection of choruses or other markers that could facilitate recognition of the form. For vocal tracks, because lyrics can cue form, randomised start points were used to avoid chorus-only bias; choruses could occur by chance but were neither targeted nor systematically included or excluded. This design encouraged participants to focus on tempo, timbre and stylistic complexity rather than formal structure, consistent with Balkwill and Thompson’s (1999) findings on cross-cultural emotion perception, thereby allowing for a more controlled evaluation of the perception of authenticity across genres. All clips featured two-second fade-ins and five-second fade-outs to create smooth transitions and minimise distraction.

## 2.6 Experimental Focus on Musical Elements

This approach aimed to minimise the influence of contextual and structural cues on participant responses. The experiment was designed around what Cambouropoulos (2010) refers to as the “musical surface”, encompassing elements such as beat, metre, grouping and harmonic organisation that emerge through the interaction of perceptual, cognitive and structural processes. Given the non-standardised nature of the tokenisation of musical elements (White, 2025) and the use of randomised excerpts to probe current systems, this method is likely to reveal detectable artefacts resulting from the encoding process. Although participants were not explicitly informed of this design intention, the selection and preparation of materials reflected this framework. Since prior research indicates that AI-generated music often lacks emotional depth and contextual expressivity due to its reliance on statistical pattern imitation rather than intentional phrasing (Tariq et al., 2022), limiting the amount of emotional information became an essential aspect of the experimental design.

## 2.7 Presentation and Grouping of Excerpts

Tracks were randomly divided into the 120 audio groups from *Udio*, *Suno* and *Spotify* using the `RAND()` function in Microsoft *Excel*, producing an even distribution of 240 tracks per source. While authorship was randomised within each genre, the genre sequence remained fixed for all participants to control for ordering effects while maintaining stylistic continuity. All tracks were presented without identifying information, keeping participants blind to the production source. Listening sessions were conducted in a controlled environment using the researcher’s personal computer and Sony MDR-7506 headphones, with playback through *VLC Media Player* and all enhancements disabled. Volume levels were standardised and remote participation was not permitted to ensure consistent conditions.

Participants engaged in stand-alone listening sessions, hearing one unique audio group of six tracks without repetition or direct comparison across styles. This design emphasised independent judgement and minimised contextual bias, focusing attention on the intrinsic musical features of each excerpt.

## 2.8 Experimental Procedure

The experiment ran from 3 March to 5 April 2025 at the New District and Main Campuses of the Sichuan Conservatory of Music in Chengdu, China. Participants completed the listening task either during class sessions or individually in faculty offices; both settings provided quiet, controlled conditions and no location differences were observed. To preserve the blind design, no familiarisation trials were provided, and participants received no information about production sources, styles, or the proportion of AI *versus* human tracks. Participants could request one re-listen per track; only seven individuals did so, and no additional re-listens were permitted after a choice was recorded. Peer discussion was not restricted since randomisation and blinding minimised the risk of systematic influence. Responses were collected using printed questionnaires to maintain environmental control. All anonymised data, including responses, track lists and group assignments, are available via a stable *Google Drive* link in Appendix A for transparency and reproducibility.

## 2.9 Justification of Listening Order

Given the limited literature on ordering heterogeneous musical styles for listening tasks, the author adopted a fixed “outside-in” sequence to balance familiarity and anticipated task difficulty. Western Classical and Chinese Traditional occupied positions one and six, to anchor the session with styles most familiar to the cohort. Positions two and five were assigned to Jazz and R&B, respectively, paired as improvisation-focused idioms whose expressive variability may obscure identification. The middle positions, three and four, were reserved for Rock and Pop, genres, respectively, for which AI generation tools often achieve strong stylistic emulation (Chia et al., 2025), thereby situating what is assumed to be the most challenging styles at the midpoint. This ordering is reported transparently as a pragmatic, pedagogically informed design choice and is acknowledged as a limitation considering the paucity of literature on style ordering for this type of experimental design.

## 2.10 Post-Listening Qualitative Interview Procedure and Analysis

After the listening task, participants who achieved a perfect score (6/6) (participants 7, 33, 74 and 91) were invited to brief interviews to explain their decision-making processes. Interviews were conducted as guided discussions rather than a formal protocol, with prompts organised according to the SPECS framework and aimed at eliciting comments on micro-timing, phrasing and perceived emotional authenticity. All interviews were conducted in Mandarin Chinese by the author, who was aware that these interviewees had achieved perfect scores but participants did not receive feedback on accuracy during the interview. During each session, the author took contemporaneous notes in English; no translation step was required.

Qualitative analysis proceeded as follows. The author served as the sole coder, applying an inductive, theme-first approach to the English notes. Initial open coding identified recurrent cues and evaluation criteria referenced by interviewees; axial coding then grouped codes into

higher-order themes aligned with the study’s constructs of timing, articulation and phrasing, timbre and mix, and perceived emotional authenticity. Theme definitions and representative excerpts were recorded in an audit trail. No second-coder reliability was computed; this is noted as a limitation. To mitigate single-coder bias, coding decisions were revisited six months after the interview and checked against session notes. The interviewer’s role and prior knowledge are reported here for transparency.

## 2.11 Methodology for Statistical Analysis

Analyses were conducted at the trial level, with participants having a binary choice of classifying each excerpt as human or AI. The primary outcome was identification accuracy, defined as the proportion correct within prespecified cells (production category, model, genre, and participant subgroups). Accuracy was estimated with two-sided 95% binomial confidence intervals and visualised against a 0.50 chance benchmark. Within-cell deviation from chance was tested using exact one-sample binomial tests (two-sided). Dependence of accuracy on model and genre was evaluated via Pearson chi-square tests of independence on contingency tables, with Cramér’s V reported as the effect size. Error structure was characterised using confusion matrices separating AI→Human and Human→AI misattributions by genre. Participant attributes were examined by stratifying accuracy and corresponding 95% confidence intervals by computer-music training and AI music generation experience. All tests used  $\alpha = 0.05$ , and figures and tables report counts, proportions, confidence intervals and  $p$ -value annotations corresponding to these procedures only. All analyses were conducted using *Python*.

### 2.11.1 Interpretation and Significance of $p$ -Value Results

While there is a growing consensus in the scientific and medical communities to adopt a more stringent threshold for statistical significance ( $p < 0.005$ ), as proposed by many, including Ioannidis (2018), this study retains the conventional  $p < 0.05$  criterion. This threshold is appropriate given the exploratory nature of the present work, which concerns perceptual judgement rather than clinical or high-stakes outcomes.

Also, in the case of this experiment, a high  $p$ -value does not necessarily indicate that the experiment was unsuccessful. As Betensky (2019) explains,  $p$ -values must be interpreted within the context of the study design, including the sample size and the meaningful effect size. In this research, which investigates listener identification of musical authorship, a low  $p$ -value suggests that responses differed significantly from random guessing, indicating either a higher rate of accurate or inaccurate identifications at a statistically significant rate. In contrast, a high  $p$ -value reflects results close to the null hypothesis, showing that listeners found the AI-generated or human-produced music indistinguishable. Therefore, within this framework, a high  $p$ -value should be interpreted cautiously because it suggests chance-level identification in the aggregate, indicating indeterminacy in authorship judgements rather than affirmative evidence of human-like quality.

### 3 Results

#### 3.1 Broad Overview of Results by Production Method

Fig. 1 and Tab. 1 show an overview of results by production method.

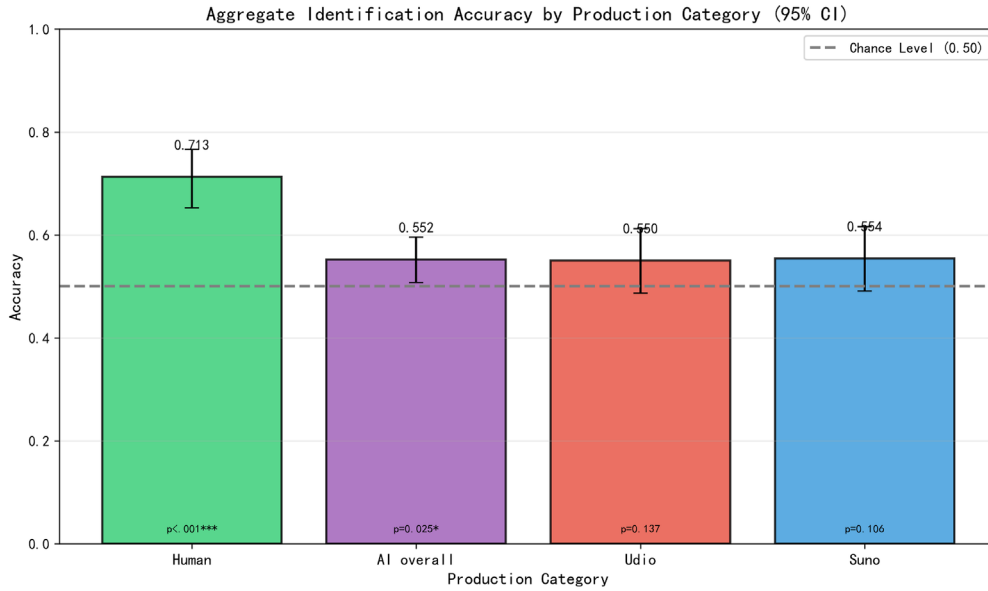


Figure 1: Broad overview of results by production method.

Table 1: Broad overview of results by production method.

Production Method	Correct	Total	Accuracy	95% CI	$p$ -value (vs 0.50)
Human	171	240	71.25 %	[0.652, 0.766]	3.59E-11
AI overall	265	480	55.21 %	[0.507, 0.596]	0.0252
Udio	132	240	55.00 %	[0.487, 0.612]	0.1375
Suno	133	240	55.42 %	[0.491, 0.616]	0.1064
Chi-square: Human vs AI overall				$\chi^2(1) = 16.57, p = 0.0000$	Cramér's V = 0.152

Across aggregate categories, identification accuracy exceeded the 0.50 chance level. Human excerpts were identified most accurately (71.25%), significantly above chance ( $p = 3.59 \times 10^{-11}$ ) and significantly higher than AI overall ( $\chi^2(1) = 16.57, p < 0.001$ ). Although this difference is statistically significant, the effect size is small (Cramér's V = 0.152), indicating a weak association and substantial overlap between conditions. AI overall was modestly above chance (55.21%,  $p = 0.0252$ ). *Udio* (55.00%,  $p = 0.1375$ ) and *Suno* (55.42%,  $p = 0.1064$ ) were likewise above the 0.50 threshold but did not differ from chance at significance levels.

### 3.2 Broad Overview of Results by Genre

Fig. 2 and Tab. 2 show an overview of results by genre.

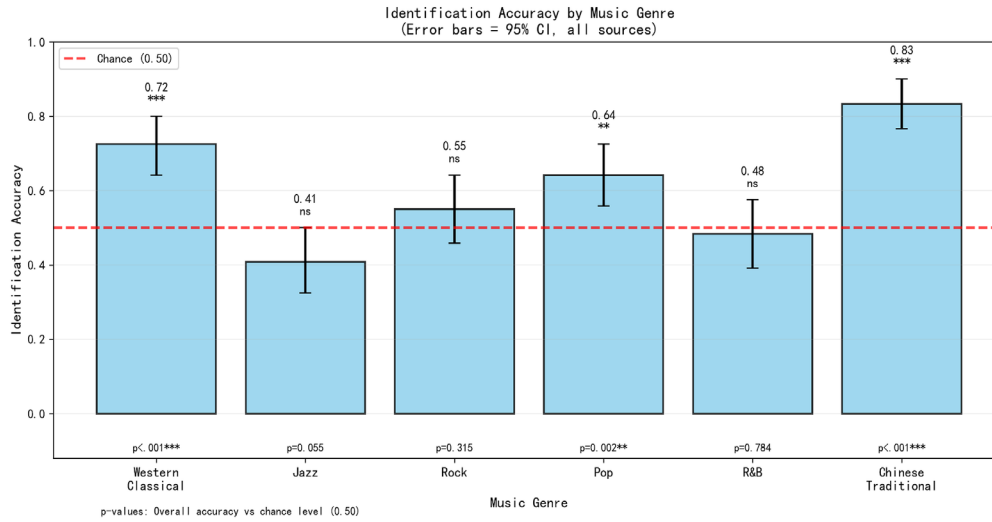


Figure 2: Broad overview of results by genre.

Table 2: Detailed results by genre.

Genre	N =	Correct	Accuracy	p-value
Western Classical	120	87	72.50%	0
Jazz	120	49	40.80%	0.0548
Rock	120	66	55.00%	0.3153
Pop	120	77	64.20%	0.0024
R&B	120	58	48.30%	0.7843
Chinese Traditional	120	100	83.30%	0

Results by genre showed identification above chance for Western Classical (72.5%,  $p < 0.001$ ), Pop (64.2%,  $p = 0.0024$ ) and Chinese Traditional (83.3%,  $p < 0.001$ ); Rock hovered modestly above chance without significance (55.0%,  $p = 0.3153$ ); Jazz fell below chance but did not reach significance (40.8%,  $p = 0.0548$ ); and R&B was indistinguishable from chance (48.3%,  $p = 0.7843$ ).

### 3.3 Model-Genre Rates of Identification Accuracy

Fig. 3 shows model-genre rates of identification accuracy and Tab. 3 shows detailed results by genre and production method.

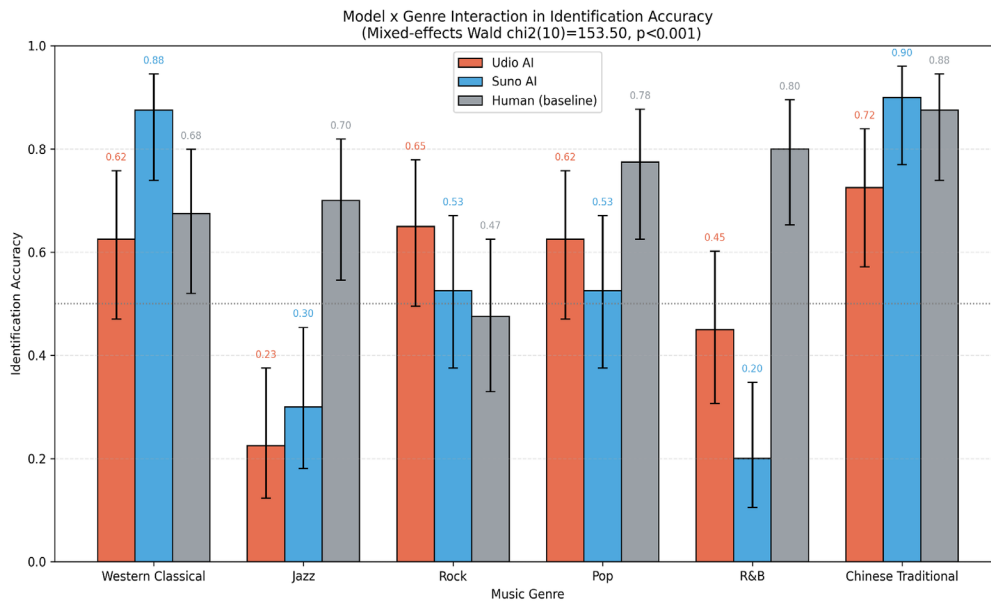


Figure 3: Model-genre rates of identification accuracy.

Table 3: Detailed results by genre and production method.

*Udio AI*

Genre	N =	Correct	Accuracy	CI 95	<i>p</i> -value
Western Classical	40	25	62.50%	[0.470, 0.758]	0.1539
Jazz	40	9	22.50%	[0.123, 0.375]	0.0007
Rock	40	26	65.00%	[0.495, 0.779]	0.0807
Pop	40	25	62.50%	[0.470, 0.758]	0.1539
R&B	40	18	45.00%	[0.307, 0.602]	0.6358
Chinese Traditional	40	29	72.50%	[0.572, 0.839]	0.0064

*Suno AI*

Genre	N =	Correct	Accuracy	CI 95	<i>p</i> -value
Western Classical	40	35	87.50%	[0.739, 0.945]	0
Jazz	40	12	30.00%	[0.181, 0.454]	0.0166
Rock	40	21	52.50%	[0.375, 0.671]	0.8746
Pop	40	21	52.50%	[0.375, 0.671]	0.8746
R&B	40	8	20.00%	[0.105, 0.348]	0.0002
Chinese Traditional	40	36	90.00%	[0.769, 0.960]	0

Human Produced

Genre	N =	Correct	Accuracy	CI 95	<i>p</i> -value
Western Classical	40	27	0.675	[0.520, 0.799]	0.0385
Jazz	40	28	0.7	[0.546, 0.819]	0.0166
Rock	40	19	0.475	[0.329, 0.625]	0.8746
Pop	40	31	0.775	[0.625, 0.877]	0.0007
R&B	40	32	0.8	[0.652, 0.895]	0.0002
Chinese Traditional	40	35	0.875	[0.739, 0.945]	0

Broadly, performance depends on both model and genre (model  $\times$  genre interaction:  $\chi^2(10) = 25.80$ ,  $p = 0.004$ , Cramér's  $V = 0.17$ ). Although this interaction is statistically significant, the effect size is small, as indicated by Cramér's  $V$ , meaning that model-by-genre differences are reliable but modest. Human-produced excerpts were identified above the 0.50 threshold in every genre except Rock: Western Classical 67.5% ( $p = 0.0385$ ), Jazz 70.0% ( $p = 0.0166$ ), Pop 77.5% ( $p = 0.0007$ ), R&B 80.0% ( $p = 0.0002$ ), Chinese Traditional 87.5% ( $p = 1.38 \times 10^{-6}$ ), with Rock at 47.5% ( $p = 0.8746$ ). For AI excerpts, higher scores indicate easier detection as AI. Detectability was highest in Chinese Traditional and Western Classical, especially for *Suno*: Chinese Traditional *Suno* 90.0% ( $p = 1.86 \times 10^{-7}$ ) and *Udio* 72.5% ( $p = 0.0064$ ); Western Classical *Suno* 87.5% ( $p = 1.38 \times 10^{-6}$ ) and *Udio* 62.5% ( $p = 0.1539$ ). Rock and Pop hovered near chance for both models (Rock: *Udio* 65.0%,  $p = 0.0807$ ; *Suno* 52.5%,  $p = 0.8746$ ; Pop: *Udio* 62.5%,  $p = 0.1539$ ; *Suno* 52.5%,  $p = 0.8746$ ). Performance was weakest in Jazz and also low for *Suno* in R&B (Jazz: *Udio* 22.5%,  $p = 0.0007$ ; *Suno* 30.0%,  $p = 0.0166$ ; R&B: *Suno* 20.0%,  $p = 0.0002$ ; *Udio* 45.0%,  $p = 0.6358$ ). In sum, identification of human music leads in most genres, while AI detectability is genre-contingent: it is high in Chinese Traditional and Western Classical, mixed in Rock and Pop, and comparatively low in Jazz and in R&B.

### 3.4 Misattribution Direction by Genre

Fig. 4 shows misattribution direction by genre and Tab. 4 shows detailed results of misattribution direction by genre.

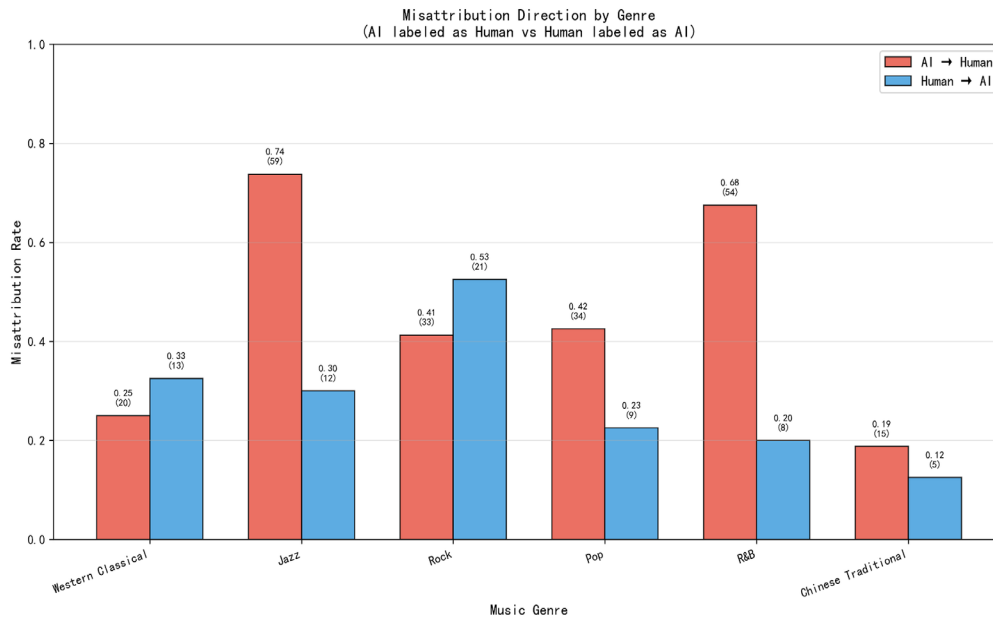


Figure 4: Misattribution direction by genre.

Table 4: Detailed results of misattribution direction by genre.

Genre	AI→AI (correct)	AI→Human	Human→Human (correct)
Western Classical	60	20	27
Jazz	21	59	28
Rock	47	33	19
Pop	46	34	31
R&B	26	54	32
Chinese Traditional	65	15	35
TOTAL	265	215	172

Continued:	Human→AI	Delta (AI→H minus H→AI)	Total
	13	7	120
	12	47	120
	21	12	120
	9	25	120
	8	46	120
	5	10	120
	68	147	720

Genre	$\chi^2$	$p$ -value	Cramér's V
Western Classical	0.4232	0.5153	0.0594
Jazz	19.3547	0.0000	0.4016
Rock	0.9470	0.3305	0.0888
Pop	3.8100	0.0509	0.1782
R&B	22.2289	0.0000	0.4304
Chinese Traditional	0.3675	0.5444	0.0553

Across genres, misattributions were asymmetric: participants mislabelled AI as human far more often than they mislabelled human as AI (AI→Human = 215 *versus* Human→AI = 68;  $\Delta$  = 147). This asymmetry was strongest in Jazz and R&B, where chi-square tests were significant with medium effect sizes (Jazz:  $\chi^2 = 19.35$ ,  $p < 0.001$ , Cramér's V = 0.40; R&B:  $\chi^2 = 22.23$ ,  $p < 0.001$ , V = 0.43). Pop showed a smaller, borderline pattern ( $\chi^2 = 3.81$ ,  $p = 0.051$ , V = 0.18). Western Classical, Rock and Chinese Traditional exhibited no reliable asymmetry, with overall lower rates of misattribution (all  $p > 0.33$ ). In broad terms, AI tracks were most often mistaken for humans in Jazz and R&B, occasionally in Pop, and much less so in Western Classical, Rock and Chinese Traditional.

### 3.5 Expertise and Experience Effects on AI Identification Accuracy

Fig. 5 shows expertise and experience effects on AI identification accuracy and Tab. 5 shows detailed results of expertise and experience effects on AI identification accuracy.

Expertise Effects on AI Identification Accuracy (95% CI)  
 AI overall: Training  $\chi^2(1)=0.90$ ,  $p=0.343$ ,  $V=0.04$ ; Experience  $\chi^2(1)=4.17$ ,  $p=0.041$ ,  $V=0.09$   
 Mixed model: ComputerMusic  $\beta=0.76$ ,  $p=0.080$ ; AIExperience  $\beta=0.05$ ,  $p=0.764$

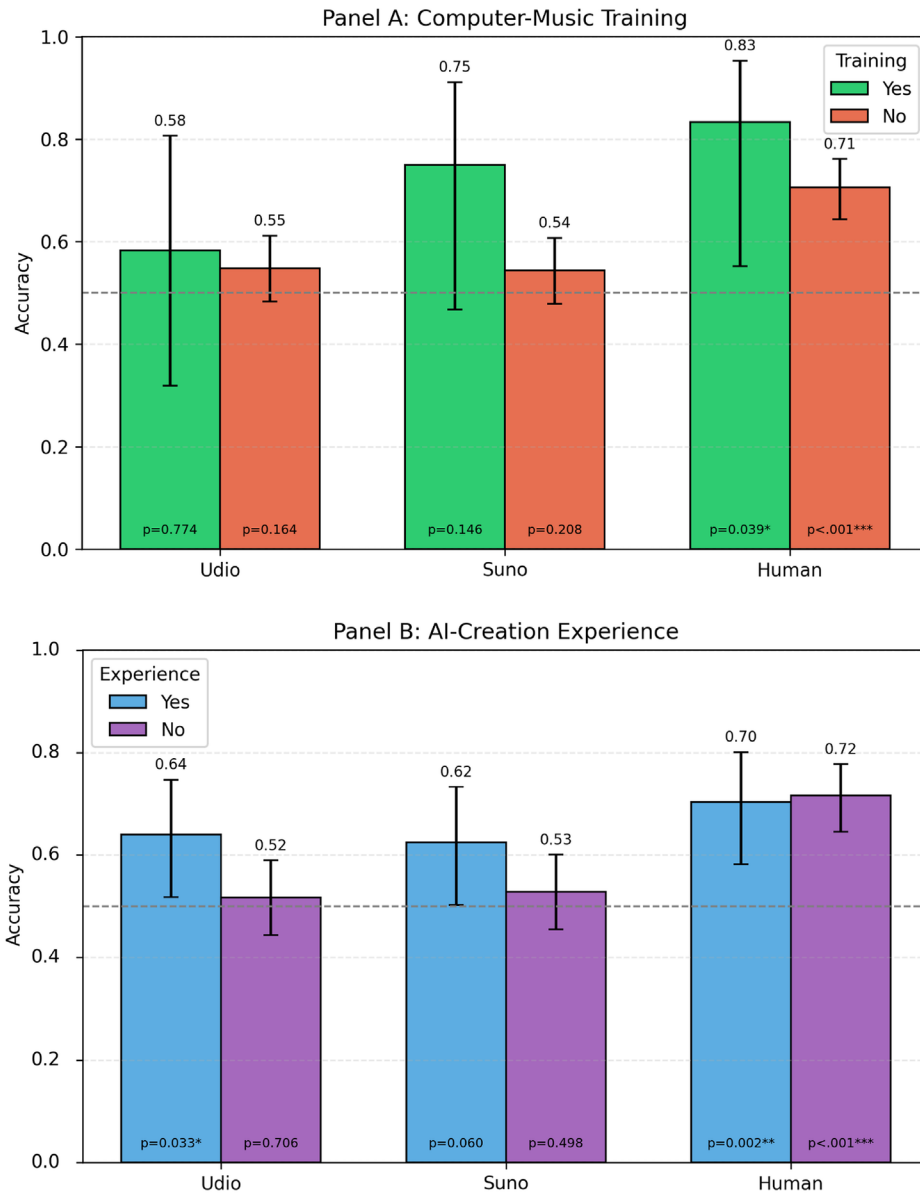


Figure 5: Expertise and experience effects on AI identification accuracy.

Table 5: Detailed results of expertise and experience effects on AI identification accuracy.

Panel	Group	N Participants	Production Method	Correct	Total	Accuracy	95% CI	p-value
A: Computer-music training	Yes	6	AI Combined	16	24	66.67%	[0.467, 0.820]	0.1516
A: Computer-music training	Yes	6	Udio	7	12	58.33%	[0.320, 0.807]	0.7744
A: Computer-music training	Yes	6	Suno	9	12	75.00%	[0.468, 0.911]	0.1460
A: Computer-music training	Yes	6	Human	10	12	83.33%	[0.552, 0.953]	0.0386
A: Computer-music training	No	234	AI Combined	249	456	54.61%	[0.500, 0.591]	0.0547
A: Computer-music training	No	234	Udio	125	228	54.82%	[0.483, 0.612]	0.1642
A: Computer-music training	No	234	Suno	124	228	54.39%	[0.479, 0.607]	0.2082
A: Computer-music training	No	234	Human	161	228	70.61%	[0.644, 0.761]	4.17E-10
B: AI-creation experience	Yes	32	AI Combined	81	128	63.28%	[0.547, 0.711]	0.0034
B: AI-creation experience	Yes	32	Udio	41	64	64.06%	[0.518, 0.747]	0.0328
B: AI-creation experience	Yes	32	Suno	40	64	62.50%	[0.503, 0.733]	0.0599
B: AI-creation experience	Yes	32	Human	45	64	70.31%	[0.582, 0.801]	0.0016
B: AI-creation experience	No	88	AI Combined	184	352	52.27%	[0.471, 0.574]	0.4240
B: AI-creation experience	No	88	Udio	91	176	51.70%	[0.444, 0.590]	0.7064
B: AI-creation experience	No	88	Suno	93	176	52.84%	[0.455, 0.601]	0.4976
B: AI-creation experience	No	88	Human	126	176	71.59%	[0.645, 0.777]	9.34E-09
A: Computer-music training	Chi-square test		Yes vs No				$\chi^2(1) = 0.90$ , $p = 0.3433$	Cramér's V = 0.043
B: AI-creation experience	Chi-square test		Yes vs No				$\chi^2(1) = 4.17$ , $p = 0.0413$	Cramér's V = 0.093

Identification accuracy showed only modest dependence on expertise. Although participants with computer-music training displayed higher accuracy descriptively, this difference was not statistically reliable when trained and untrained participants were compared ( $\chi^2(1) = 0.90$ ,  $p = 0.343$ ), and the effect size was very small (Cramér's  $V = 0.043$ ). In contrast, prior AI-creation experience was associated with better AI detection: participants with experience were more accurate in identifying AI tracks than those without, but also had a small effect size (AI combined 63.3% *versus* 52.3%;  $\chi^2(1) = 4.17$ ,  $p = 0.041$ ,  $V = 0.093$ ). Within models, the experienced group exceeded chance for *Udio* and for AI combined and trended above chance for *Suno*, while the non-experienced group hovered near chance on AI tracks. Human identification remained high in both groups.

## 4 Discussion

This study examines four questions: (RQ1) how accurately musically trained listeners identify AI-generated *versus* human-produced excerpts (authorship recognition); (RQ2) how genre shapes authorship judgements; (RQ3) how prior experience with AI music or computer-based composition affects detection accuracy; and (RQ4) whether misattribution skews toward Human→AI or AI→Human and what this implies for expectation and trust. Two conditions frame

the interpretation: participants received no evaluative instructions, so responses reflect intuitive listening; and brief post-listening interviews supply qualitative context on the emotional criteria and cognitive judgements listeners reported.

#### 4.1 Evaluation of Hypotheses

**H1. Structural Predictability And Recognition Accuracy:** Supported. As noted in Fig. 2, genres with clearer hierarchical structure showed higher identification accuracy, with Western Classical at 72.5% ( $p < 0.001$ ) and Chinese Traditional at 83.3% ( $p < 0.001$ ). Improvisation-forward styles showed lower accuracy, with Jazz at 40.8% ( $p = 0.0548$ ) and R&B at 48.3% ( $p = 0.7843$ ). These outcomes align with the prediction that tonal and formal predictability facilitates recognition and are consistent with Miller’s (2020) “Turing test for jazz” account of improvisation.

**H2. Misattribution in Pop and Rock:** Partially supported. As noted in Tab. 4, misattribution of AI-generated music as human occurred frequently in mainstream styles: in Pop, 42.5% of AI trials were labelled human; in Rock, 41.25%. Human Rock excerpts were correctly identified at only 47.5%, indicating elevated ambiguity of authorship in these genres. H2 is marked as partially supported since Jazz and R&B, both music with prominent improvisation and of an African-American tradition, showed higher rates of misattribution: Jazz, 73.75% and R&B, 67.5%, showing that the AI-generated music in these styles was more convincing than human-produced music to the participant pool. Notably, for the aggregate results of the misattribution direction by genre, both Jazz and R&B had a  $p$ -value of 0 with a  $\chi^2$  of 19.3547 and 22.2289, respectively, and a Cramér’s  $V$  of 0.4016 and 0.4304, respectively, denoting a moderate association.

**H3. AI Music Generation Experience:** Partially supported. As shown in Fig. 5 and Tab. 5, participants with AI music-creation experience identified AI tracks more accurately than those without (AI-combined 63.28% versus 52.27%;  $\chi^2(1) = 4.17$ ,  $p = 0.0413$ , Cramér’s  $V = 0.093$ ), with higher point estimates for *Udio* (64.06% versus 51.70%) and *Suno* (62.50% versus 52.84%). Human-track accuracy remained high in both groups. By contrast, computer-music training yielded only a non-significant trend toward higher accuracy ( $\chi^2(1) = 0.90$ ,  $p = 0.3433$ ,  $V = 0.043$ ). These results align with exposure-based accounts of perceptual fluency for AI-creation experience, but provide insufficient evidence for a comparable advantage from formal computer-music training. To clarify causality and adaptation effects, larger longitudinal studies with dedicated participants are needed.

**H4. Chinese Traditional Music Detectability:** Supported. As shown in Fig. 3 and Tab. 3, Chinese Traditional produced the highest identification accuracy at 83.3% ( $p < 0.001$ ), significantly above chance and exceeding Western Classical at 72.5% ( $p < 0.001$ ) and all other genres. Both AI models were most readily detected in this genre, with *Suno* at 90.0% ( $p = 1.86 \times 10^{-7}$ ) and *Udio* at 72.5% ( $p = 0.0064$ ), while human excerpts were also correctly identified at high rates, 87.5% ( $p = 1.38 \times 10^{-6}$ ). These outcomes are consistent with the proposed mechanisms of greater participant familiarity with their home tradition (Huron, 2006), and the noted underrepresentation of Global South repertoires in common AI training data (Mehta et al., 2025).

#### 4.2 Application of the SPECS Methodology

This study aligns with Jordanous’s (2012) Standardised Procedure for Evaluating Creative Systems (SPECS) by assessing generative AI systems against perception-based indicators of creativity. Here, creativity is operationalised as perceptual authenticity, namely the extent to which listeners perceive AI-generated music as human-produced. Evaluation centred on stylistic fidelity and emotional expressivity across the six genres, using five adapted criteria grounded in post-listening interview evidence from participants 7, 33, 74 and 91, each of whom achieved a perfect score (6/6) on the listening task (see Appendix B).

- **Aesthetic appeal:** Interviewees frequently reported low aesthetic engagement, for example describing Pop and Rock as “boring”, and limited appeal where drums felt overly uniform.

- **Technical proficiency:** Reports identified overly regular micro-timing, identical snare hits and residual high-frequency artefacts that would normally be rolled off in post-production, these possibly due to the tokenisation techniques employed by the models.
- **Stylistic competence:** Several excerpts were described as MIDI-like rather than performed, and Chinese Traditional tracks were judged to lack characteristic rhythms and idiomatic pitch-bending, also most likely due to tokenisation.
- **Emotional expressivity:** Pop and R&B were often said to convey positive affect, whereas Jazz and Chinese Traditional were judged to lack convincing emotional nuance compared with expectations.
- **Novelty:** Interviewees generally reported that nothing sounded particularly novel, despite surface-level stylistic competence.

A practical benchmark underlies this framework. If expert or trained listeners cannot reliably distinguish AI-generated from human-produced music, the AI may be said to attain perceptual creativity in listener reception. However, this form of creativity is better understood through Boden’s concept of psychological creativity (P-creativity), in which an idea is novel to the agent that produces it; as opposed to historical creativity (H-creativity), which requires the production of something genuinely new in a broader historical sense (Boden, 2004). In the present context, AI systems generate outputs that may be novel within the internal logic of their learned processes; however, because these outputs are derived from training data based on existing musical works (Mycka & Mańdziuk, 2024), they cannot be regarded as historically creative. Correct identification rate therefore serves as a quantitative indicator of authenticity perception in relation to P-creativity, while interview evidence clarifies which dimensions, such as micro-timing regularity, timbral artefacts, and idiomatic gesture, most strongly shaped listeners’ judgements.

### 4.3 Interpreting Statistical Significance and Small Effect Sizes

Across analyses, several comparisons reached statistical significance, but the associated effect sizes indicate modest practical differences. As shown in Fig. 1 and Tab. 1, the aggregate Human *versus* AI comparison was statistically significant ( $\chi^2(1) = 16.57, p < 0.001$ ), yet the effect size was small (Cramér’s  $V = 0.152$ ). This pattern suggests that, while the data provide strong evidence of a difference, the practical separation between human- and AI-produced excerpts is limited: many excerpts remained difficult to classify, with substantial overlap and confusability across production methods and genres. Accordingly, these findings are best interpreted as demonstrating limited aggregate discriminability rather than a strong categorical distinction between human and current AI music. Similarly, Fig. 3 and Tab. 3 show a statistically significant model  $\times$  genre interaction ( $\chi^2(10) = 25.80, p = 0.004$ ), again with a small effect size (Cramér’s  $V = 0.17$ ), indicating that model-by-genre differences are reliable but modest and are better understood as uneven shifts in detectability across genres rather than a consistent separation between conditions. Finally, Fig. 5 and Tab. 5 indicate that expertise-related effects were small in practical terms. The computer-music training comparison (Yes *versus* No) was not statistically significant ( $\chi^2(1) = 0.90, p = 0.3433$ ) and showed a very small association (Cramér’s  $V = 0.043$ ). AI-creation experience (Yes *versus* No) reached statistical significance ( $\chi^2(1) = 4.17, p = 0.0413$ ), but the effect size remained very small (Cramér’s  $V = 0.093$ ), indicating only a modest shift in AI identification accuracy. Overall, prior experience contributed limited explanatory power relative to broader sources of variability (e.g., genre- and excerpt-level confusability), and the computer-music training result should be interpreted cautiously given the small number of trained participants ( $n = 6$ ). Taken together, these small effect sizes suggest that statistically reliable differences in this study translate into only modest practical advantages for listeners. In other words, identification performance appears to be shaped more by excerpt- and genre-specific features (and the resulting confusability between human and AI outputs) than by broad categorical factors such as model type or participant expertise alone. Future research should examine musical training and related forms of expertise using more targeted, adequately pow-

ered designs and more refined pre-listening questionnaires (e.g., years of training, primary studied genre, and listening/production experience) to clarify whether specific training profiles meaningfully influence AI detection performance.

#### 4.4 Broader Implications of AI-Generated Content

As audiences grow accustomed to algorithmically shaped musical aesthetics, standards for creativity and authenticity shift and authorship becomes less a matter of source and more an ambiguous judgement. Frequent misidentification of human works as AI points to changing listening criteria shaped by increasingly capable DAW tools and standardised sonic traits such as compressed timbres, quantised timing and platform-optimised mixes. As these features become normalised, a hybrid workflow is likely to emerge (White, 2025), in which composition, generation, editing and selection are integrated, blurring authorship with curation. These dynamics unfold alongside ongoing copyright debates surrounding AI-generated music, including whether the use of training data confers any claim of authorship over model outputs (Sturm et al., 2019; Bulayenko et al., 2022; Berkowitz, 2024; Shroff, 2024). In this setting, creativity and authorship are increasingly evaluated by affective plausibility and genre-typical surface cues rather than by deliberate craft, raising questions about agency, labour visibility and cultural capital in the music economy. These trends call for integrated theory-building across music cognition, media psychology and cultural studies to explain how familiarity, expectation and socio-technical framing shape judgements of authenticity.

#### 4.5 Stylistic Misattribution of Rock Music

An interesting result concerns Rock music, which was the only genre in which human-produced excerpts fell below the 0.50 threshold (see Fig. 3 and Tab. 3). Rock also showed the greatest Human→AI misattribution, 21 of 40 songs, yielding a correct identification rate of 47.5% (see Fig. 4 and Tab. 4). Although AI excerpts were identified with modest accuracy (*Udio* 65%; *Suno* 52.5%), the extent of misattribution for human Rock raises questions about emotional expressivity and listener expectations in this genre. This pattern aligns with Chia et al. (2025), who argue that genres such as Rock and Pop are overrepresented in training data; however, their account does not consider misattribution of human material. The finding may also reflect hybrid workflows, described by Hasan (2024), in which AI tools embedded in DAWs shape writing, recording and mixing, thereby obscuring human cues in the creative process. Additionally, Rock's relatively formulaic structures (Fornäs, 1995) may obscure perceptual markers of human-produced music.

#### 4.6 Jazz and R&B: Participants' Familiarity

A parallel emerged between identification rates for Jazz and R&B. As shown in Tab. 3, human-produced excerpts remained relatively high in accuracy (*Jazz* 70%,  $p = 0.0166$ ; *R&B* 80%,  $p = 0.0002$ ), whereas AI excerpts were identified with substantially lower accuracy in both genres (*Jazz*: *Udio* 22.5%,  $p = 0.0007$ ; *Suno* 30%,  $p = 0.0166$ ; *R&B*: *Udio* 45%,  $p = 0.6358$ ; *Suno* 20%,  $p = 0.0002$ ). A definitive explanation is beyond the scope of the present study, but one plausible hypothesis is differential familiarity with the genres. Given that both genres originate in African-American musical traditions and that the participant pool comprised mainland Chinese students and faculty, participants may have had comparatively less exposure to stylistic nuances required for accurate identification. Future work should test this familiarity account by (1) recruiting more demographically and culturally diverse participant samples to assess generalisability; and (2) including additional genre conditions that vary in cultural proximity for the same participant pool (for example, Latin American or Caribbean styles such as tango, samba, or reggae) to evaluate whether similar patterns emerge.

#### 4.7 Suggestions for Academic Policy and Practice

Conservatory programmes may benefit from the development of courses based on AI music generation so that students become familiar with its sonic signatures and typical artefacts. This study indicates that prior exposure improves the ability to identify AI-produced music, supporting curricula that combine hands-on prompting with critical listening to micro-timing, phrasing, timbre, and mix cues, thereby reinforcing authorship recognition (RQ1) and clarifying genre-

contingent cues of stylistic authenticity (RQ2). Given the rapid evolution of these tools, institutions should research their pedagogical impact and teach integration into existing DAW and creative workflows, strengthening technological literacy and documenting how experience shapes perceptual accuracy (RQ3). Courses should include a hybrid workflow (White, 2025) literacy across composition, generation, editing and curation, alongside reflective seminars on copyright, authorship, and training data, to frame expectations and reduce attribution bias and trust gaps (RQ4). Assessment rubrics should anticipate AI use, specify permitted forms of assistance, and evaluate compositional intent and perceptual plausibility as separate criteria, aligning evaluation with RQ1 and RQ2. Faculty development is recommended so instructors can model responsible use, set clear evaluation protocols, and align teaching with institutional policy, providing a consistent environment for measuring learning gains on recognition, genre effects, experience effects and attribution patterns across RQ1–RQ4.

#### 4.8 Alignment with Existing Research

The study corroborates recent literature (White et al., 2025; van Schaik, 2024; Hernandez-Olivan & Beltran, 2021), indicating that listeners increasingly struggle to differentiate human and AI-generated music, while also observing both emotional engagement with AI compositions (van Schaik, 2024) and recognition of their structural deficiencies (Hernandez-Olivan & Beltran, 2021). The findings further affirm the influence of listener characteristics, including musical training, cultural background and prior exposure to AI-generated music, highlighted by Olayeni (2023), and they support ethical considerations regarding authorship and creative agency raised by Canyakan (2024) and Patil (2023).

#### 4.9 Study Limitations and Future Directions

Although the study recruited 120 musically trained participants, all were drawn from a single institution, which limits generalisability. Future work should broaden sampling across cultural and educational contexts, include additional AI models (for example, *Mureka AI*, released after data collection), and test unfamiliar or hybrid genres to yield more comprehensive results. Longitudinal designs are warranted to examine perceptual adaptation, assessing whether priming or repeated exposure to AI music leads to desensitisation or improved discrimination, thereby offering a dynamic account of how auditory cognition evolves in response to AI-generated sound. In parallel, rapid advances in AI music generation raise methodological considerations. The challenges with audio source separation of AI-generated music can introduce artefacts that confound timbre, balance and micro-timing, and the nature of these artefacts shifts as algorithms evolve. Subsequent studies should document the separation pipeline in detail, including model choice, versioning and parameters, and, where feasible, should compare analyses on raw stems *versus* separated mixes to quantify separation-induced bias over time, aligning reporting with representation-level concerns in tokenisation and musical encoding (White, 2025).

### 5 Conclusion

The findings of this study demonstrate that musically trained listeners identify human-produced music more accurately than AI-generated music; that detectability is strongly genre-dependent; and that prior AI creation experience improves AI detection rates. Chinese Traditional and Western Classical excerpts were most often identified correctly across production methods; Pop and Rock hovered near chance; and Jazz and R&B exhibited the lowest AI detectability. Human-produced Rock was frequently misidentified as AI-generated. These results warrant genre-specific benchmarks, routine reporting of confusion matrices, and treating the direction of misattribution as a primary outcome. For researchers, they motivate trial-level models with clear effect sizes and genre-sensitive thresholds; for musicians and educators, they support curricula combining hands-on generation practice with feature-level listening; and for the creative sector, they call for richer metadata and context-aware disclosure in production workflows. Taken together, these patterns support a broader interpretive claim: judgements of musical “creativity” are ultimately grounded in perceptual authenticity, that is, listeners’ beliefs about human agency and their social connection to the artist (White, 2025). Thus, even as systems advance and may suffice for many commercial purposes, lasting musical engagement will remain rooted in a personal, emotionally mediated connection to a human creator.

## Ethical Approval Statement

This study complied with established ethical standards for research involving human participants. The research protocol was approved by the Ethics Committee of the Pop Music Department, Sichuan Conservatory of Music (No. 01/2025). All participants provided informed verbal consent before taking part and had the option not to participate.

## Ethical Use of AI Disclosure Statement

This article was developed with the assistance of generative artificial intelligence tools, specifically *ChatGPT*, *Gemini* and *Grammarly*. These tools were used exclusively for non-empirical tasks such as idea generation, conceptual brainstorming, background research (which was subsequently verified for accuracy) and proofreading. No AI tools were used for data collection, statistical analysis, interpretation of results, or the creation of graphs, tables, or figures. All research design, data handling and analytical content reflect the independent work of the author.

## Appendix A: Google Drive Links to Raw Data, Data Analysis, Audio Groups and Human-Produced Songs

### Raw Data

<https://drive.google.com/file/d/1BIrZPzgpt72AuNJPbz9nCzQsNRymmsrb/view?usp=sharing>

### Audio Groups

<https://drive.google.com/file/d/1nw4wkw9mquK3nkUpeB2q1ur88oC3lBzw/view?usp=sharing>

## Appendix B: Pre-Listening Participant Background Data Questionnaire (Translated from Chinese)

Audio Group Number \_\_\_\_\_

1. **Age:** \_\_\_\_\_
2. **Gender:**
  - Male
  - Female
3. **Musical Background:**
  - Western Classical
  - Jazz
  - Pop
  - Other (please specify): \_\_\_\_\_
4. **Primary Instrument:** \_\_\_\_\_
5. **Do you have perfect pitch?**
  - Yes
  - No

6. **How familiar are you with AI Music technology?**
- Not at all
  - Somewhat familiar
  - Very familiar
7. **Have you previously created AI-generated music?**
- Yes
  - No
8. **Cultural Background and Ethnic Group:** \_\_\_\_\_

## References

- Aljanaki, A. (1987). *Emotion in music: Representation and computational modelling* [Doctoral dissertation, University of Utrecht]. Utrecht University Student Theses Repository. <https://dspace.library.uu.nl/bitstream/handle/1874/339995/Aljanaki.pdf?sequence=1>
- Balkwill, L.-L., & Thompson, W. F. (1999). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception*, 17(1), 43–64. <https://doi.org/10.2307/40285811>
- Bell, A. P. (2025). Pedagogy of the prompt: Music education, artificial intelligence, and big tech magic. *Action, Criticism & Theory for Music Education*, 24(3), 226–229. <https://doi.org/10.22176/act24.3.202>
- Berger, S., & Neitsch, J. (2023). Investigating and comparing remote recording methods. In M. S. Lundmark, A. Asadi, S. Berger, & O. Niebuhr (Eds.), *Proceedings of the 13th Nordic Prosody Conference* (pp. 200–211). Sciendo. <https://doi.org/10.2478/9788366675728-017>
- Berkowitz, A. E. (2024). “Gimme some truth”: AI music and implications for copyright and cataloging. *Information Technology and Libraries*, 43(3), 2–11. <https://doi.org/10.5860/ital.v43i3.17072>
- Betensky, R. A. (2019). The *p*-value requires context, not a threshold. *The American Statistician*, 73(sup1), 115–117. <https://doi.org/10.1080/00031305.2018.1529624>
- Boden, M. A. (2004). *The creative mind: Myths and mechanisms* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203508527>
- Bown, O. (2025). Music AI’s global reach. *Journal of Popular Music Studies*, 37(2), 112–116. <https://doi.org/10.1525/jpms.2025.37.2.112>
- Bulayenko, O., Quintais, J. P., Gervais, D. J., & Poort, J. (2022, February 28). AI music outputs: Challenges to the copyright legal framework. SSRN. <https://doi.org/10.2139/ssrn.4072806>
- Cambouropoulos, E. (2010). The musical surface: Challenging basic assumptions. *Musicae Scientiae*, 14(2), 131–147. <https://doi.org/10.1177/10298649100140s209>
- Canyakan, S. (2024). The role of AI in creative processes: Ethical and legal perspectives in the music industry. *Journal of Music Theory and Transcultural Music Studies*, 2(2), 143–158. <https://doi.org/10.5281/zenodo.15031855>
- Carnovalini, F., & Rodà, A. (2020). Computational creativity and music generation systems: An introduction to the state of the art. *Frontiers in Artificial Intelligence*, 3(14), 2–16. <https://doi.org/10.3389/frai.2020.00014>
- Casini, L., Cros Vila, L., Dalmazzo, D., Kaila, A.-K., & Sturm, B. L. T. (2025, September 15). *Data-driven analysis of text-conditioned AI-generated music: A case study with Suno and Udio* [preprint]. arXiv. <https://arxiv.org/abs/2509.11824>
- Chen, Y., & Sun, Y. (2024). The usage of artificial intelligence technology in music education system under deep learning. *IEEE Access*, 12, 130546–130556. <https://doi.org/10.1109/ACCESS.2024.3459791>
- Cheng, L. (2024). The use of digital technology in school music education: Artificial intelligence and emerging practices. In J. L. Aróstegui, C. Christophersen, J. Nichols, & K. Matsunobu (Eds.), *The Sage handbook of school music education* (pp. 381–390). SAGE. <https://doi.org/10.4135/9781529674842.n29>
- Chia, S., Hartanto, A., & Tong, E. M. W. (2025). Do listeners devalue AI-generated pop music? Exploring negative biases in listeners’ responses to AI-labelled vs human-labelled pop music. *Computers in Human Behavior: Artificial Humans*, 6, 100217. <https://doi.org/10.1016/j.chbah.2025.100217>
- Chuan, C.-H. (2013). A multimodal approach to song-level style identification in pop/rock using similarity metrics. In *Proceedings of the 12th International Conference on Machine Learning and Applications* (Vol. 2, pp. 321–324). IEEE. <https://doi.org/10.1109/ICMLA.2013.143>

- Deruty, E., Grachten, M., Lattner, S., Nistal, J., & Aouameur, C. (2022). On the development and practice of AI technology for contemporary popular music production. *Transactions of the International Society for Music Information Retrieval*, 5(1), 35–49. <https://doi.org/10.5334/tismir.100>
- DeVereaux, C., Höhne, S., & Tröndle, M. (2025). Cultural management and policy transformations: Music and AI in dialogue. *Journal of Cultural Management and Cultural Policy*, 11(2), 311–314. <https://doi.org/10.1177/27018466251376755>
- Drott, E. (2021). Copyright, compensation, and commons in the music AI industry. *Creative Industries Journal*, 14(2), 190–207. <https://doi.org/10.1080/17510694.2020.1839702>
- Fiorino, S.(2025). *Ajuste de modelos de difusión para la generación de audio* [Thesis, University of Buenos Aires]. University of Buenos Aires Repository. <https://gestion.dc.uba.ar/media/academico/grade/tesis/Tesis de Licenciatura - Santiago Fiorino.pdf>
- Fornäs, J. (1995). The future of rock: Discourses that struggle to define a genre. *Popular Music*, 14(1), 111–125. <https://doi.org/10.1017/S0261143000007650>
- Guo, J., Suttachitt, N., & Charoensloong, T. (2024). Examining the role of cultural influences in melody creation methods and their relationship with improvisational ability in musicians. *Cultura: International Journal of Philosophy of Culture and Axiology*, 21(3), 162–188. <https://culturajournal.com/submissions/index.php/ijpca/article/download/477/392>
- Hasan, M. R. (2024). The influence of technology on modern music production. *International Journal of Humanities and Information Technology*, 6(4), 19–25. <https://doi.org/10.21590/ijhit.06.04.03>
- Hassink, N. (2024). *AI music vs. human music: Exploring artistic and economic value judgments in music* [Master's dissertation, University of Utrecht]. Utrecht University Student Theses Repository. <https://studenttheses.uu.nl/handle/20.500.12932/47602>
- Hazzard, A., Vear, C., & Moroz, S. (2024). Towards a hierarchy of trust in human-AI music-making. *GenAICHI: CHI Workshop on Generative AI and HCI*. [https://generativeaiandhci.github.io/papers/2024/genaichi2024\\_28.pdf](https://generativeaiandhci.github.io/papers/2024/genaichi2024_28.pdf)
- Hernandez-Olivan, C., & Beltran, J. R. (2021, August 27). *Music composition with deep learning: A review* [preprint]. arXiv. <https://arxiv.org/abs/2108.12290>
- Hong, J. W., Peng, Q., & Williams, D. (2021). Are you ready for artificial Mozart and Skrillex? An experiment testing expectancy violation theory and AI music. *New Media & Society*, 23(7), 1920–1935. <http://dmitriwilliams.com/wp-content/uploads/2020/06/aimusic.pdf>
- Hsu, E. (2025). *“All I miss is soul”*: Negotiating legitimacy, identity, and taste in the age of AI music [Master's Thesis, Lund University]. Lund University Repository. <https://lup.lub.lu.se/luur/download?func=downloadFile&recordId=9188606&fileId=9194545>
- Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. MIT Press. <https://doi.org/10.7551/mitpress/6575.001.0001>
- Ioannidis, J. P. A. (2018). The proposal to lower *p* value thresholds to .005. *JAMA*, 319(14), 1429–1430. <https://doi.org/10.1001/jama.2018.1536>
- Johnson, T. (2020). Chance the rapper, Spotify, and musical categorization in the 2010s. *American Music*, 38(2), 176–196. <https://doi.org/10.5406/americanmusic.38.2.0176>
- Jordanous, A. K. (2012). *Evaluating computational creativity: A standardised procedure for evaluating creative systems and its application* [Doctoral dissertation, University of Sussex]. University of Sussex Repository. [http://sro.sussex.ac.uk/44741/1/Jordanous,\\_Anna\\_Katerina.pdf](http://sro.sussex.ac.uk/44741/1/Jordanous,_Anna_Katerina.pdf)
- Kratochvil, R. (2025). *Generative AI music as a disruptive force in music consumption – Opportunities and Challenges* [Master's Thesis, Technical University Vienna]. Technical University Vienna Repository. <https://repositum.tuwien.at/bitstream/20.500.12708/224141/1/Marx%20Zsolt%20-%202025%20-%20Generative%20AI%20Music%20as%20a%20Disruptive%20Force%20in%20Music...pdf>

- Krumhansl, C. L. (1991). Music psychology: Tonal structures in perception and memory. *Annual Review of Psychology*, 42(1), 277–303. <https://doi.org/10.1146/annurev.ps.42.020191.001425>
- Kumar, A., & Sarmento, P. (2023, April 18). *From words to music: A study of subword tokenization techniques in symbolic music generation* [preprint]. arXiv. <https://arxiv.org/abs/2304.08953>
- Lecamwasam, K., & Chaudhuri, T. R. (2025, June 3). *Exploring listeners' perceptions of AI-generated and human-composed music for functional emotional applications* [preprint]. arXiv. <https://arxiv.org/abs/2506.02856>
- Lin, Y., & Weatherly, K. I. C. H. (2024). Standardizing “Yikao”: An analysis of the reform in arts college entrance examination (ACEE) in China in 2024. *International Journal of Music Education*, 0(0), 1–18. <https://doi.org/10.1177/02557614241269062>
- Ma, N., & Yu, D. (2025). Generative AI and the evolution of artistic creativity. In V. Geroimenko (Ed.), *Human-computer creativity: Generative AI in education, art, and healthcare* (pp. 177–201). Springer. [https://doi.org/10.1007/9783031865510\\_10](https://doi.org/10.1007/9783031865510_10)
- Mehta, A., Chauhan, S., & Choudhury, M. (2024, December 5). *Missing melodies: AI music generation and its “nearly” complete omission of the Global South* [preprint]. arXiv. <https://arxiv.org/abs/2412.04100>
- Miller, B. A. (2020). “All of the rules of Jazz”: Stylistic models and algorithmic creativity in human-computer improvisation. *Music Theory Online*, 26(3). <https://mtosmt.org/issues/mto.20.26.3/mto.20.26.3.miller.html>
- Mycka, J., & Mańdziuk, J. (2024). Artificial intelligence in music: Recent trends and challenges. *Neural Computing and Applications*, 37, 801–839. <https://doi.org/10.1007/s00521-024-10555-x>
- Nayar, V. (2025). The ethics of AI generated music: A case study on Suno AI. *GRACE: Global Review of AI Community Ethics*, 3(1), 1–22. <https://doi.org/10.60690/pm0vte39>
- North, A. C., & Hargreaves, D. J. (2008). *The social and applied psychology of music*. Oxford University Press.
- Olayeni, S. (2023). *The impact of artificial intelligence (AI) in music business industry* [Thesis, Centria University of Applied Sciences]. Centria University Repository. <https://urn.fi/URN:NBN:fi:amk-2023102327854>
- Owen, K. R. (2025). *The creation and life of “artificial intelligence and music”: How practices of design and use shape the direction of AI and work in a creative industry*. [Doctoral Thesis, University of Southampton]. University of Southampton Repository. [https://eprints.soton.ac.uk/505917/1/KO\\_Thesis\\_Final\\_pdf\\_a3.pdf](https://eprints.soton.ac.uk/505917/1/KO_Thesis_Final_pdf_a3.pdf)
- Pujari, V., & Wilson, B. (2023). Copyright and authorship in AI generated music. *Journal of Emerging Technologies and Innovative Research*, 10(12), 351–354.
- Rahman, M. A., Hakim, Z. I. A., Sarker, N. H., Paul, B., & Fattah, S. A. (2024, August 26). *SONICS: Synthetic or not – Identifying counterfeit songs* [preprint]. arXiv. <https://arxiv.org/abs/2408.14080>
- Ribeiro, T., & Marins, P. (2024). The integration of artificial intelligence in music education: Initial categorisation and potential applications. In E. Himonides, C. Johnson, H. Prior, & A. King (Eds.), *Proceedings of Sempre MET2024 International Conference on Music-Education-Technology* (pp. 12–14). International Music Education Research Centre (iMerc) Press. <https://doi.org/10.17605/OSF.IO/WE2AV>
- Rickard, E. (2022). *Generating music using AI* [Master’s Thesis, Lund University]. Lund University Repository <https://lup.lub.lu.se/luur/download?func=downloadFile&recordId=9093922&file-Id=9093927>
- Riedl, L. (2025). *Videos mit künstlicher Intelligenz gestalten: Kreative Tools und professionelle Workflows für die Videoproduktion*. Springer Vieweg. <https://doi.org/10.1007/978-3-658-46663-3>
- Scherer, K., & Zentner, M. (2008). Music evoked emotions are different – more often aesthetic than utilitarian. *Behavioral and Brain Sciences*, 31(5), 595–596. <https://doi.org/10.1017/s0140525x08005505>

- Schwartz, E. H. (2025, September 30). I tried Suno studio the new platform that mixes AI music generation with hands-on editing – Like GarageBand, but smarter. *TechRadar*. <https://www.techradar.com/ai-platforms-assistants/i-tried-suno-studio-the-new-platform-that-mixes-ai-music-generation-with-hands-on-editing-like-garageband-but-smarter>
- Shank, D. B., Stefanik, C., Stuhlsatz, C., Kacirek, K., & Belfi, A. M. (2023). AI composer bias: Listeners like music less when they think it was composed by an AI. *Journal of Experimental Psychology: Applied*, 29(3), 676–692. <https://doi.org/10.1037/xap0000447>
- Shroff, L. (2024). AI & copyright: A case study of the music industry. *Stanford OJS Journals*, 2(1), 3–4. <https://doi.org/10.60690/7pqxkr18>
- Stevens, C. J. (2012). Music perception and cognition: A review of recent cross-cultural research. *Topics in Cognitive Science*, 4(4), 653–667. <https://doi.org/10.1111/j.1756-8765.2012.01215.x>
- Strauss, H., Vigl, J., Jacobsen, P., Bayer, M., Talamini, F., Vigl, W., Zangerle, E., & Zentner, M. (2024). The Emotion-to-Music Mapping Atlas (EMMA): A systematically organized online database of emotionally evocative music excerpts. *Behavior Research Methods*, 56(4), 3560–3577. <https://doi.org/10.3758/s13428-024-02336-0>
- Sturm, B. L. T., Iglesias, M., Ben-Tal, O., Miron, M., & Gómez, E. (2019). Artificial intelligence and music: Open questions of copyright law and engineering praxis. *Arts*, 8(3), 3–4. <https://doi.org/10.3390/arts8030115>
- Surbhi, A., & Roy, D. (2024). Tunes of tomorrow: Copyright and AI-generated music in the digital age. In D. Roy & G. Fragulis (Eds.), *ETLTC2024 International Conference Series on ICT, Entertainment Technologies, and Intelligent Information Management in Education and Industry*. AIP Conference Proceedings, 3220(1), 050003. AIP Publishing. <https://doi.org/10.1063/5.0234946>
- Susino, M., & Schubert, E. (2018). Cultural stereotyping of emotional responses to music genre. *Psychology of Music*, 47(3), 342–357. <https://doi.org/10.1177/0305735618755886>
- Tariq, S., Iftikhar, A., Chaudhary, P., & Khurshid, K. (2022). Examining some serious challenges and possibility of AI emulating human emotions, consciousness, understanding and “Self”. *Journal of NeuroPhilosophy*, 1(1), 55–75. <https://doi.org/10.5281/zenodo.6637757>
- van Schaik, S. (2024). *Can AI-generated music induce emotions? A mixed-methods research study on the emotional impact of AI-generated music* [Master’s Thesis, Stockholm University]. Stockholm University Repository. <https://www.diva-portal.org/smash/get/diva2:1955553/FULLTEXT01.pdf>
- Vengathattil, S. (2025). Collaborative AI in music composition: Human-AI symbiosis in creative processes. *International Journal of Management Science and Information Technology*, 5(1), 253–262. <https://doi.org/10.35870/ijmsit.v5i1.4085>
- Vila, L. C., Sturm, B. L. T., Casini, L., & Dalmazzo, D. (2025). The AI music arms race: On the detection of AI-generated music. *Transactions of the International Society for Music Information Retrieval*, 8(1), 179–194. <https://doi.org/10.5334/tismir.254>
- Wang, C. (2025). *Beyond authorship: Human-AI collaboration and open creative processes in music*. Hal Open Science. <https://hal.science/hal-05102298>
- White, C. W. (2025). *The AI music problem: Why machine learning conflicts with musical creativity*. Taylor & Francis.
- White, C. W., Kapoor, K., Cosme-Clifford, N., Symons, J., & von Mutius, L. (2025, January). *Humans perceive AI-generated music as less expressive than comparable human-made content* [preprint]. SSRN. <https://doi.org/10.2139/ssrn.5087035>
- Wittschen, M. (2025). AI vs. human-produced music: Technical, perceptual, creative, and ethical dimensions. *Capstone, The UNC Asheville Journal of Undergraduate Scholarship*, 38(1). <https://jane-way.uncpress.org/capstone/article/id/2075/download/pdf>
- Xiong, Z., Wang, W., Yu, J., Lin, Y., & Wang, Z. (2023, August 26). *A comprehensive survey for evaluation methodologies of AI-generated music* [preprint]. arXiv. <https://arxiv.org/abs/2308.13736>

Youvan, D. C. (2025). *Tokenese: Designing a human language optimized for AI tokenization and efficient communication* [preprint]. <https://doi.org/10.13140/RG.2.2.16342.77123>

Zenieris, R. (2023). *Perception and bias towards AI music*. University of Twente Repository. <https://essay.utwente.nl/96214/1/Digital%20Transformation%20of%20Music%20%287%29.pdf>